Lifelogging with *SAESNEG*: A System for the Automated Extraction of Social Network Event Groups

Benjamin Robert Howard Blamey

This research was conducted at
Cardiff Metropolitan University, and is submitted
in partial fulfilment for the award of
of Doctor of Philosophy, University of Wales.

June 2015

Director of Studies: Dr. Tom Crick

Supervisor: Dr. Giles Oatley

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references.
A bibliography is appended.

Signed .... (candidate)

Date ....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ....

Abstract

This thesis presents *SAESNEG*, a *System for the Automated Extraction of Social Network Event Groups*; a pipeline for the aggregation of the personal social media footprint, and its partitioning into events, the “event clustering” problem. *SAESNEG* facilitates a reminiscence-friendly user experience, where the user is able to navigate their social media footprint. A range of socio-technical issues are explored: the challenges to reminiscence, lifelogging, ownership, and digital death.

Whilst previous systems have focused on the organisation of a single type of data, such as photos or Tweets respectively; *SAESNEG* handles a variety of types of social network documents found in a typical footprint (e.g. photos, Tweets, check-ins), with a variety of image, text and other metadata – *differently heterogeneous* data; adapted to sparse, private events typical of the personal social media footprint.

Phase A extracts information, focusing on natural language processing; new techniques are developed; including a novel *distributed approach* to handling temporal expressions, and a parser for social events (such as birthdays). Information is also extracted from image and metadata, the resultant annotations feeding the subsequent event clustering. Phase B performs event clustering through the application of a number of pairwise similarity *strategies* – a mixture of new and existing algorithms. Clustering itself is achieved by combining machine-learning with correlation clustering.

The main contributions of this thesis are the identification of the technical research task (and the associated social need), the development of novel algorithms and approaches, and the integration of these with existing algorithms to form the pipeline. Results demonstrate *SAESNEG*’s capability to perform event clustering on a differently heterogeneous dataset, enabling users to achieve lifelogging in the context of their existing social media networks.

Contents

Listings

List of Figures

List of Tables

Chapter 1

Introduction

> “In 10 years, every human connected to the Internet will have a timeline. It will contain everything we’ve done since we started recording, and it will be the primary tool with which we administer our lives. This will fundamentally change how we live, love, work, and play. And we’ll look back at the time before our feed started – before Year Zero – as a huge, unknowable black hole.” (Croll, 2015)

1.1 Introduction

This thesis documents the development of *SAESNEG*: a System for the Automated Extraction of Social Network Event Groups, for organising a user’s social media footprint into *events*, to facilitate a reminiscence-friendly user experience of that data – adapting and extending existing work to this novel context.

It builds on two key notions:

- The **need** for a lifelogging or reminiscence-friendly user experience in the context of online social networks, and the heterogeneous social media footprints of their users.
- The **existence** of a range of algorithms and systems for event detection – in uniform datasets such as collections of images or text documents.

Having read this first section, the reader will have an overview of the data-mining task at the core of the thesis, and an introduction to how it was approached, through the means of a clear, real-world example. Subsequent sections of this introductory chapter will examine in more detail the rationale (section 1.3), formal problem statement (section 1.4), source dataset (section 1.5) and an outline of the contributions (section 7.2). Detailed discussion of formal definitions and related work is intentionally deferred, to deliver a short introductory chapter – however, there are extensive references to related sections in the thesis.

A person may have multiple accounts across different social networks, to which content will be posted regularly – this aggregated set of documents can be referred to as that user's *social media footprint*. Given this (potentially large) set of social media documents belonging to a user, the task is to partition this set according to the underlying real-world *events* related to each item.

1.2 Motivating Example

As an example, figure 1.1 shows three documents from the author's social media footprint. The three documents are associated the same event – a friend's wedding, and were created and uploaded by 3 different users:

(A) A photo, uploaded to Flickr, by 'A'.

(B) A photo, uploaded to Facebook, by 'B'.

(C) A status message, created on Facebook, by 'C'.

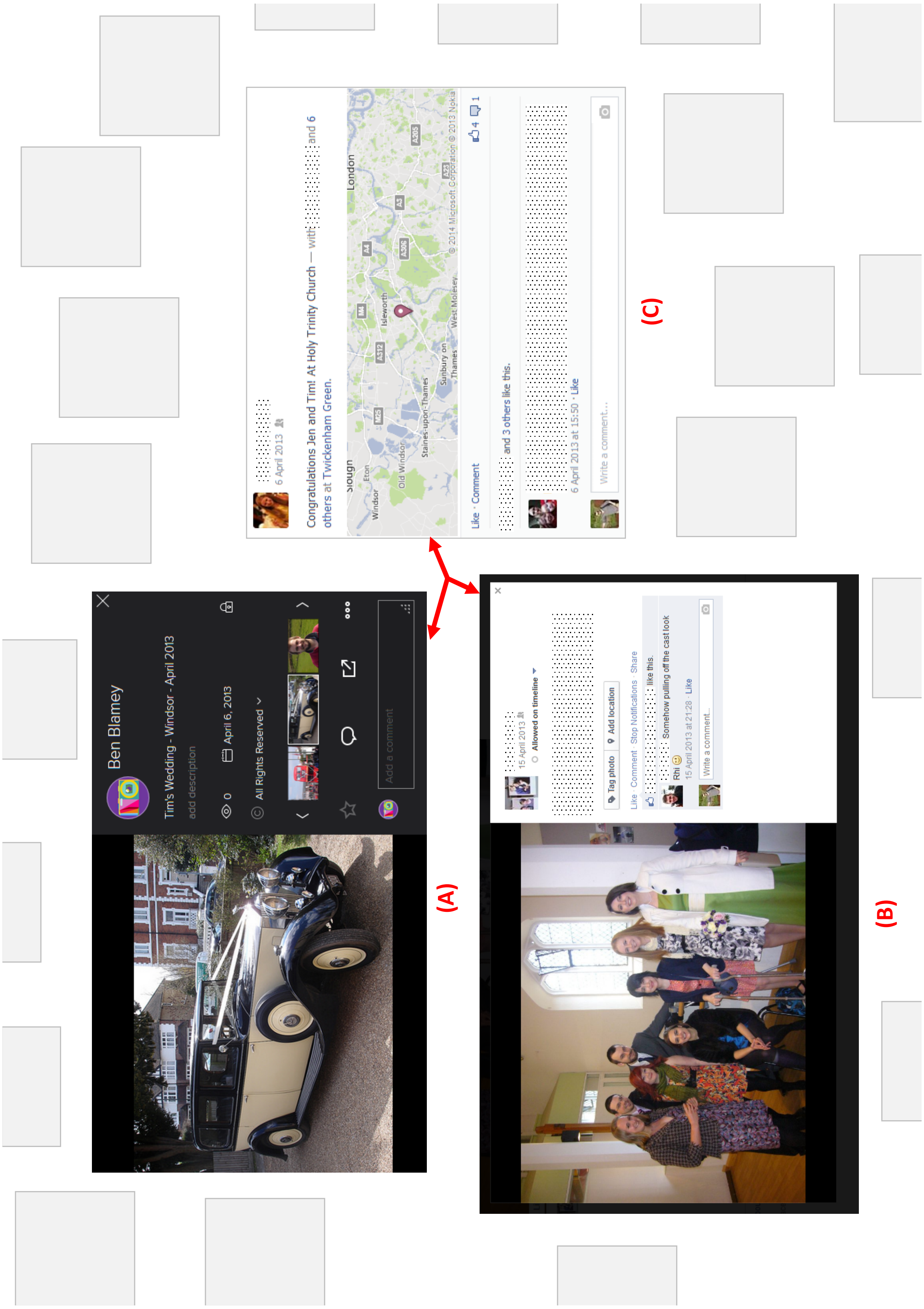
The research task is to develop a system able to identify, for example, that these three items of content relate to the same event (from amongst other posts in the larger social media footprint), so that they can be presented together, rather than being separated in the stream of content. It is argued that by partitioning a user's entire social media footprint into such *event groups*, the resulting sequence creates a form of *life story* for the user.

Initially, the documents reside at their original OSN (online social network), and might be displayed individually in chronological order. Conversely, a system capable of document event clustering enables event-based visualisation and navigation of the data, perhaps for reminiscence; or other purposes – or as a preliminary step in a variety of other applications. Event-based visualisation is arguably recognised as effectively the *de facto* means of displaying personal content for reminiscence purposes, whilst various authors have identified problems with the way the OSNs currently display data (see section 2.3).

Having downloaded the documents from the user's social networking account(s), there are a number of strategies that could be used to inform the event partitioning decision, some of which can be adapted from existing work:

Friends – Several people have been tagged[1] in (B) and (C), with several people in both items of content. 'A' who uploaded (A), is also tagged in both (B) and (C). Alternatively, people may be mentioned in text by name – named entity extraction (section 3.2.6.4) could be used to gather this information. This information may indicate attendance of the individuals concerned at the underlying events(s). Therefore, if documents have multiple people in common, it may be an indication that the content relates to the same real-world event. One must not overlook the person who took the photo or wrote the status message, which similarly suggests presence at the event.

[1] Generally, tagging means that an annotation has been added to the document; in the case of Facebook photos, it specifically refers to an annotation with the name of a person, and where geometrically they appear in the photo – not to be confused with geo-tagging, associating the document with a place.



Ben Blamey

Tim's Wedding - Windsor - April 2013

add description

0

April 6, 2013

All Rights Reserved

Add a comment

[redacted]

15 April 2013

Allowed on timeline

Tag photo · Add location

Like · Comment · Stop Notifications · Share

[redacted] like this.

Rhi 😊 Somehow pulling off the cast look

15 April 2013 at 21:28 · Like

Write a comment...

[redacted]

6 April 2013

Congratulations Jen and Tim! At Holy Trinity Church — with [redacted] and 6 others at Twickenham Green.

Like · Comment

[redacted] and 3 others like this.

[redacted]

6 April 2013 at 15:50 · Like

Write a comment...

Figure 1.1: Example Document Clustering Scenario

Spatial – The status message (C) was tagged with a location in the metadata, whereas (A) includes the nearby place 'Windsor' in the title. After processing the text with NER (named entity recognition, section 3.2.6.5), the geographical proximity of these places may indicate that they relate to the same event.

Temporal – The status message (C) was created on 6th April, whilst photo (B) was uploaded on the 15th April (so the event certainly started before the 15th). Because (B) originates from Facebook, the EXIF metadata (which may contain the photo creation time) is not available, but the event in question must have occurred at or before the photo upload time. (A) originates from Flickr (where the EXIF metadata is available) indicating the photo was actually taken on the 6th April, the same date as (C). Photo (B) being uploaded around a week afterwards may indicate it originates from the same event. In other cases, the text may contain temporal expressions which could be parsed (section 5.5) and compared with those extracted from metadata. This mixture of different kinds of documents introduces complexity into handling the temporal dimension – temporal information in different documents can have different semantics. With uniform datasets comparison can be easier: temporal data obtained from a single method could be used – in the simplest case simply using the timestamps of the data, as with GPS-labelled data (Gong et al., 2006). This temporal aspect is seen as crucial in existing systems, and is the focus of a novel approach developed in this thesis (section 5.5.1).

Scene – Two photographs of the same (or similar) scene may indicate that they relate to the same event, depending on other evidence. Techniques from the field of content-based image retrieval could be used to compare image content to facilitate such reasoning. However, photographs could be substantially different, and still relate to the same event, as the example illustrates.

Other strategies are perhaps not so obvious, and are developed in this thesis:

Well-Known Event Semantics – (A) mentions the word 'Wedding' in the textual description – other wedding-related content might originate from the same event, whereas content relating to different kinds of well-known events such as a birthdays would not. By parsing additional semantics such as the names of the people getting married or the age associated with the birthday, further reasoning may be possible. Techniques from natural language processing (NLP), such as named entity recognition (section 3.2.6.5) could be adapted for this purpose (section 5.6).

User Structures – If any of these items of content were already organised into a structure by the user (such as a photo album), it may suggest that other content also relates to the same event. Because of this, photo albums are sometimes used as ground truth in existing work, but such assumptions are not always accurate – an album may not represent an event, and may simply be 'Summer' or 'Random', or a combination of two events. Textual data associated with the album could feed into other strategies, perhaps containing temporal or spatial information.

Type – When comparing two items of content, to decide whether they relate to the same event, the types of document could be used to determine an *a priori* probability of the documents relating to the same event, or otherwise used as a starting-point in that decision making process.

Each of these strategies may yield positive or negative evidence for event commonality (or indeed no evidence at all). The intention is that by combining such evidence the system would be able to perform the document event clustering task (chapter 6) on the social media footprint. This work is built on a significant body of research on systems for event detection, in a variety of domains – and the techniques that support them. In particular, systems for detecting events in textual data (section 3.4.6) and collections of images (section 3.4.4).

The remainder of the chapter is as follows: section 1.3 gives an overview of the rationale and context for the study (in an overview of chapter 2) to establish the key contributions of the thesis. Section 1.4 formally sets out the research problem, whilst section 1.5 discusses the source data, with discussion of the characteristics which make this thesis a unique challenge. There are three areas of key contributions (section 7.2): the wider socio-technical motivation for the system, the system itself: *SAESNEG*, the work on individual algorithms and techniques that support it – and finally, the results and performance evaluation. The chapter concludes with section 1.8 – which outlines the structure of thesis.

1.3 Rationale

This section introduces the motivations for the project – and its key contributions to the domain. It is broadly an overview of chapter 2, the first of two literature review chapters, which examines existing work in relation to the wider social context and rationale, as well as reviewing related systems from a more socio-technical view-point. Because of the diverse fields, techniques and existing systems upon which the thesis draws, a second literature review (chapter 3) is devoted entirely to a more technical review of related work.

Chapter 2 introduces the key theme of *reminiscence*: people have traditionally looked back on personal photographs and documents, and attach value to such artefacts – photographs have long been central to reminiscence-related activities (Banks, 2011). With digital photography, there was a growing trend towards greater volumes of photographs. With so many photographs on our devices, systems for their management and experience became a research focus in the 1990s and into the next decade. Research developed algorithms for organising and displaying these photos to facilitate a better user experience: events were often central to such systems, a number of which are reviewed from a user experience perspective in section 2.3.8. The notion of the event (Westermann and Jain, 2007) continues as a ubiquitous model for organising personal data to the present day and features in a range of existing work reviewed in chapters 2 and 3.

This thesis is concerned with reminiscence in the context of online social networks (OSNs), and is preceded by a history of systems relating to personal data, such as these photo-management systems. However, there are three trends, or characteristic features associated with the use of OSNs which make this context a new and unique challenge: (i) a greater volume of data, (ii) the fragmentation of data across multiple services and devices, and (iii) a trend towards more diverse kinds of social media data – other than photographs. These trends create distinct, recognised challenges for the way people experience the past through social networks. The commercial and academic spheres have responded to these challenges with a number of solutions, with both of their approaches critically reviewed. From there, the thesis argues that there exists an as-yet unaddressed need for the development of system capable of organising a variety of different *kinds* of social networking data into events – the development of a theoretical framework, and implementation in *SAESNEG*, are the overarching aims of this thesis.

A second key topic to introduce is *thanatosensitivity* – the issue of death in the context of personal digital information, and its far-reaching implications for issues of data ownership, privacy, user interface design, security, and backup.[2]

Chapter 2 continues with discussion, and sets out the relevant differences between these personal photo collections and a user's social media footprint as three distinct trends, creating challenges which make *SAESNEG* distinct from previous work.

The coming of online social networks (OSNs) in the 2000s, and their rapid subsequent growth in popularity ever since has done little to dampen the trend towards **greater volume** (section 2.3.1) – not least because it is no longer one's own content that needs to be displayed, it is the aggregate content of a network of friends or contacts – each publishing a constant stream of media. Meanwhile, camera-equipped smart phones have become ubiquitous, with photos easily uploaded to social networks direct from smart phones. Rather than delivering a reminiscence-friendly user experience, authors have argued that services "focus on the now" (Gerlitz, 2012), displaying the most recent activity, thereby encouraging users to regularly add more content to the platforms.

The migration to online social networks has also lead to **fragmentation** (section 2.3.2). With digital photography, users would typically store their photographs on their own devices: the transition to social networks has meant that this single, personal repository no longer exists – personal documents are distributed amongst a collection of social networks and cloud services, and across multiple devices – with obvious implications for risks of losing (or simply forgetting about) personal content fragmented across these platforms, as well as obstructing a unified view of all the data for reminiscence purposes (sections 2.3.2 and 2.3.7).

Of crucial importance to this thesis is another change brought about by the use of social networks: the collections of personal documents are no longer restricted to photographs; social networks (and their corresponding mobile apps) allow the creation of **many different kinds of data**. The typical social media footprint now includes 'Check-Ins' record visits to particular places and venues, short status/micro-blog messages, as well as the usual stream of photographs – now accompanied with comments, likes, descriptions and other metadata (location, timestamp, etc.); the characteristics of the source data is discussed in more detail in section 4.10.

Chapter 2 continues with a critical review (section 2.3.8) of solutions to these challenges, from both academic and commercial realms, such as password services (2.3.9)- where passwords can be saved for release after the death of the owner of a social network account. Clearly, securing the passwords does not secure the data, and such

[2] Other developments sought to bridge the gap between digital and physical in different ways – including digital photo frames, and systems for the production of printed photo-books from digital photographs. Automating printed photo-book creation from digital photographs (including those on social networks) has been a research focus, see section 2.3.13.2. This existing work, and other similar experiments are discussed in subsequent chapters.

arrangements may contravene the conditions of particular OSNs.

OSNs have attempted to encourage a reminiscence-friendly experience by means of a large chronological list of content, such as Facebook Timeline. It seems to rely on explicitly-created metadata (such as work and relationship history), and does not visibly employ much data-mining. Content relating to different events is muddled together according to the basic chronological ordering (as is the case in the Twitter timeline), whilst being fragmented across various social networks.

Commercial OSN backup services go some way to addressing the issue of data loss, with content aggregated from multiple social networks. Such services may have limited functionality for automatically organising the content, usually relying on chronological ordering, and user-created structures for presentation and navigation (as with Facebook's timeline). Furthermore, due to the volatility of tech start-ups, the services can be subject to the same business pressures as other OSNs, and are similarly susceptible to closure: two examples of such services, *Loccit* and *Memolane* have closed down in recent years. Systems which store data locally on devices obviate these issues somewhat, but still lack an event-centric user experience. The key distinction between such services and the goal of this work, is that although these services collect a variety of data, none (currently) of them organise it into events. The system developed in this thesis should be equally applicable to such backup services as to OSNs themselves.

Other event-based systems (which do not handle the complete social media footprint) are reviewed. This serves two purposes: to identify techniques and approaches that might be transferrable to this task, and to establish the precedence of event-centric user interfaces (and the systems supporting them) for the display of personal data. This includes desktop photo systems, capable of organising photographs into events based on features timestamps and CBIR techniques (content-based image retrieval, section 3.3). Achieving a comparable event-centric user-experience in the context of social networking data is a key goal of this thesis.

In the context of the multimedia data-mining community, the notion of the event is so widespread that attempts have been made to formalise the notion, and standardise its representation. Other areas of existing work include printed photo books (Sandhaus et al., 2010), wearable cameras such as the *SenseCam* (Hodges et al., 2006).

Chapter 2 also introduces another key concept: lifelogging. Lifelogging is a term used to encompass various well-established research efforts, which can perhaps best be characterised as relating to the *intentional, systematic collection and digital archiving of personal data* (not necessarily social networks).

Despite this seemingly broad definition, the label is applied to a few distinct research areas. One such topic concerns systems where individuals (typically evangelists) comprehensively digitise and archive as much information as possible – recording phone calls, scanning letters, etc. Wearable cameras are also strongly associated with lifelogging, many research projects have involved the *SenseCam*, with algorithms developed to organise the huge collection of photographs taken automatically by the cameras. A strongly-related topic is that of the quantified self movement, arguably part of a wider trend towards wearable computing has gained more commercial traction, where devices are worn to record physiological data – perhaps to track fitness progression. This thesis is critical of the lifelogging research community for failing to embrace the opportunity for research into lifelogging systems based on data held in online social networks, with established means of capture – instead focusing on separate 'end-to-end' solutions.

In summary, the rationale for *SAESNEG* (and hence this thesis) is that reminiscence is hughely important, there is a long history of systems to facilitate reminiscence-related activities for our personal photo collections – such systems are mature and generally event-centric. Separately, authors have been critical of OSNs for not providing a similar reminiscence-friendly user experience, with OSN-backup services suffering a similar shortcoming. The research challenge of this thesis is to develop a system capable of enabling the event-centric organisation (which has proven successful for personal photo collections), to the larger and more diverse social networking footprint.

1.4 Formal Problem Statement

With a significant body of research relating to events, event detection and event clustering – this research task requires a precise formulation. Becker et al. (2010) describes the abstract task well:

> "Problem Definition: Consider a set of social media documents where each document is associated with an (unknown) event. Our goal is to partition this set of documents into clusters such that each cluster corresponds to all documents that are associated with one event." (Becker et al., 2010)

According to the above definition, **an event is a set of documents** – this is the representation adopted in this study. More specifically, the research task of this thesis is:

> To develop an extensible framework for the creation of a life-story from a users social-media footprint, by utilising a range of new and existing techniques for extraction of features from that footprint, and event clustering, with the application to a system for lifelogging.

The mathematical formulation of the problem is relatively simple – and introduces a number of assumptions to render the problem tractable:

- All documents relate to an event;
- Events are disjoint, this implies there are no sub-events;
- Determining exactly what lies inside the social media footprint is a subjective issue, let alone one that is easy to mathematically formalise. Instead, a simple heuristic is used;

Representation of event facets are not formalised at this stage – their representation is developed through the literature review and is linked to the development of specific algorithms later in the thesis.

$$\begin{aligned}
d &= \text{ A social network document} \\
u &= \text{ A social network user} \\
F_{complete}(u) &= \text{ The user's social media footprint} \\
F_{complete}(u) &= \{d : I(u,d) = 1\}
\end{aligned}$$

where I is the indicator function for the OSN:

$$I(u,d) = \begin{cases} 1 & d \text{ was uploaded by } u \\ 1 & u \text{ was tagged in } d \\ 1 & d \text{ was added to a structure created by } u \text{ (such as a photo album)} \\ 0 & \text{ otherwise} \end{cases}$$

There is no need to collect every single item of social networking data ever generated by the user, nor is it sensible from the standpoint of ground truth assembly – instead a sample, $F(u)$ is used.

$$F(u) \subset F_{complete}(u)$$

Having formally defined the source dataset, the notion used for events is as follows:

$$\begin{aligned}
e_{u,i} &= \text{an event} \\
e_{u,i} &= \{d_0, \ldots, d_n\}
\end{aligned}$$

For a given user u', the set of events for a user are a disjoint partitioning of the user's social media footprint:

$$\begin{aligned}
\bigcup_i e_{u',i} &= F(u') \\
e_{u',j} \cap e_{u',i} &= \emptyset \, \forall \, i,j : i \neq j
\end{aligned}$$

Furthermore, this set of event is known as the *life story*:

$$\begin{aligned}
L(u) &= \text{ The user's life story} \\
L(u) &= \{e_{u,0}, ...\} \\
L_{GT}(u) &= \text{ The ground-truth event partitioning} \\
L_C(u) &= \text{ The computed event partitioning}
\end{aligned}$$

The research task is to develop an algorithm/system capable of generating the event partitioning $L_C(u)$, given the set of source documents $F(u)$:

$$\begin{aligned}
f_{\text{SAESNEG}} &: \{d_0, ...\} \to \{\{d_0, ...\}, ...\} \\
f_{\text{SAESNEG}} &: F(u) \mapsto L_C(u)
\end{aligned}$$

This is referred to as the *document event clustering* task. The matter of 'accuracy' of the generated clustering, in comparison to the ground-truth clustering, is discussed in section 6.11.

1.5 The Social Media Footprint

The challenge of organising a lifetime worth of content is not new: there has been more than a decade of research into the organisation and presentation (typically large) collections of personal photographs. There is a similarly long history of the extraction of events from text sources, such as 'newswire'. As with photographs, the arrival of OSNs has created new research tasks – such as the extraction of events from micro-blogging corpora such as those from Twitter. In order to build on this existing literature, and identify which techniques are useful, it is necessary to analyse the task in more detail, especially the source data used for evaluation, in order to draw meaningful comparisons with existing work.

Chapter 4 presents the overall methodology for the study, and includes a detailed analysis of the source data for this study: a person's social media footprint (SMF). As formalised in the previous section, the footprint is analysed by considering it to be a set of small documents or *documents* related to the user.

Various kinds of documents may be present in the SMF – photographs, check-ins, events, Tweets, etc. By simply inspecting examples of these documents, one quickly recognises that each document can contain a variety of information, in different forms.

It is crucial to draw a distinction between the various *kinds* of documents that may be found in the social media footprint, the various *forms of data* contained within them – both distinct from the *knowledge* that may be extracted from this data. Finally, the *strategies* for comparing documents for the computation of document event commonality are again distinct, chapter 6 explains how the strategies draw on various categories of extracted information, meaning there is not a one-to-one relationship between the two layers in the theoretical model.

The distinction between kinds of document and forms of data is important. For example, a 'photo' uploaded to a social network is, in actuality, a document containing a mixture of structured metadata, various textual information, along with the image itself. A status message or Tweet is comprised of a similar mixture of information; certainly text and metadata; possibly images also[3].

Through this simple analysis, it is clear that image content and textual information do not exist in isolation, and are mostly accompanied by metadata, with images often occurring in combination with textual information. Simply by looking at representative examples of data, it is clear that social networking data is heterogeneous, documents each contain a mixture of information, and may also be completely different from each other.

[3] Aside: If all kinds of documents can be understood in this way, it raises the question as to why not dispense with the notion of the document *kind* altogether. This thesis argues that the kind of document itself contains implicit information, with underlying semantic meaning: a Tweet and a Flickr photo, whilst containing a similar mixture of data – may have been uploaded for different reasons, have a different relationship with the underlying event, and represent a different use of the social network by the user. This is crucially important for the interpretation of temporal data, see earlier in the introduction for a concrete example.

1.6 Differently Heterogenous Data

The goal of this thesis is to combine techniques for event detection/clustering in photos and multimedia with techniques for event detection/clustering in textual sources and thus to create a framework capable performing event clustering on a social media footprint containing a variety of kinds of data. These two groups of systems (and the techniques supporting them) are reviewed extensively in chapter 3. In this existing work, systems for detecting or organising content into events have (so far) tended to focus on a particular kind of source document. The two main examples being collections of photos and collections of Tweets, each of these two respective collections is at once *uniform* (i.e. all the elements in the collection are of the same kind) and *heterogeneous* (each document contains a mixture of text, image and metadata); because of this, such datasets could be called *uniformly heterogeneous*.

That the datasets in this existing work are *uniformly heterogeneous* (or at least treated as such) is a key difference to the research problem studied in this thesis. The social media footprint is considered *differently heterogeneous* because the documents themselves contain a mixture of forms of data (they are heterogenous), whilst the documents in the footprint represent different kinds of documents, with different schemas, different metadata, and different semantics (section 1.6).

This is a fastidious neologism; but also a crucial and necessary distinction between the research problems investigated in the existing work, and the one investigated in this thesis. It has a range of implications for the architecture and constituent components of the data-mining pipeline, choice of database(s), and the suitability of machine-learning techniques.

1.7 Public and Private Events

In addition to the differently heterogenous (section 1.6) nature of the data mentioned above, another key difference between much of the existing work is the distinction between public and private data/events. This study focuses on the personal social media footprint, where data (and events) tend to be non-public. Many previous studies have addressed the event identification task in the context of large, public datasets; originating from Flickr and Twitter, for example. The abundance and redundancy characterising these datasets means that high performance can be achieved with machine-learning algorithms, based on features extracted by comparatively simple techniques, focusing on ‘public’ events, whilst excluding background ‘noise’.

- External knowledge bases may contain information about public events that can be cross-referenced with the data. Conversely, your neighbour’s BBQ is unlikely to have a Wikipedia article! For private events, ground truth needs to be collected in a different way.
- For public events, with large numbers of attendees, social networking data may be available publicly. This is less likely for smaller, more private gatherings.
- Consequently, less data is available for training of machine-learning models.

- Similarly, sparse data for individual events could make the event-detection task more difficult.
- A particular challenge with Facebook data is the stringent protection of personal data, only partial information about documents may be available (see examples in appendix B).

To develop the system, the goal was to combine existing techniques into a single system capable of performing the document event clustering task on the personal social media footprint. Chapter 4 describes how these components were integrated into *SAESNEG* proceeded by details of the various developments and improvements, in chapters 5 and 6. The next chapter summarises the key contributions of this thesis, in an overview of chapter 7.

1.8 Summary and Thesis Structure

This research project addresses a real-world need using state-of-the-art data-mining techniques, combining existing work from a variety of areas, so the literature review is divided into two chapters to accommodate the wide range of work that is relevant to the study. Consequently, some of the existing systems are mentioned and reviewed in more than one chapter; in chapter 2 there is discussion of how a system relates to wider context, and the specific problem or task it is designed for, as well as user interface, while discussion of technical issues and algorithms are discussed separately in the relevant sections of chapter 3.

In chapter 2, (the first literature review chapter), there is an in-depth discussion of the important motivating concepts: lifelogging, social networks, reminiscence, and user interfaces. The chapter explains how these ideas can be combined to present an opportunity for development of a system such as that presented in the later chapters, forming the rationale for this study. There is also discussion of the some of the concepts that underpin related fields, most important being the notion of the event. The historical influence of the concept is discussed, and how it has been interpreted in different fields, and review the historical precedence of reminiscence-based systems.

Chapter 3 (the second literature review chapter) explores the existing work in more detail, and thoroughly reviews topics that underpin the functionality of these existing systems with the discussion organised into three themes: techniques, systems and approaches for handling text (i.e. NLP, *natural language processing*), techniques and approaches for handling photo/image data. The focus is on NLP (specifically a focus on information extraction techniques), noting issues such as *unnatural languages* with critique of established techniques for approaching temporal expressions, and the pros and cons of various approaches to named entity recognition. Following review of these techniques and approaches, decisions were made about what is appropriate to the problem being studied, including the realisation that volume, or 'spike'-based approaches would be unsuitable for the sparser personal social media footprint.

The next three chapters (4, 5 and 6) document the development of the system, the development of the techniques used in it, their integration and evaluation. Chapter 4 focuses on the selection of source data, and inspecting output at various stages of the pipeline, as well as a variety of other useful support tools. Storage and representation of data at the various stages in the pipeline are also discussed.

Chapter 4 presents the methodology; and gives an overview of the system, *SAESNEG* used to conduct the experiment. Some of the more trivial implementation details are discussed in this chapter. There is discussion of the infrastructure relevant to the evaluation of the entire system, the collection of ground truth, and the user interface system. It also includes details of how data (and ethical approval) was collected from users, a description of the web-based ground truth creation tool, (complete with an instructional video).

Subsequently, in the empirical work, the various parts of the system need to be described, many of which contain work which is novel in its own right. Hence, rather than presenting a separate results chapter, results are presented for different components of the system where their implementation is described.

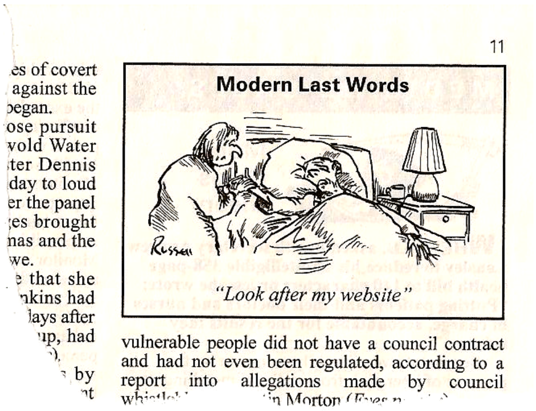
In chapter 5, the algorithms for 'Phase A' are discussed: the extraction of information from text and images. Some of the more novel techniques have their own set of results presented in this chapter. There is also separate discussion of the handling of text, image and metadata.

In chapter 6, the novel approach to event clustering is presented. The key component of our system, the integration of new and existing algorithms as the set of document event commonality strategies is followed by a final step to combine evidence from the different strategies to produce the final document event groupings.

Contributions are detailed in chapter 7; the nature of the study leads to contributions in a number of ways: *SAESNEG* as a framework, software architecture and experimentation environment, specific algorithms developed for particular components of the pipeline, as well as wider socio-technical contributions such as the addressing of the original motivating problems.

Chapter 2

Literature Review: Context



11

es of covert
against the
began.
ose pursuit
vold Water
ter Dennis
day to loud
er the panel
es brought
nas and the
we.
that she
nkins had
lays after
up, had
by
nt

Modern Last Words

"Look after my website"

vulnerable people did not have a council contract and had not even been regulated, according to a report into allegations made by council whistl in Morton (Eyes n

Figure 2.1: *Modern Last Words*, reproduced from Private Eye Magazine, January 29, 2012.

2.1 Introduction

This chapter seeks to establish clear and definitive motivation for the development of *SAESNEG* and to establish some of the key areas of literature to which this thesis contributes.

The issues raised in this chapter underpin the wider contribution: they are key to establishing the theoretical contribution of *SAESNEG* and this thesis (in addition to the isolated contributions of individual algorithms and techniques mentioned elsewhere in the thesis).

This chapter is the first of two literature review chapters, and sets the background context for the thesis. It explores why personal documents (such as photographs) are valued, the current trends toward online social networks – and resulting challenges this brings to the reminiscence experience – effectively problematising reminiscence in the context of online social networking. Chapter 2 continues by discussing some proposed solutions to these challanges.

Chapter 3, the second literature review, focuses on more technical issues: reviewing existing systems, and the data-mining techniques which underpin them. Chapter 4 draws on the literature review to present *SAESNEG*, the proposed system; incorporating a selection of techniques identified in chapter 3; whilst meeting the socio-technical challenges set out in chapter 2.

2.2 Reminiscence, Photos, Lifelogging and Thanatos

In this section, two key concepts are introduced: reminiscence and lifelogging. The importance of reminiscence is established through discussion of the widespread use of photographs for this purpose, and the changes brought about by digitisation of the means of capturing, storing, and viewing these photographs. The activity of lifelogging is explained, and the various distinct practices and research disciplines that have been associated with this label are identified. Finally, aside from lifelogging, and reminiscence, digital death[1], i.e. the legacy of personal documents (figure 2.1), and the new challenges of death in a digital context are pertinent topics which underlie much of this chapter.

2.2.1 Reminiscence and Photos

The importance of reminiscence as a human activity is well-established: “That the desire to store, organise and interact with such sentimental objects is a key human value is attested to in the many years of related anthropological and sociological research.” (Kirk et al., 2010, p. 2).

There is a history of using photographs in reminiscence related activities, as they are an undeniably ubiquitous means of capturing moments in our lives. Banks explores the topic extensively, drawing on personal experience and fieldwork: “my grandfather didn’t really start exploring his memories until he was quite old...He surrounded himself more with them.” (Banks, 2011, p. 7). This reminiscing is not necessarily a solitary act: “One of the most common and enjoyable uses for photographs is to share stories about experiences, travels, friends and family” (Landry, 2008, p. 1).

Aside from photos, Banks gives examples of other physical objects encountered in his fieldwork: diaries, a “charred plastic gear”, and an Oxo tin (Banks, 2011, p. 3), however, the focus of this thesis is on reminiscence with personal documents – specifically social networking data. First, a discussion of reminiscence with digital photos is discussed, as it leads us to extensive relevant literature – with many of the issues raised being pertinent to OSN-based reminiscence.

There has been much commercial interest in developing systems to recreate interaction similar to that with printed photos, such as story-telling (Rowe, 1989), with photographs in digital form: Microsoft has been trying to develop suitable devices for some time (Swan and Taylor, 2008; Kirk et al., 2010), but have not achieved commercial success to rival simple digital photo frames. Indeed, users have complained that some systems for photo management have too many features (Rodden and Wood, 2003).

Aside from the now ubiquitous photo frames, there has been research involving a

[1] *Thanatos* is the daemon representing death in Greek mythology.

range of specialist hardware, such as tabletop interfaces. Apted et al. (2006) developed a *SharePic* application for tabletop photo-sharing, exploring human-computer interaction and digital inclusion issues relating to elderly people. Some devices blur the boundaries between physical and digital; such as the Living Memory Box (Stevens et al., 2003), and the journalist who created a "physical representation of Twitter" by printing out his timeline on index cards (Weingarten, 2010; Harvilla, 2010)

Aside from reminiscence (Mulvenna et al., 2009), these objects often outlive their original owners, consequently "Material artefacts are passed down across generations of family members as a way of sustaining social relationships and bolstering ideas of shared heritage, history and values." (Odom et al., 2012, p. 7). Therefore, issues of thanatosensitivity (Massimi and Charise, 2009), i.e. those surrounding death, bereavement and legacy are highly relevant to the context.

A determining factor in how people interact with photos is how they choose to organise them. The arrival of digital cameras, with their ease of use, meant that people began "to take and accumulate more and more digital photos." (Lim et al., 2003, p. 2). Managing these often large[2] collections of personal photographs has been a topic of human-computer interaction research for more than a decade. These persistent efforts are evidence of a genuine need to develop systems for review and reminiscence of digital photographs; they are examined in more detail later in this chapter, and in the next.

2.2.2 Lifelogging

Lifelogging is a label which has been applied to a variety of practices and research tasks. The term has been defined as "collecting, storing and displaying one's entire life, for private use, or for friends, family, even the entire world to peruse" (trendwatching.com, 2004)[3]. Despite this extremely broad definition, the term has mostly been applied to a few distinct classes of projects and practices:

1. Comprehensive Systems such as MyLifeBits (Gemmell et al., 2006), where users (often lifelogging evangelists) have developed their own systems for the capture and storage of a broad range of personal digital information, including digitised versions of physical artefacts – e.g. email, telephone calls, physical mail, and a miscellanea of digitised documents. Such systems include means to retrieve documents through search, and often include novel visualisations of the data.

2. The use of wearable cameras, such as the *SenseCam* (Hodges et al., 2006), equipped with a variety of sensors to capture a series of photographs recording the user's daily activities. The task of segmenting these images into distinct activities or events (and identifying the activity being undertaken) is a popular research task.

3. The quantified self movement; where the body is equipped with wearable sensors to record exercise performance and physiological data; has become highly

[2] Banks computes that he will acquire 200,000 photos in his lifetime (Banks, 2011, p. 3)

[3] The referenced article is actually a definition of life-caching, a related term. There are mixed views on whether lifelogging and life-caching have the same meaning – Wikipedia suggests they are slightly different, in other contexts they appear to be synonymous. The term lifelogging is used in this thesis because it is the more popular term.

popular, as part of a wider trend in wearable computing.

Note that in these cases, digital information is being recorded, and it is not necessarily one's entire life that is being captured, so perhaps a more precise definition would be: 'lifelogging is any practice to systematically capture and store digital information about one's own life for later display'.

Bell and Gemmell (2009) go further, discussing 'e-memory'; their vision is to be applauded, but a little over hyped: technology is yet to biologically integrate with the human memory. Having a computer store your biography might help your brain to remember it (even if you think you have forgotten), but it is ultimately your brain that holds the memory, not the computer. In this respect, the '5 Rs' model (Sellen and Whittaker, 2010) is a more tractable framework for understanding the purposes of these systems.

To be precise about the definition, only in the comprehensive systems is an attempt made to capture one's entire life – digitising all available content, the other systems typically try to capture a single kind of information; such systems could be used to contribute data to larger, more comprehensive systems. Systems for the management and visualisation of collections of digital photographs pre-date the term by decades, and are not generally associated with the label (but associated literature is extremely useful, and is discussed extensively in the next section).

The quantified self movement is part of a wider trend towards wearable computing, explained by Scoble (2012): "Wearable computers and sensors like the *Nike FuelBand*, *FitBit*, *UP Jawbone* and soon the *Google Glasses*." Personal heart monitors and pedometers are obviously nothing new, it is integration with social media which is the new development. The *Autographer*[4] is perhaps the most recent attempt to launch wearable cameras into ubiquity. Arguably, sport has seen greater adoption of wearable computing, *GoPro*[5], popularly worn for skiing, and adventure sports[6], yet some in the media have questioned whether the a more widespread wearable computing revolution will ever happen: "we already have general purpose smartphones...there's no killer app, which translates into significantly greater convenience" (Orlowski, 2014).

These practices remain niche; only a handful of individuals have developed their own comprehensive systems. Wearable cameras are yet to achieve significant use outside academic studies: Gurrin et al. (2013) remarked that the *SenseCam* has "not yet in widespread use at the population level", and investigated the use of smart phones as alternatives to the *SenseCam* (in the context of healthcare), finding that many of the features could be replicated, albeit with some issues. Wearable sensors have entered the mainstream, but are far from ubiquity.

Having introduced the concept of lifelogging, the following section will explore the relationship between lifelogging and online social networks. This thesis proposes a system to facilitate a lifelogging user experience built on online social networking data (often captured through smart phones) – the next section outlines the challenges and opportunities for doing so.

[4] http://www.autographer.com

[5] http://gopro.com/

[6] https://www.youtube.com/results?search_query=gopro

2.3 Trends, OSNs and Challenges

In this section, the trend away from desktop personal media collections towards online social networks (OSNs) is discussed, and how it greatly increases the amount of personal data available for reminiscence, whilst introducing challenges to gather, store and display the data.

Instead of focusing on new means of capturing data directly with specialist apps and wearable computing, existing social networking data represents a huge opportunity for lifelogging. By collecting, organising, and displaying the user's footprint of social networking data appropriately, the many millions of users of social networking services will have access to a lifelogging experience. Working towards that goal, the concept of lifelogging can be used as a lens through which to evaluate OSNs, and critique their current user experience, establishing whether all this new data is valuable from a reminiscence perspective, and discussing some wider issues.

2.3.1 Greater Volume

> Let's imagine, for a moment, that the year is 2019, and we have dragged ourselves into the future with a minimum of apocalypse. Picture yourself sitting in front of your news-o-scope (my patent is pending) when up pops word that a website you were really into a decade ago is shutting down. "Facebook!" you exclaim. "I remember Facebook! I posted 250,000 pictures to Facebook. My lost youth!" (Tossell, 2009).

The transition to digital photography, and then to online social networks (where other forms of data are shared), has greatly increased the volume of personal data people own: the advent of digital photography had made it easier to take photographs, online social networks had a big impact as well, making it easier to share them. According to Facebook user data, in 2011, each active user uploaded an average of 7.5 photos per month (Schenkel and Spaniol, 2011). It is not just the user's own photos that form part of their life-story, many people will be 'tagged' in photos that their friends' have uploaded, meaning there is a arguably a much larger set of photos to preserve than before.

As social networks were becoming popular towards the end of the last decade, the ubiquity of phones equipped with cameras made photo sharing easier than ever, changing the way people socialise: "with the advent of mobile technologies including camera phones, people increasingly use them to facilitate their social life outside the work environment" (Stelmaszewska et al., 2008, p. 1). Now, smart phones are available, allowing people to share photos directly from their phone, uploading them to the social networks using apps, again, contributing to a greater volume of content. Facebook now have over 1.19 billion active mobile users[7], at the point of their IPO 300 million new photos were being shared on the network every day (Facebook, 2012b).

This increase in volume is further magnified by the fact that content is shared amongst different users of online social networks. Previously, each of us took our own photographs, on a social network, for every photo the user uploads, there are many more

[7] As of March 2015. `http://newsroom.fb.com/company-info/`

which other people have uploaded for the same event, where perhaps the user is tagged in the photo.

2.3.2 Multiple Data Sources: Networks and Devices

Users tend to have multiple social network accounts for different services, even if they are only actively using a subset. This creates various challenges, as users move between networks, and stop using some services, this data can become lost, or passwords forgotten, email addresses change – content might be deleted without the user even realising. Users may post content to multiple platforms, presenting a de-duplication task.

2.3.3 Multiple Kinds of Data

This trend towards more diverse data about ourselves has continued as more of the so-called lifelogging technologies achieve predicted commercial adoption: "Just two years into our ten year prediction, business formation has been nothing short of amazing" (Bell and Gemmell, 2009, p. 11). Recent years have seen the rise of location-based social networks, such as FourSquare, recording 'check-ins', and apps to record physical exercise, such as Endomondo which now has 12 million users[8] (Endomondo, 2012).

2.3.4 The Question of Value

Physical mementos can have obvious value: Banks speculates on the implied significance (at least to their original owner) of the objects' survival through his grandfather's spring-cleaning, "these photos survived all those decision points and made it to me. That seems important." (Banks, 2011, p. 5)[9].

This chapter discusses photos extensively, because of their long history as a means of reminiscence, but all kinds of digital documents are available in social networks. It is not only photos that are important to preserve, anyone that has used these services will know that they contain a "tremendous amount of content such as text, images, audio or video" (Aggarwal, 2011, p. 4). Facebook alone has a huge variety of different objects that users can create, with similar range on other networks. The number of classes of objects is only set to increase, with innovations such as Facebook's shared albums (Facebook, 2014a).

It is important to consider whether this content is worth preserving. Authors have advised archiving their Tweets and status messages: "In the short term we may not see their value but in the long term they may offer a valuable resource not just for reminiscing but also for telling the day-to-day story of our lives to our offspring" (Banks, 2011, p. 7)

[8] `http://blog.endomondo.com/about/`, the number of active users might be considerably lower.

[9] The significance is not always clear to the bequeathed: Banks "met a person who had been left a box of rocks by his grandfather. Why he'd received this bequest was a mystery" (Banks, 2011, p. 0).

There is good reason to be sceptical; whether the preservation of data associated with a users 'friends' has any value to the user is an open question, as is the strength of the underlying personal relationships: "We show that interaction activity on Facebook is significantly skewed towards a small portion of each user's social links... [this] casts doubt on the assumption that all social links imply equally meaningful friend relationships." (Wilson et al., 2009, p. 13) (they go on to propose an 'interaction graph' which more accurately models the relationships).

Indeed, even for our own content, there are mixed views on its value: Van House (2007, p. 9), in interviews of Flickr users from UC Berkeley, USA found that Flickr photos were considered "transitory, ephemeral, 'throwaway', a stream, not an archive. Their primary interest was in recent images, their own and their contacts'. Images are archived elsewhere.".

Sinn and Syn (2014) surveyed people about how they perceive their use of Facebook, in documenting their daily lives. They found that participants "did not agree that they used Facebook to document their everyday lives", and were "not nostaglic about old posts", and that many users see Facebook as a tool for networking rather than documenting. However: "Many participants indicated that they cared about saving content posted on Facebook".

The long history of systems and research into the presentation of photos means that they have clear value, whether check-ins, status messages and all the other kinds of data have definitive value remains to be seen, but there are various commercial and academic efforts have sought to archive develop means of archiving this material. Another motivation for preserving non-photographic content is that it preserves the original context and story surrounding the treasured photographs: "All that matters about it really, is the story that's tied to it. Once the story dies, so, to some extent does the value of the item" (Banks, 2011, p. 14).

Great value is attached to even the most ephemeral data belonging to a deceased relative – evidence of sentimental value in content was found by Massimi when interviewing about the use of technology by the bereaved "respondents found themselves combing through hard drives full of the deceased's files, trying to find important pieces of information", and that (39%) valued 'journals or written works', with some interviewees eulogising the deceased on memorial websites or Facebook. (Massimi and Baecker, 2010, p. 6).

The value (or lack thereof) that individuals associate with their own content may only become clear over time, and the value of reminiscence is difficult to estimate until it has been implemented. A simple argument would be that if the content is worth sharing, it is likely to be worth preserving.

2.3.5 Challenge 1: Handling the Volume: Focus on "Now"

The ever increasing volume of information being published creates a challenge for the user experience: firstly, it greatly increases the amount of content that needs to be organised and displayed in the social network, for reminiscence purposes or otherwise. The user experience of social networks tend to focus on the 'now', the recent content.

Hence, users are incentivised to always add new content (and thus commercial value) to the services (Gerlitz, 2012). Conversely, from a lifelogging perspective of looking back at one's life, the user experience is poor: "While users perform increasing amounts of activities and connections, platforms offer only limited possibilities to make sense of one's own data and often turn activities into fleeting objects on streams and promote immediate interaction without organised access to the past" (Gerlitz, 2012, p. 1).

Facebook's solution to the volume challenge, the Facebook Timeline, is discussed later in the chapter. However; instead of the timeline, users are presented with the 'newsfeed' when logging in, an example of an infrastructure "in which users are asked to continually add value by building content, creating connections." (Gerlitz, 2012, p. 2), leading to a cyclic system where "platforms not only encourage users to respond now, like now, Tweet now and share now, but also to like again, share again and Tweet again" (Gerlitz, 2012, p. 3).

The work of Sinn and Syn (2014) claims that people do not use Facebook in order to document their lives, but instead to share content with their networks. Yet, the apparent contradiction is that users see value in safeguarding their content, whilst not using Facebook for reminiscence. This disconnect hints at a user experience which is ill-suited or even discourages the activity, or perhaps that users wish to preserve the content for future reminiscence activities.

2.3.6 Challenge 2: More Networks, Devices, Applications and Users

Facility for bulk-downloading of content, such as via an API, is provided at the whim of the provider, so may not be available. Where access is available, each network has its own idiosyncrasies. Data is structured in different ways, with different information available, Facebook for example remove EXIF metadata from photos, while Flickr does not. Twitter only allows access to the most recent 3,200 Tweets (Twitter, 2015). Although such APIs are typically a collection of HTTP endpoints, beyond that there is little consistency; data may be published in various data formats, with XML and JSON being popular, with different mechanisms for obtaining access[10].

Users will each use a variety of different applications and devices can publish information to social networks; in doing so, they can leave clues in the metadata, for example, specific Facebook applications will post photos to particular named photo albums, or have particular wording in messages, which can be detected and used to infer additional knowledge about the content, effectively creating new layers of metadata within the data.

The multitude of networks creates challenges for collecting and organising the data related to an individual. Different people may upload content related to the same event to different social networks. Even if the social network is the same, if it is uploaded separately there may not be any metadata which directly and unambiguously links the content. Innovations like Facebook's shared albums can obviate this issue somewhat, but does little for the millions of photographs in the backlog; in these cases, the matching will need to be inferred from clues left in the content and its metadata. Other studies have made progress in addressing the fragmentation of the

[10] The OAuth standard, is implemented slightly differently on many of the social networks.

social networking footprint. Steiner et al. (2012) identified this problem: "for one and the same event, the event-related user-generated data may be shared on a plethora of social networks."

2.3.7 Challenge 3: Risk of Network Closure & Data Loss:

The survival of any commercial service, especially in the fast-moving domain of social media, is far from guaranteed. Many authors have noted the termination of the *GeoCities* service, and the emigration from *MySpace* (Tossell, 2009; Bell and Gemmell, 2009; Banks, 2011). Over the course of our lifetime, we will use many different services, and archiving our data from all these different services is a key challenge in preserving our social media footprint (Garfinkel and Cox, 2009).

Secondly, it raises issues related to ownership of content, with complications if people wish to leave social networks, or when a relationship ends (Sas and Whittaker, 2013). Other issues potential causes of data loss have been noted; suppose there are some photos of you, or some significant event in your life, 'owned' by someone else's user account that may be deleted if that user leaves the network (Robertson, 2014).

Ownership is not just about ensuring data is not lost; control of deletion of data is another consideration: Cheng (2012) investigated how long it took photos to cease to be available from Facebook's servers after 'deleting' the content from various social networks, Twitter and Flickr "took mere seconds", MySpace took "several months" and Facebook took "more than a year", with some readers reporting a delay of "three years or more", with discussion various situations where deleted data has "reappeared".

Unhelpfully, services aggressively seek to control our content: Twitter requires that content is not "rendered with non-Twitter content (e.g., comments, updates from other networks)" (Twitter, 2014), whilst in Facebook, 'bulk export' to another social network is forbidden (Facebook, 2014b).

2.3.8 Solutions: A Better User Experience

In the proceeding sections, a variety of classes of systems are discussed which in some way overcome the challenges raised in the previous section, outlining the shortcomings of some solutions. Having discussed these systems, and their shortcomings, the discussion moves to event-centric means of presenting personal data. The sub-section begins with the history of older systems for the organisation of personal photos, key to the development of the event, having been influential on more recent systems for organising photos from sources such as online social networks, and from wearable cameras such as the *SenseCam*. This section demonstrates the strong body of literature supporting the proposition that events are an effective way to present large quantities of personal data, whilst identifying areas of work with useful techniques for detecting events in such data, hence justifying the decision to adopt this approach for the organisation of social networking data, as well as leading into a review of the event detection techniques themselves in the following chapter.

2.3.9 Password and Memorial Services

Perhaps the simplest approach to dealing with digital death is to simply give friends and relatives access to your data when you die: by leaving a copy of your access credentials to your key services. This being a relitively simple service to implement, it is offered by a large number of companies. The service is often bundled with other features, such as sending out a last email, and planning funeral services; examples include: Death Switch[11], PasswordBox[12].

As these services only store credentials, they do little to actually secure the data – content is still vulnerable to deletion, and specific companies have their own policies regarding account closure in the event of death.

Some social networks provide special features to handle death; Facebook has a memorial mode (Facebook, 2014c). This is not without issues, because new friendships cannot be created, and privacy settings are locked, family members can be excluded from the content. There are also a number of websites which offer a facility to create an online memorial to the deceased, a form of content curation[13].

2.3.10 Web Archiving and Blogs

Since online social networks are often websites, or offer websites as part of their overall service, the preservation of social media content can be considered to be a web archiving task – a task with a long history and associated body of research literature. Web archival offers another perspective on the issue of social media preservation and reminiscence.

Banos et al. (2012) note the popularity of blogging, noting that "there is still no effective solution for ubiquitous semantic weblog archiving, digital preservation, management and dissemination", and that "to the best of our knowledge, no current Web Archiving effort has ever developed a strategy for effective preservation and meaningful usage of Social Media". The remainder of their paper is effectively a functional specification for blog preservation.

Archival of web content comes with its own unique challenges and issues, especially so in the context of social media – is it the content that is being preserved, or the complete experience – and how to divide the two. McCown and Nelson (2009) discuss the difficulties of archiving one's social networking data, discussing the possibility of Facebook shutting down, noting the issue of preserving the experience of the site as well the content itself, and the additional complexity this brings. After discussing various methods of archiving Facebook, they develop a browser extension for Firefox which preserves complete page snapshots automatically, overcoming any complications of privacy restrictions. Such an approach is an improvement on simply storing the password, but may not be fully automated; in any case, the original experience is being preserved, without an obvious means of integrating data from multiple platforms.

[11] http://www.deathswitch.com/

[12] https://www.passwordbox.com/

[13] A more extensive list of these services can be found at: http://www.thedigitalbeyond.com/online-services-list/

Larger projects have attempted to tackle digital preservation at an organisational or community level (see Edelstein et al., 2011 for a review), rather than for an individual: LiWA – Living Web Archives – web content capture, and ARCOMEN, which "is about memory institutions like archives, museums and libraries" and aims to "transform memories into collective memories", uses events.

2.3.11 Chronological Document Timelines: Facebook Timeline and Twitter

Online social networks have developed some of their own means of letting users experience the past by viewing historic data. The Twitter timeline is perhaps the canonical example of such a system: a chronological list of items.

The Facebook 'wall' was originally presented in a similar way to Twitter, with a simple chronological ordering of items. A attempt to overcome this issue on the part of Facebook, has been the timeline feature (Facebook, 2012a), which is a little more elaborate. Whilst still a list of every item of content in the persons profile, navigation is possible according to month and year, content such as photos is presented alongside status messages, and semantic events such as birth, and periods of employment are displayed (when they are explicitly entered into the profile), giving a richer experience, without apparent need for data-mining.

However, surveys have found the public reaction to be overwhelming negative (Choney, 2012); it seems the problem was not that the new feature did not give an overview of the user's copious data – it was that it did. The fact people could now easily browse this information; meant that it became clear how much data they had shared on the service, one user Tweeting: "I ditched Facebook the first time I tried timeline, scrolled down to the bottom and had a sobering social media moment" (Cluley, 2012).

Regardless of how well-designed the Facebook user experience becomes, the experience is always going to be fragmentary: content from other networks cannot be integrated, and ownership of the data (and so the choice on how to present it) lies with the host network, and it remains a hassle for the user to safeguard against data loss. Croll (2015) predicts a single, aggregated timeline including not just our social media data but financial, health: all the digital data in our lives, with ownership and control being a key issue.

2.3.12 Commercial OSN Backup Services

To tackle the risk of data loss, a number of companies offer services to back up the social networking data. Many of these services are able to combine data from multiple platforms and allow the data to be viewed as a single collection. Superficially, these services can appear to overcome two of the key challenges: the data is secured from accidental loss, and the footprint is no longer fragmented, it can be experienced as a combined whole. On closer inspection, these issues are only partially solved.

Memolane offered such a service[14] – content could be imported from a variety of

[14] The service was shutdown in 2013. `http://tinyurl.com/b7mjkgs`

different social networks and is presented together (see figure 2.2).

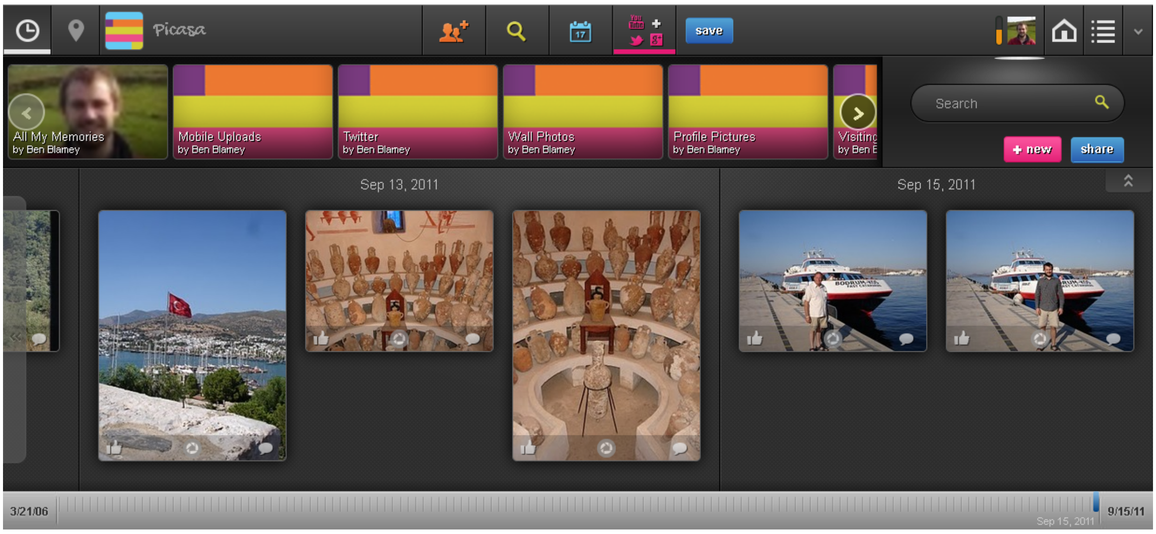


Figure 2.2: Viewing the author's data in Memolane.com, (November 2012).

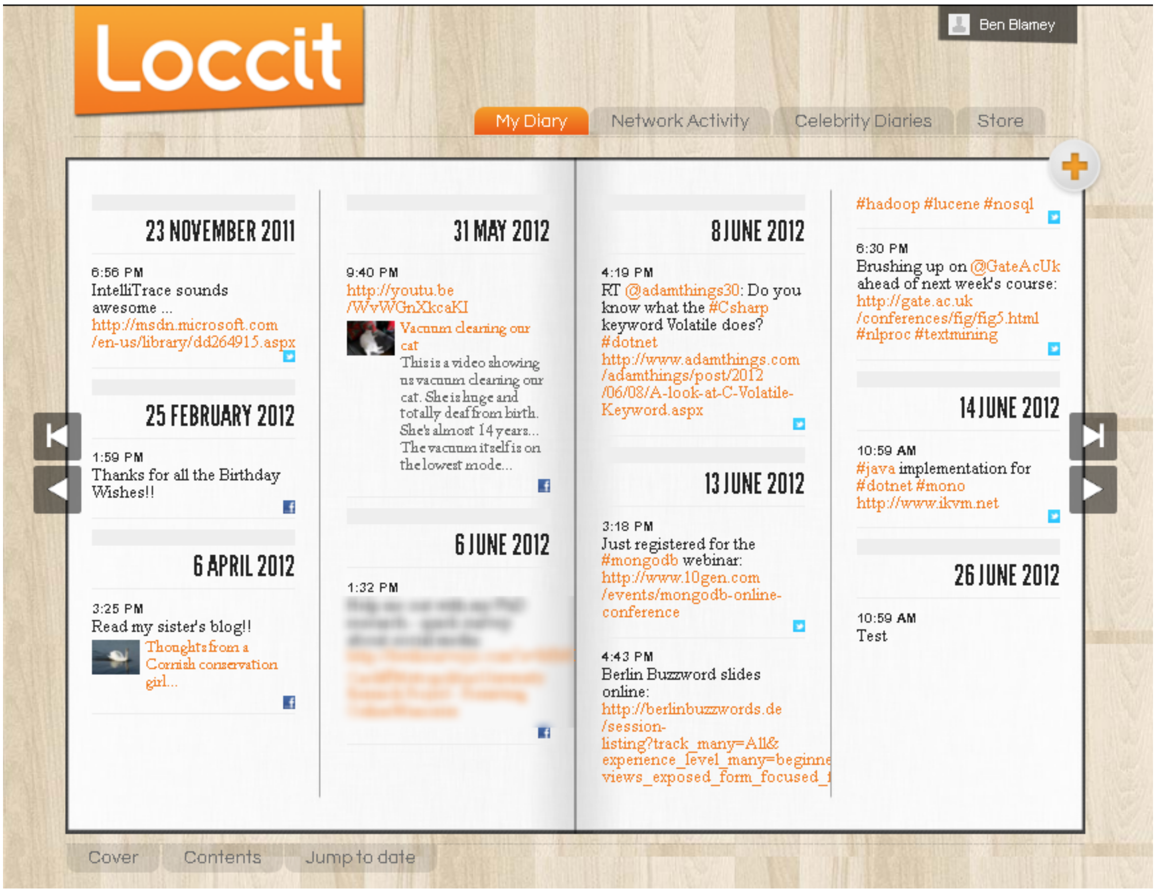
Loccit.com was a similar service, packed with features, presenting the gathered content in a novel 'diary' view, with facility for adding private 'reflections', as well as the ability to create a variety of printed goods from the imported social media content (see figure 2.3).

Ironically, these online services, intended to aggregate this data can themselves be closed, as was the case with both *Loccit* and *Memolane*. Indeed, such tech startups may indeed be more vulnerable to closure than larger social networks. SocialSafe[15], offers an alternative approach – licensing desktop software to allow users to download data to their own machine. Whatever happens to the company, the software remains installed on the user's harddrive, along with the data, obviating much of the risk.

Platforms such as SocialSafe go a long way in solving the issue of safeguarding data loss; a solution in the private cloud would be even better. Users may have the ability to manually group content, similar to the *Loccit* 'chapters', and the 'lanes' in *Memolane* – intended for different strands of the user's life. Or, existing structures can be imported wholesale from the originating network, as in the case of photo albums in SocialSafe. In *Memolane*, automatically created 'lanes' are simply based on the underlying organisation of the data in the social networks, such as wall photos from Facebook. Naturally, there is facility for faceted search and filtering, but the user interface is actually primitive when compared against an established class of systems.

In contrast, much of the existing systems for organising personal photos are based on organising content into 'events'. The technical challenge undertaken in this thesis is to develop a system to organise content into events, in pursuit of established user interface design practice, but to do so on a complete social networking footprint, rather than merely a collection of photos. The automated organisation of the social networking documents into a series of events, should enable better visualisation, search and navigation of the aggregated data.

[15] `http://www.socialsafe.net/`



23 NOVEMBER 2011

6:56 PM
IntelliTrace sounds awesome ... http://msdn.microsoft.com/en-us/library/dd264915.aspx

25 FEBRUARY 2012

1:59 PM
Thanks for all the Birthday Wishes!!

6 APRIL 2012

3:25 PM
Read my sister's blog!!
Thoughts from a Cornish conservation girl...

31 MAY 2012

9:40 PM
http://youtu.be/WvWGnXkcaKI
Vacuum clearing our cat
This is a video showing us vacuum clearing our cat. She is huge and totally deaf from birth. She's almost 14 years... The vacuum itself is on the lowest mode...

6 JUNE 2012

1:32 PM
[redacted]

8 JUNE 2012

4:19 PM
RT @adamthings30: Do you know what the #Csharp keyword Volatile does? #dotnet http://www.adamthings.com/adamthings/post/2012/06/08/A-look-at-C-Volatile-Keyword.aspx

13 JUNE 2012

3:18 PM
Just registered for the #mongodb webinar: http://www.10gen.com/events/mongodb-online-conference

4:43 PM
Berlin Buzzword slides online: http://berlinbuzzwords.de/session-listing?track_many=All&experience_level_many=beginne views_exposed_form_focused_f

#hadoop #lucene #nosql

6:30 PM
Brushing up on @GateAcUk ahead of next week's course: http://gate.ac.uk/conferences/fig/fig5.html #nlproc #textmining

14 JUNE 2012

10:59 AM
#java implementation for #dotnet #mono http://www.ikvm.net

26 JUNE 2012

10:59 AM
Test

Figure 2.3: Viewing the author's data in Loccit.com, showing the aggregated social media footprint presented as a 'diary' (November 2012).

2.3.13 Event-Based Systems

Early research was quick to acknowledge the importance of the 'event'; over the proceeding subsections, and into the next chapter, the chapter will examine how the concept has been applied repeatedly (if not consistently) in a large number of systems for the organisation of personal data.

2.3.13.1 Desktop Photo Management Systems

The notion of an 'event' has been a key model for photo organisation for more than a decade – early systems were designed to assist users in viewing their personal collections of digital photos. Some of the earliest examples coincide with content-based image retrieval techniques gaining popularity; a selection of examples are reviewed here. This section seeks to establish the motivation for an event-centric approach, technical details of the algorithms involved are deferred until chapter 3.

Searchable image databases have been around a long time e.g. Kuchinsky et al. (1999), describing their 'FotoFile' application cite numerous commercial systems. Their desktop photo organisation system was developed to investigate user interaction with various search and visualisation features (some of which were third party controls), including facial recognition, and a 'hyperbolic tree' where photographs are organised into a hierarchy depending on a chosen metadata facet. The work was amongst the first to highlight the importance of the event: "we can use narrative structure underlying the events captured in photos as a source of their organisation and annotation", yet there is little discussion of how the event itself is represented or conceptualised, or any list of event facets.

Wenyin et al. (2000) propose their *MiAlbum* system, desktop PC software for organising and exploring a personal photo collection. In their work, the 'event' is one of a selection of organisational axes for photographs: 'places, date, time, event' – i.e. the kind of event, rather than a overarching notion comprising these facets. It included early content-based image retrieval techniques for search functionality, with users able to give feedback on search results, in an application of previous work (Lu et al., 2000). Using scanned photographs results in a lack of metadata, Banks's TimeCard system (Banks, 2011, p. 74) (a chronological slideshow of scanned photos) requires users to type in notes relating to the photo[16].

The focus on image content continued in the work of Lim et al. (2003). They explicitly outline their definition of an event: who, what, where, and when. For the 'what' facet i.e. the kind of event, they propose to 'approximate event by visual event'. They outline a complex system, with neural networks for recognising keywords associated with image regions, producing a semantic graph describing the photo, which is then matched with similar photos to assign an overall label for the photo to one of many event kinds in their taxonomy.

Motivation for these techniques is supported by an early comparative investigation into personal management of physical and digital photos found that "[physical] albums are mostly classified by specific events, such as holidays" (Rodden and Wood, 2003, p. 3), with analogous organisation of digital photos into folders. Although not all participants found this necessary: "I download them and I put them in a folder, and I label that with the date that I download it on, and that gives me as much organisation as I want, really." (Rodden and Wood, 2003, p. 3).

By 2007, research on content-based image retrieval techniques has advanced significantly, with a range of techniques and approaches; Suh and Bederson (2007) comprehensively review the existing work, before introducing two of their own: hierarchical events, and clothing-based person recognition – built into their photo-management system: *SAPHARI*, motivated by their previous system – *PhotoMesa* (Bederson, 2001). Through questionnaires, the authors found that "participants were significantly positive with semi-automatic annotation interfaces compared to manual annotation interfaces".

In the same year, Jain and Westerman recognised the significance of the event concept, especially in multimedia applications, and proposed a theoretical framework for standardising its representation:

> "A common model would also provide a glue for integrating and syndicating events and media from largely isolated multimedia applications and data sources, making possible novel cross-application and cross-data source multimedia services" (Westermann and Jain, 2007)

They also noted that "from an academic perspective, a common multimedia event model provides a reference framework for classifying, comparing, and evaluating event

[16] Notes on the back of old family photographs, can thought of as a precursor to EXIF.

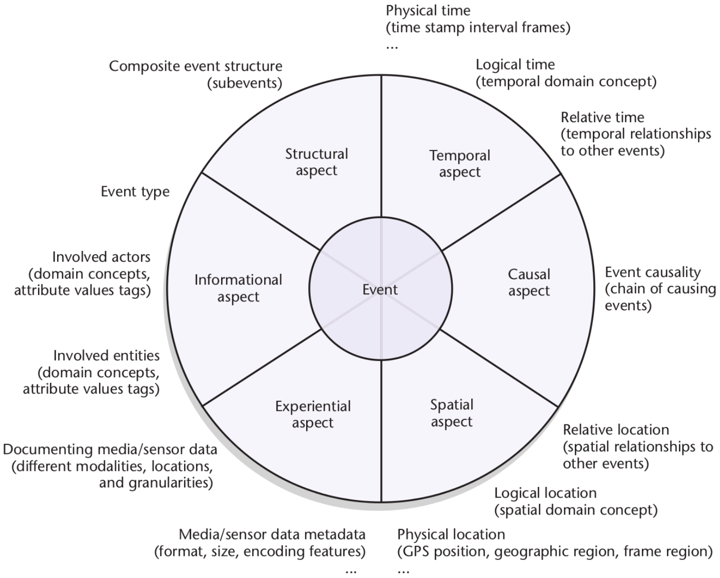


Figure 2.4: Event aspects, reproduced from Westermann and Jain, 2007

support in different applications". They give examples of "life logs" and "event-centric media managers", suggesting a common language – the common event model – for processing events in different contexts, stopping short of doing so, they instead problematise the concept, discussing a range of issues, effectively a specification that such a model should satisfy. For example, an event can be 'discrete' (i.e. an instant) or 'continuous' , such as a football match. Their theoretical framework for describing events contains six various aspects (see figure 2.4) – they use the example of a football match.

- Temporal – an absolute timestamp/duration, or whether the event is described to other events, e.g. "the brawl occurred during half time".
- Spatial – the latitude and longitude of the football pitch.
- 'Informationational' – i.e. who was involved, e.g. players shirt numbers, goals scored, tagging of entities involved. Facebook tags (people) would be a modern example of this.
- Experiential – the associated items of media.
- Structure – recognition that events can contain sub-events, and the desirability of being able to aggregating both high and low-level metadata from different levels in the event.
- Casual relationships between events.
- Association – such as between different event systems, such as different views of the same event in CCTV – the authors see domain knowledge, formalised through ontologies and the like, as key to this: "It is not enough to implement associations [as] pointers or foreign key attributes".
- Uncertainty – given the imprecision of event detection, their imagined common event model must capture uncertainty: "event rule and event-notification languages must incorporate probabilistic methods to handle event detection and inference in the presence of uncertainty". They go on to discuss issues of extensibility and knowledge awareness.

Some issues were overlooked: the authors seem to focus on public events, with their example of the football match (the importance of more private and personal events is neglected) especially when the paper cites work relating to personal photos. There can be private photos taken at a public event, as well as media for a private event which is later made public. Social media raises some new issues, such as the blurring boundary between the event and its representation in media, user interaction with social media becomes an intrinsic part of the event.

Subsequent research has attempted to fomalise the event representation, many of them using Resource Description Framework (RDF[17]). These representations tend to be domain specific, examples include:

- Raimond and Abdallah's Event Ontology (for music events), figure 2.5 [18] (used by `http://musicontology.com`)

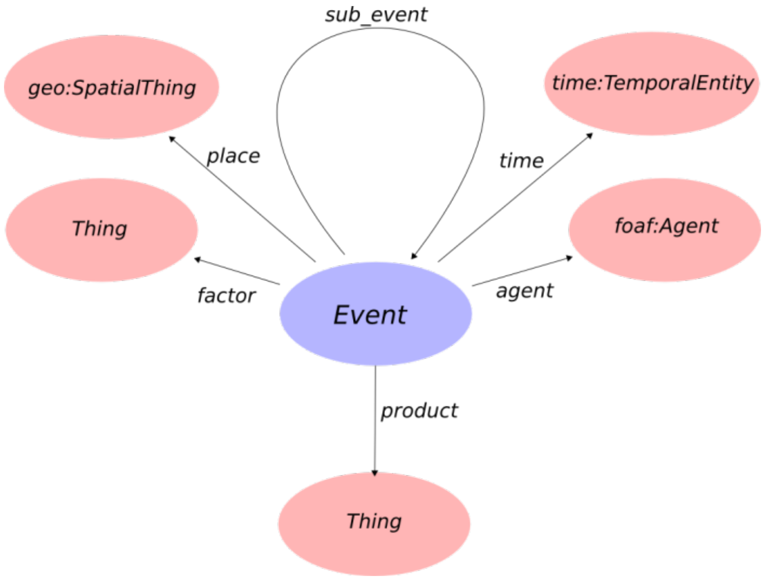


Figure 2.5: The Event Model, reproduced from `http://motools.sourceforge.net/event/event.html`.

- The EventCube ontology – for *SenseCam* data (Wang, 2012).
- Schema.org's Event `https://schema.org/Event` (intended for a "concert, lecture, or festival").

The remainder of this chapter will explain how many systems have failed to acknowledge many of the more nuanced aspects of Jain and Westerman's proposed common event model, with the more simplistic who/what/where/when model gaining more popularity.

2.3.13.2 Printed Photos and Scrapbooking

Another relevant literature topic concerns products created by printing photographs.

[17] `http://www.w3.org/RDF/`

[18] `http://motools.sourceforge.net/event/event.html`

Scrapbooks and scrapbooking have been a popular means of preserving photographs and other ephemera for a long time (Buckler, 2006), and the production of physical photo books from digital photos, as well a virtual scrapbooks are natural progressions of the practice. Rabbath notes the desirable characteristics of physical, printed products: "[physical photobooks] have always been the preferred way to reliably preserve the memory of important moments...in contrast to the inherently dynamic nature of the social community sites...[content] could be deleted or the [site] itself could vanish from the internet" (Rabbath et al., 2010, p. 2).

Many of the challenges for designing a software user experience for photographs and other documents, are pertinent to the design of these printed products: a decision needs to be made regarding what content to include, how it is to be structured and organised over multiple pages in a suitable way, and finally, the selection of an aesthetically pleasing layout for the content. Making these design choices can be laborious; consequently, several authors have investigated the automation of these tasks as a research problem. Fageth, discusses various means of automating the selection of images for printing, as well as the inclusion of relevant third party images, from sources such as Flickr (Reiner et al., 2008). Additionally by geo-coding the text from the user-specified product summary (e.g. 'Visit to Oldenburg') to find the geographic location, copy from the associated Wikipedia page can be included. Rother et al. (Rother et al., 2006), from Microsoft Research, presented their system *AutoCollage*, which is able to build a collage of photos with blending and intelligent cropping of photos around key features such as faces.

However, it is the use of events that is of interest to this thesis – events are crucial to image selection and layout – the scope of the printed product might be a single event, such as a holiday, or alternatively, each event may have a dedicated page or section of a photo book.

Rabbath, Sandhaus and Boll have published extensively on these research problems (2010, 2011, 2012, 2013), investigating the whole pipeline of photo book creation, everything from approaches to handling metadata (Boll et al., 2007), a system analysis yielding a set of canonical processes and phases in photo book creation (Sandhaus et al., 2008); in the same way this thesis seeks to standardise a system for event detection in social media. They have analysed how people use photos of events in photobooks (Sandhaus and Boll, 2009), The group's 2011 review paper (Sandhaus and Boll, 2011) focuses on data-mining online photo collections, and how social context can yield additional semantics.

Organising our digital mementos can be slow: "making a photo book as a special gift to your beloved can be very time-consuming" (Boll et al., 2007, p. 1); indeed many of us never get around to it. Automating these processes is the key goal underling much of the work in this chapter. More recently, experiments to detect events in user's Facebook photos (Rabbath et al., 2012) – the work is the state of the art for detecting events in Facebook photos is reviewed in the next chapter.

2.3.13.3 For Photos (Mobile)

With the ubiquity of mobile devices, and especially camera-equipped smart phones, there is a growing need to display large collections of personal photos on the phone. At their Worldwide Developers Conference (WWDC) 2013 keynote, Apple announced new features for the 'Photos' app bundled with their iOS operating system[19]. In the presentation[20], Craig Federighi (Apple's Senior Vice President, Software Engineering) explained how most users have an "endless unorganised stream on their camera roll", and demonstrated new functionality in their app which can organise photos into "moments". These moments (effectively events) can themselves be grouped into collections, or simply by year. There is an allusion that the algorithm is based exclusively on when and where facets, and is seems to be designed for photos that were actually taken with the phone. Although the technology appears to be quite simple, it is evidence for the challenge of volume, and re-enforces the event model as the de facto standard for displaying personal content.

2.3.13.4 For *SenseCam* Data

Other authors from the discipline of lifelogging have identified a need for modelling event semantics for effective user experience of personal history data, such as that originating from the *SenseCam* [21] wearable camera. Anecdotally, the popularity of wearable cameras as a research task within lifelogging is such that the topic has become almost synonymous with the term. The *SenseCam*, worn throughout the day takes photos automatically (based on strategies such as time lapse, change in light levels, and the like), resulting in a stream of photographs for each day. To allow navigation of this enormous volume of content requires automated organisation of the photos. Again, the event has been at the centre of the user experience:

> "An understanding of the 'what' or the semantics of an event would be invaluable within the search process and would empower a user to rapidly locate relevant content." (Byrne et al., 2010, p. 2).

There has been extensive work on developing algorithms and techniques to automatically segment these photo streams into events, with the task made easier by uniform metadata including timestamps and GPS location information for every photo. State-of-the-art systems, such as that implemented by Wang (2012), are able to achieve impressive results. The equivalent task for OSN data, the focus of this thesis, is a little different. Instead of a stream of photos all with uniform metadata, the OSN data can be patchy, and differently heterogeneous (section 1.6) – documents are of various kinds, with various fragments of metadata available, creating a number of challenges, as noted by Dietze et al. (2012) "user-generated content and social media...characterized by a high degree of diversity, heavily varying quality and heterogeneity". Although many of these challenges are not pertinent to *SenseCam* data, some of this work is reviewed in the next chapter.

2.3.13.5 Single Data Kind – OSN

The sharing of photographs online has arguably been crucial to the growth in popularity of OSNs. This trend has introduced new challenges for the archive, organisation

[19] `http://www.apple.com/ios/whats-new/#photos`

[20] `http://www.youtube.com/watch?v=SRmjUzcpLO0`

[21] `http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/`

and display of this content, but has meant that vast datasets of photos, often annotated, are available for large-scale machine-learning. With all this extra information available, it became possible to overcome the limits of what information could be extracted in the days of digital photography – using tags associated with images on (for example) Flickr, for unsupervised training of software for image content detection, at large scale. However, there is reason not be over-confident: many of these datasets are based on 'public' events, where many different people have all attended the same event and post content online, resulting in many photos associated with the event. Technical details of these studies are discussed in the next chapter.

2.3.13.6 ...for Multiple Data Kind/Source

Much of the discussion so far has focused exclusively on photos, and their organisation into events. Other authors have confronted the task of handling a mixture of different kinds of data, often on the mobile platform.

That the mobile phone is an ideal lifelogging platform: Rhee et al. (2010) present their *Life Diary* app to aggregate various activity logs on the phone into a single timeline. They note that the life diary serves a useful function aside from offering a reminiscence experience: returning calls, sending digital content to friends, etc.

In response to the perceived deficiencies, some authors have to better integrate the fragmented user experience of OSNs, and introduce more lifelogging functionality into the user interfaces of these systems. Such a study was conducted by Rhee et al. (2010), developing a prototype GUI for a multi-platform OSN which they call *LifeDiary*, following a user-centric design process. Life-caching (a synonym of lifelogging) develops into an important aspect of their design, with opportunities on the mobile platform: noting the ubiquity of the mobile phone, they explain how the platform presents new opportunities for lifelogging:

> "A mobile phone becomes a useful gadget to collect, record, and store daily events with the help of affordable storage space. The present study argues that a mobile phone may help organise special occasions by collecting daily experience"

They see "[the emergence of] recording and sharing events in daily life as a mega-trend", where people "fulfil their needs by collecting, editing, and sharing contents via mobile devices, PCs, and the web". Their motivation is that mobile web services did not meet user expectations, and they note a lack of interaction between mobile and web.

Initially, for a group of 32 people, they conduct "shadowing", where users' interactions with existing OSNs are observed. The shadowing process is claimed to gain "rich, in-depth, and accurate insights into how people use products, processes, and procedures". Together with semi-structured interviews, a number of key themes emerge, which are developed into a number of *personas* (a concept in user-centric design). One of these, the "product wizard", representing users who "put a value on content management" and "love to collect games or music by interacting with friends or attending events and need simple ways to discover and organise content". This persona forms the basis for the development of the lifelogging related functionality of their system. Wireframes for mobile, PC, and website platforms are developed. User "diaries" are the core feature: a private diary for friends and family, and a public diary. Their design reflects the importance of the mobile platform to life-caching, nevertheless, the content will be seamlessly shared and synchronised between the different platforms. User feedback on the prototypes is positive.

In their work, an event is a single document e.g. a phone call, a photo; they do not group the data into according underlying real world event, which would ultimately result in a large volume of documents. Over time, this unstructured volume would not be conducive to a reminiscence experience.

Storing and presenting this additional data, of various kinds, presents a challenge, but also an opportunity. This richer information can help us to add meaning and build a more detailed picture of the user's life, as Banks explains:

> "It is not just the details embedded in a modern digital image that provide new forms of context that we didn't have in the past. Much of the data we create every day, from the mundane to the deeply personal, is stamped with a creation time. Emails, text messages, blog posts, and Facebook and Twitter status messages are logged all the time. All of this detail, brought together in time, can create a supporting context for a moment...Each item provides reinforcing detail that helps flesh out what took place at that moment." (Banks, 2011, p. 5).

The author is not aware of any such system which is yet capable of clustering a dataset containing a variety of different kinds of documents into event clusters.

2.3.13.7 Other Projects and Approaches: Presentation of Life Stories

Similarly, presenting content around 'stories' can is known to help the user navigate digital content. Alahmari, whilst studying the psychology of event timeline user-interfaces, interviewed users about marking 'temporal landmarks' including "birthdays, weddings, relationships, public vacations, and news" (Alahmari, 2012, p. 9)). Based on responses, he concludes that "adding personal pictures and personal temporal landmarks on the timeline is one of the best keys that help user to remember temporal information. [sic]" (Alahmari, 2012, p. 1).

Other systems are noteworthy for their novelty; Cho et al. (2007) developed a system called *AniDiary* which summarises a lifelog as a cartoon strip, based on a variety of data collected by a mobile phone app (e.g. GPS information, photos, call logs) – ComicDiary (Sumi et al., 2002) is a similar system published a few years earlier.

2.3.13.8 Beyond the Event

Gonzaga explains that "identity theory assumes that multiple roles can make up the self, which can impact role-related behaviour" (Melcombe, 2011, p. 8), and interviewed women, known to be make up more than half the users of most social networks (Cashmore, 2009), about the representation of their identifies in social media. She found that women typically had presented a ‘social’ persona on Facebook during college years, but since graduation, they “had to transform their identity to fit a professional role, thereby changing various information on their profile, taking down certain information, and un-tagging pictures that could be considered inappropriate for the professional world”. An earlier study attempted to categorise people according to the set of identities they were presenting online (DiMicco and Millen, 2007). In a qualitative study (Zhao et al., 2013), working ‘Hogan’s exhibition approach’ investigated the way we interact with our personal archive on Facebook, finding tensions between public/private ‘regions’.

Earlier work supports these ideas, Belli (1998) gives an overview of theories about the structure of biographical memory, which make good inspiration for what is arguably a system of artificial memory – such as the interplay between events and themes – such as ‘relationship’ and ‘work’ in the theories of Conway (1999).

This is clearly a complex issue, outside the scope of this work, but the opportunity for this research is that if digital content can be organised into roles, it may prove to be a useful means of navigation, and perhaps inform the design of privacy settings in the future.

2.4 Summary

This chapter has presented first of two literature reviews, discussing the context of the study, whilst deferring discussion of technical details until the reader is aware of the scope of existing work which is relevant to the study. It began with definition of lifelogging, and discussion of reminiscence, and our history of preserving photos, and noted issues of thanatosensitivity (the importance of death in user interface design).

It looked at the migration towards online social network, and how this increases the amount and heterogeneity of our data, as well as its fragmentation across various services. Next, the precise challenges created or exacerbated by these trends were outlined, before moving on to a range of services and research areas that can be considered solutions or responses to one or more of these challenges. These solutions are critically reviewed, shortcomings are pointed out, in that many solutions address only a subset of the challenges (the password storage services), create new challenges (as in the case of online backup services), or simply offer room for improvement, as in the case of Facebook and Twitter timelines.

Finally, the concept of the event was introduced, and its long-standing influence on user interfaces for solving the challenges outlined earlier. Shedding light on the breadth of event-centric user interfaces systems sets the scope for a more technical review of the underlying data-mining techniques, to inform the development of *SAESNEG*

The remainder of the thesis pursues the development of a system for the unsupervised organisation of a user’s personal social media ‘footprint’ into a set of events, facilitating a user interface conducive to a reminiscence experience. The proposed system brings a lifelogging user experience to the online social network ecosystem, and the user data it already contains.

Wider issues have been discussed, including the non-ubiquity of wearable cameras, and the convergence of mobile phones, social networking and lifelogging. As social media becomes the de facto medium for the storage and exchange of personal information, the tasks of organising sensor data (from mobile phones) and organising social media data gradually converge. The need for a universal, extensible system for handling heterogeneous social networking data is clear.

This chapter maps the landscape in which the major contribution lies: lifelogging in the context of social media, whilst cementing beyond doubt that such a system will be genuinely useful and address a real opportunity. Novelty is also established; whilst this chapter has identified several classes of related systems documented in existing work, to the author's knowledge, no existing system is able to organise different kinds of social networking data into events.

The next chapter will review more technical aspects of some existing systems, as well as survey wider techniques for handling images – and introduce another relevant research area – natural language processing, for handling textual data in social media.

Chapter 3

Literature Review: Technical

> "We note that such social networks are extremely rich, in that they contain a tremendous amount of content such as text, images, audio or video." Aggarwal, 2011, p. 4

3.1 Introduction

The previous chapter discussed the wider context, lifelogging, and the challenges of preserving the personal social media footprint, as well as reviewing a number of proposed solutions for the organisation of personal documents. To complement this, the current chapter critically reviews existing, relevant work from a technical perspective, reviewing the broad set of topics which are pertinent to the development of a system of this scope: the typical social media footprint contains a variety of different kinds of data, techniques to extract information from this existing data must be reviewed. Some of the existing systems have previously been discussed in the previous chapter; here they approached from a technical angle, and a range of information extraction techniques are reviewed.

The central argument of this thesis is the recognition that techniques employed for organising photographs into events i.e. detecting events in images (e.g. photos), can be *combined* with techniques for detecting events in textual data (newswire, microblogging) in a single pipeline for performing event detection and clustering on an entire social media footprint. The next two sections are concerned with techniques to extract information from these two kinds of source data, the so called 'Phase A' of the system.

The analysis of text, and the field of natural language processing (NLP) is the focus of many of the algorithms developed for *SAESNEG*. Section 2, starts with an overview of the field, before drilling into tasks and topics relevant to the development of the system, and more recent work is reviewed. Frameworks and toolkits, which combine multiple tools for different processing steps are reviewed. Issues such as the handling of text from social networks (and the issue of *unnatural language*) are discussed in section 3.2.4 and throughout.

Regarding images, the previous chapter discussed the long history of systems for organising collections personal photographs, much of the discussion focused on the organisation of collections of personal photographs, with more recent work in the context of online social networks. Clearly, social media contains an abundance of photographs, so techniques from the topic of content-based image retrieval (CBIR) to extract information from those images will be useful for the event clustering task. Section 3.3 discusses some of the techniques which have proven useful, and used by the photo management systems discussed in the previous chapter. Frameworks which combine some of these different techniques are evaluated, as a means of incorporating functionality into *SAESNEG*.

Building on this broad foundation in the key techniques and concepts; section 5 reviews examples of systems, some of which have already been mentioned in the previous chapter. It explores how these existing systems have combined techniques on the event clustering task (on a variety of datasets), and outlines what is effectively the state-of-art for these related systems (with a focus on implementation). This section explores how existing work has combined information relating to event facets, and strategies to the computation of document and event similarity.

In reviewing these existing systems, and discussing the state-of-the-art, a basis is formed to establish the novelty of the proposed system, through its scope (of source data), and application. This section forms the starting point for the design of the system architecture in the following chapter: deconstructing the confectionary of techniques, algorithms, and strategies comprising the related systems, so that they can be soundly and cleanly integrated into a single, extensible framework. The notion of the event, and understanding its varied interpretation across this existing work, is key to this process.

3.2 Text and Natural Language Processing

3.2.1 Introduction

This thesis is built on a body of existing work on the detection of events in textual data, spanning several years. Existing systems which extract events from text are underpinned by natural language processing (NLP). An overview of NLP is presented, with discussion of key concepts and important historical influences.

More recent issues are covered: developments in the field have shifted some of the attention from traditional corpora such as newswire, towards web and social media sources such as micro-blogs, bringing challenges to existing tools, with the issue of 'unnatural language'. Key NLP techniques are explained, followed by deeper discussion of the techniques and approaches most pertinent to the event detection task in the area of information extraction. The review will show that extracting information for the respective 'who/what/where/when' event facets is the basis of the majority of approaches to event detection and extraction. The kinds of data associated with these facets, such as the names of people, temporal expressions, places, each have their own specialised techniques, each of which are critically reviewed.

The unvarnished W*[1] event model used by many existing systems contrasts with

[1] An abbreviation used in this thesis to mean various selections of

the nuanced vision of Jain and Westerman (Westermann and Jain, 2007); discussed in the previous chapter, perhaps presenting opportunities to extend the state-of-the art: wider areas of NLP are briefly reviewed, such as opinion mining and sentiment analysis – to inform discussion of more complex event models, and other developments. More importantly, existing pipelines and systems have sought to integrate this pipeline of algorithms (both for general purposes, and specific applications). Significant examples, including GATE, Stanford CoreNLP, and TwitIE; are reviewed and evaluated.

NLP is a key theme of the thesis (more so than CBIR); this section gives a critical review of relevant topics in the field, both supporting later contributions to specific areas, and informing discussion of the design and implementation of *SAESNEG* but the section starts with some key definitions.

3.2.2 Natural Language Processing

Natural languages "are the languages that real people speak" (Jurafsky and Martin, 2009, p. 5) yet NLP can be hard to define clearly. A wide range of tasks are described as NLP in the literature: converting from natural language to machine-readable form, vice-versa, perhaps understanding and/or generating speech, translating between natural languages, or simply counting the number of words in document.

A presentation by Cardie (2011) leads us to a clear definition of the field: the study of algorithms that have natural language as their input or output. With this range of work, it is no surprise that NLP is said to build on a variety of fields, with contributions from experts in computer science, electronic engineering, and linguistics (Jurafsky and Martin, 2009, p. 3).

Simple techniques can be highly effective: Bayesian classifiers using bag-of-words features are used for the sentiment analysis task, indeed simply counting the number of words in a document is an example of NLP according to this definition.

In her overview of NLP, Cardie (2011) says that most of the research effort in NLP is done on algorithms with NL as their input; and output non-NL data, specifically those algorithms that take NL input in the form of text strings[2]. These are also the topics most relevant to this thesis, and we focus on these topics for the remainder of the chapter.

However, a huge range of topics come under this category, for example: information retrieval, information extraction, knowledge retrieval, knowledge extraction, knowledge discovery, summarisation, opinion analysis, opinion mining, sentiment analysis, natural language understanding, document classification, text mining, semantic annotation. Further categorisation is necessary to understand the literature and identify relevant work.

Helpfully, Cardie (2011) goes on to suggest a sensible taxonomy, sub-dividing the tasks and work in this area into two key groups: **information extraction** (broadly defined as concerning the extractions of facts and factual information) and **opin-**

Who?/What?/Where?/When?/How?/Why? as used by other authors.

[2] This perhaps explains why definitions of NLP often overlook topics such as speech synthesis and optical character recognition.

ion analysis (dealing with subjectivity and opinions). This division emphasises the high-level applications, but perhaps overlooks the set of core techniques underpinning both areas[3]. These core techniques, such as **tokenisation** (identifying words within a string of text), and tagging grammatical **part-of-speech** are introduced first, followed by a deeper survey of pertinent work on information extraction. Discussion of these core techniques and a scrutiny of information extraction, follows proceeding sections on history and the context of OSNs, and language models.

3.2.3 History and Paradigms

The early decades of computer science saw the establishment of two of the early ideas in natural language processing: the automaton (Kleene, 1951) and generative probabilistic models of text (Shannon, 1948). Various research goals and approaches were favoured in the subsequent decades, but the goal of achieving 'understanding' was particularly influential, these Chomskyan theories (Chomsky, 1956) are seen as a contrast to the statistical approaches more popular today which actually have more in common with the foundational work.

Understanding is a subjective concept, and techniques from machine-learning mean that algorithms can perform a range of useful tasks on natural language, without understanding in any deep sense. The goal of achieving understanding (in a human-like sense) perhaps fuels a perception that that NLP is intrinsically hard (and, by implication "easy" tasks are not real NLP). Obviously, there are certainly a lot of tasks within the field that are exceptionally hard (this is true of all disciplines), but the literature is full of effective algorithms built on techniques that are not. But this view is widespread, with universities teaching that "NLP is hard" (Copestake, 2007; Barzilay, 2010) – perhaps reminiscent of Chomskyan goals and ideas.

The history of the field is often described (Gold, 2012) in terms of these two rival 'Chompskyan' camps (Jurafsky and Martin, 2009). But, presenting it this way is perhaps a simplified view. Norvig, seen as an evangelist of statistical NLP, actually sees this as a "false dictomy" between the two camps (Halevy et al., 2009), yet his views are often misrepresented (Amatriain, 2012). Pragmatically, they see a higher level of abstraction, deconstructing both approaches[4].

Aside from this discussion, two undeniable trends of recent decades are a growth in the volume of data available for analysis, through growth of the web. Also, the use of datasets which do not require human annotation, such as the use of emoticons and consumer 'star ratings' as labels of document sentiment (Read, 2005), and multi-lingual documents for the machine translation task. These large datasets were ideally suited to machine-learning techniques, growing in popularity. Weikum et al. (2012) give an overview of some of these techniques, applied to 'big data'.

[3] Named entity recognition, for example, is often part of an opinion-analysis system.

[4] However, the authors are critical of RDF.

3.2.4 NLP in Online Social Networks and 'Unnatural' Language

The focus of the thesis is on text from the social media media footprint. A topic often discussed in the context of OSN-sourced text is the phenomenon of 'unnatural' language, defined as "informal expressions, variations, spelling errors ... irregular proper nouns, emoticons, unknown words" (Hagiwara, 2010)

Existing NLP tools are known to struggle with such language: "social media poses a number of challenges for language analysis tools due to the degraded nature of the text" (Dietze et al., 2012), on tasks such as part-of-speech processing (discussed in section 3.2.6.2). Twitter, by virtue of the fact it is a popular, public global micro-blogging platform makes it an overwhelmingly popular choice of OSN corpora. Examples of work on other social networks are less common, and often see the wider issue as having secondary or even nugatory importance, perhaps focusing instead on a single phenomenon. Robertson et al. (2009) studied the walls of candidates for the 2008 US Presidential election, to plot sentiment-time graphs. To the author's knowledge, his 2012 study (Blamey et al., 2012) was the first to use Facebook photo comments as a corpus. With a handful of studies using Facebook status messages, such as Tang et al. (2012), to overcome what they see as the inherent problems associated with short-document corpora built from Facebook status messages or Tweets.

With much of the attention on Twitter, it is worth investigating the supposed uniformity of language characteristics across different social networks. Unnatural language has perhaps become a 'catch-all' term for a variety of phenomena not found in standard English. There has been little effort to scrutinise the definition of the term, and no evidence to suggest that the phenomenon is consistent, or even relevant to all social networks.

Clearly, there are social, cultural, and practical issues associated with notions of 'misspelling' (Oatley et al., 2015). The presence of emoticons is one issue, text speak presents an entirely different set of challenges. Twitter has unique concerns – aside from hashtags (e.g. '`#belieber`') and mentions (e.g. '`@ladygaga`') – the 140-character limit undoubtedly has some (unique) effect on the language. Twitter is a global forum, users converse in a multitude of languages, whereas a particulr community of Facebook friends may have a completely different demographic. Nevertheless, it would be foolish to generalise across all networks and all users. This study will need to balance the adoption of new techniques effectively developed for Twitter, with more mature techniques which may actually give reasonable performance on an alternative social network corpus. An in-depth study of the pervasiveness of these issues awaits future research.

3.2.5 Language Models

At its simplest, a language model is a means of computing 'the probability of a string of text'. The n-gram model of language, arguably as old as computer science itself (Shannon, 1948), can be formulated in two ways. Text can be modelled as a sequence of words (or overlapping sequences of consecutive words), or, as a sequence of characters, without consideration of individual words (or overlapping n-length strings[5]).

The character n-gram model has proven successful in a number of scenarios, both outside NLP (compression, cryptography), and within it: information extraction (Klein et al., 2003), Chinese word segmentation (Xue, 2003), author attribution (Peng et al., 2003b), language identification (Cavnar and Trenkle, 1994), sentiment analysis (Habernal et al., 2014), sarcasm detection (Ptácek et al., 2014).

The simplicity of the character n-gram has a number of advantages over the word-based model. Peng et al. cite drawbacks of the standard text classification methodology including "language dependence" and "language-specific knowledge", and notes that "In many Asian languages ... identifying words from character sequences is hard" (Peng et al., 2003a, p. 110). Software for character n-gram classification has been available for years: the Stanford Classifier[6] (Manning and Klein, 2003), LingPipe[7] (Alias-i, 2008), amongst others.

These views are supported by empirical findings: Carpenter, who has stated publicly that he favours character n-gram models (Daumé, 2006; Carpenter, 2010), software is able to achieve performance equivalent to word-based classification of movie reviews in another study (Carpenter, 2011; Pang and Lee, 2004) with his LingPipe framework. Whilst Peng et al. apply the character n-gram model to a number of NLP tasks, including text genre classification observing "state of the art or better performance in each case". (Peng et al., 2003a, p. 116). More recently, Rybina (2012) does binary sentiment classification on a selection of German language web corpora, finding that character n-grams consistently outperforms word n-grams by 4% on F1 score. Such work motivated Blamey et al. (2012) to experiment with character n-grams (in the context of OSN text), finding little improvement over the word-based model, contrary to Rybina's work.

The word-gram alternative is also used for document classification, often with the so-called bag-of-words model. Aside from these applications, the word-based model is ubiquitous across information extraction and opinion analysis disciplines. The additional step of splitting the string into words or tokens is known as tokenisation, and is discussed in the following section.

3.2.6 Tasks and Techniques

3.2.6.1 Tokenisation

Tokenisation "is the task of separating out (tokenizing) words from running text" (Jurafsky and Martin, 2009). At first glance, this task seems deceptively simple: just break up the words where there are spaces[8]. Punctuation complicates the matter a period may delimit a sentence, or form part of a number. Numbers, incidentally, may contain spaces. There is no mutually agreed definition of what constitutes a token;

[5] Another language model is that of the skip n-gram, where the set of skip-N-grams for a given string are "any N words in sentence order" Also, S-skip-N-grams (vary s to a maximum of S, where s is number of words skipped in total over an N-word-gram sequence), have been found to give greater coverage (Guthrie et al., 2006).

[6] `http://nlp.stanford.edu/software/classifier.shtml`

[7] `http://alias-i.com/lingpipe/`

[8] Certain languages, such as Chinese, are written without spaces, further complicating the task.

Jurafsky and Martin (2009) put forward examples such as *New York* and *rock 'n' roll*.

Emoticons (which contain punctuation characters) further complicate this process – the emoticons themselves can contain spaces, such as :) and : (used by Go et al. (2009). Wikipedia has an extensive list of both 'Western' and 'Eastern' emoticons[9].

For example, the emoticon :D should be found in '*He finally asked:D*', but not in '*He asked:Did you see the new movie?*' (Narr et al., 2012). Such 'unnatural language' text presents difficulties for long-established tools

In response to these challenges, new tokenizers have been developed: O'Connor et al. (2010) developed a regex-based tagger capable of recognizing "hashtags, @-replies, abbreviations, strings of punctuation, emoticons and unicode glyphs (e.g. musical notes)" as tokens. See Ptaszynski et al. (2011) for a comprehensive survey of techniques for emoticon extraction. The GATE gazetteer can also be set up for emoticon detection, with an appropriate gazetteer list[10].

Related tasks include that of URL breaking – identifying individual words from inside URLs (Salvetti and Nicolov, 2006), and handling repetition of characters, so that 'LOOOOOOOOL' becomes 'LOL' and 'hahahahaha' becomes 'hahaha', (see Blamey et al. (2012) for a discussion). Despite this, tokenisation is sometimes considered such a simple task "precise details are often thought too tedious for publication" (Blamey et al., 2012). Subsequent processing for IE or opinion analysis often rely on knowledge of sentence boundaries, so pipelines commonly include a sentence splitter, (Cunningham et al., 2001–2014, sec. 6.4).

3.2.6.2 Part-of-Speech Tagging

Part-of-speech tagging is the task of annotating tokens with their grammatical 'part-of-speech', such as noun or verb – each divided into various sub-categories depending on the precise scheme used.

POS tagging is an important preliminary step in many applications. For machine-learning learning techniques, the part-of-speech label associated with a word is often used as a feature for further classification. However, this is not always the case. None of the Stanford NER models use part-of-speech tags "because the features used by the Stanford POS tagger are very similar to those used in the NER system, so there is very little benefit to using POS tags" (The Stanford Natural Language Processing Group, 2014). The part-of-speech task is discussed because of its use in many of the popular text-processing pipelines, and because many of the key techniques, approaches, and developments mirror that of the NER task: identifying named entities is very similar to identifying proper nouns. As in NER, there are broadly three approaches, often used in combination.

Dictionary-based, or lexical approaches, in which one can simply have a list of words associated with each part-of-speech. The Moby Lexicon (Ward, 1996) includes such a resource. This approach has clear intrinsic limitations, as words frequently have senses corresponding to different parts of speech; a simple response is to "pick the tag

[9] http://en.wikipedia.org/wiki/List_of_emoticons

[10] http://greenwoodma.servehttp.com/svn/repos/open-source/show/blog-code/emoticons/western.lst?revision=HEAD

which is the most common for that word" (Charniak et al., 1993, p. 2), which can be highly effective (their study achieved accuracy of 90%).

Creating (and updating) such lists can be prohibitively expensive, consequently such resources can become out of date, especially in the case of proper nouns. However, such lists are a component of more state-of-the-art POS taggers (Owoputi et al., 2013), just as gazetteer lists are used in many NER systems (see section 3.2.6.8).

For decades, rule-based systems were developed (Klein and Simmons, 1963; Greene and Rubin, 1971). This is clearly a huge and difficult task, as noted by Brill "the rules in rule-based systems are usually difficult to construct and are typically not very robust" (Brill, 1992, p. 1). As with lexical resources, rules are used in the NER task. GATE includes the JAPE [11], with similar technology in the Stanford toolkit[12].

In the 1990s, statistical-based approaches rapidly achieved better performance than rule-based systems. The Brill Tagger is a famous example (Brill, 1992), – noted for the fact that rules are learnt automatically, using orthographic features (and POS tags) of surrounding words. Into the next decade, a variety of machine-learning techniques were being applied to the task, often using labelled corpora to train the machine classifiers, such as Hidden Markov Models (Brants, 2000) and maximum entropy (Ratnaparkhi, 1996), and averaged perceptron (Spoustová et al., 2009) (itself developed by Collins, 2002). Such techniques soon achieved per-token accuracy of over 90%, and began to mature.

However, these mature techniques are known to struggle when applied to text from Twitter. Ritter et al. have "demonstrated that existing tools for POS tagging, Chunking and NER perform quite poorly when applied to Tweets" (Ritter et al., 2011, p. 1532). After examining the reasons for this performance drop, they manually annotate a Twitter corpus and train their own tagger: T-POS, employing conditional random fields (CRF) (Lafferty et al., 2001), performing Brown clustering (Brown et al., 1992), to supplement features – a means of achieving inexact word matching. They add POS classes for hashtags, URLs, @-mentions, and reTweets (which they say can be reliably detected with regular expressions alone).

Whilst "[lengthening words] is a common phenomenon in Twitter" (Brody and Diakopoulos, 2011, p. 569), presenting a problem for lexicon-based approaches. Owoputi et al. found similar issues – noting the difficulties of proper noun detection in Twitter, because "neither correct capitalisation nor spelling can be used as reliable indicators" (Owoputi et al., 2012). They improve the performance of an existing Twitter-specific POS tagger (Gimpel et al., 2011) by adding various name lists, various orthographic features and, again, incorporating hierarchical word clustering (Brown et al., 1992) (using Liang's implementation 2005). Dietze et al. (2012) suggest other strategies, such as: "using techniques from SMS normalisation, retraining language identifiers, use of case-insensitive matching in certain cases, using shallow techniques rather than full parsing". Whether these issues associated with Twitter also apply to text from other social networks, is an open research question. The focus on Twitter is discussed back in section 3.2.4.

[11] `http://gate.ac.uk/sale/tao/splitch8.html#chap:jape`

[12] `http://nlp.stanford.edu/software/tokensregex.shtml`

3.2.6.3 Information Extraction

This section gives an overview of relevant topics in information extraction an umbrella term for a number of closely related (and often equivalent) tasks, concerned with the extraction of (structured) factual information from (unstructured) text. This area is distinct from the extraction of opinions and related information from text, referred to as opinion analysis, with tasks such as sentiment analysis (this division is discussed in section 3.2.2).

Various kinds of mentions of factual information in text have been the focus of research. The DARPA-funded Message Understanding Conferences (MUCs) types (used for their IE competitions) illustrates this:

ORGANISATION: named corporate, governmental, or other organisational entity

PERSON: named person or family

LOCATION: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)

DATE: complete or partial date expression

TIME: complete or partial expression of time of day

MONEY: monetary expression

PERCENT: percentage

(Chinchor, 1997, section 2)

This scheme has been adopted as a standard in systems such as GATE (Cunningham et al., 2001–2014, sec. 6.7). For the event-detection goal, there are two sets of entities strongly associated with informational facets of the event: named entities, i.e. people, locations, and perhaps organisations (these relate to the who/where/what facets), and temporal expressions – dates, times etc. (the 'when' facet). Named Entity Extraction (i.e. people and locations) is discussed next (sections 3.2.6.4-3.2.6.10), followed by temporal expressions (section 3.2.7).

3.2.6.4 Named Entity Extraction and Sub-Tasks

This section reviews a significant body of research concerning the extraction of named entities in text (excluding temporal expressions). There is no single formal definition of named entity extraction; a number of terms are used to describe a number of related tasks: named entity recognition, reference resolution, co-reference resolution, entity-linking, record-linkage, relation extraction, entity resolution, etc. As a step towards understanding these many terms (and their associated literature) two key goals are worth identifying:

1. Detecting a reference to a named entity, identifying its boundaries within the text. Sometimes known as 'named entity recognition' (NER).

2. Annotating that reference with structured information – this could simply be its category (i.e. person or place), a sub-category (i.e. capital city), or even a reference or link to some external knowledge base – such as the relevant Wikipedia article.

These two goals may be achieved in a single step, such as by using a gazetteer, with separate lists for people and places. Alternatively, it can be considered as two separate tasks: named entity recognition, followed by named-entity linking. As with part-of-speech tagging, ambiguity can present a challenge to both goals – deciding whether or not a word is being used a proper noun, and if it is, to which specific entity is it referring (known as named entity disambiguation).

Clearly, the twin goals are intrinsically linked, and approaches often tightly integrate them. Hence, the review presented here is structured around the techniques used, rather than attempting to separate the two tasks – the focus is on a few examples, so that they can be discussed in moderate detail. Other tasks, including co-reference resolution (dealing with multiple references to the same entity within a document) and relation extraction (building a knowledge base of factual information) are briefly discussed afterwards.

3.2.6.5 Named Entity Recognition

3.2.6.6 Rule-Based Approaches

As with the part-of-speech tagging task (discussed earlier in section 3.2.6.2), hand-crafted rules are a simple but effective approach for detecting mentions of named entities in text. JAPE, part of the GATE Framework (Cunningham et al., 2001–2014, chapter 8) is a system that allows users to specify rules to match sequences of tokens, so that particular tokens can be annotated as named entities. Rules can match based on part-of-speech, orthographic features, with a variety of matching functionality.

3.2.6.7 Sequence Models

Sequence models apply machine-learning techniques to learn features of words that tend to surround mentions of named entities (in training data), effectively modelling the sequence of words as a Markov chain. Two established approaches to this general task are the discriminative maximum entropy Markov model (MEMM), and the generative hidden Markov model (HMM); the distinction between these two general classes of probabilistic models are discussed below:

Discriminative and Generative Probability Models are two classes of probabilistic models for machine-learning.

This distinction is key to many discussions on the efficacy (and hence selection) of machine-learning techniques, in many of the domains relevent to this thesis.

Machine-learning tasks can often be characterised as computing the most likely label or output variable for a given set of input data. In such a scenario, we wish to model the relationship between an unobserved variable y, and observed (i.e. input) variables x.

Discriminative models (also known as conditional models) do not attempt to model the relationships between input variables, instead taking the input data as given (modelling $p(x|y)$). Conversely, generative models compute the probability of the combination of the output data and the input data, $p(x, y)$ (this requires iterating through all the output values, hence, there must be a finite set). Lafferty et al. (2001) discusses the distinction between the two classes. See Jordan and Ng (2002) for a discussion.

Examples of discriminative models: Logistic Regression, Support Vector Machines (Corinna Cortes, 1995), Conditional Random Fields (Lafferty et al., 2001). Examples of generative models: Naive Bayes, Hidden Markov Model (Baum et al., 1970).

Lafferty et al. (2001) proposed the conditional random fields approach, and demonstrated that their CRF-based model outperforms HMM and MEMM-based models, by combining the advantages of both models. A reason for their better performance was that CRFs are discriminative, thereby obviating issues related to the interactions between input models (which can become intractable for long-range dependencies). Their approach also overcomes a limitation of maximum entropy Markov models (MEMMs) (and other similar models): the label bias issue, where states with many outgoing edges are penalised, and conversely, states with fewer outgoing edges are favoured. They empirically demonstrate the label-bias problem exists, and present parameter estimation algorithms to overcome the problem, as well as demonstrating performance improvement over HMM and MEMM approaches on the POS tagging-task. Applying the approach to the NER task achieved performance of 84%, (McCallum and Li, 2003).

The development of CRF, with their effectiveness at the NER task work is seen as a key milestone: the Stanford CoreNLP NER component, seen as state-of-the art pipeline, is built with CRF classifiers.

3.2.6.8 Gazetteer-based Approaches

Turning away from sequence models, using lists of words is a simple and highly effective technique: searching for occurrences in the text, as with the part-of-speech task. A key advantage of this approach is its simplicity, making it much faster and easier to implement, and computationally cheaper. Such dictionaries are called gazetteers. The popular GATE framework includes a gazetteer in its standard pipeline (ANNIE), which uses simple plaintext word lists for various categories of named entities (Cunningham et al., 2002, chapter 13).

Usually, the gazetteer is built from some external resource, such as an atlas or encyclopaedia.

An important advantage of the approach is that a mention which has been matched with an entry in the gazetteer is implicitly associated with the corresponding external entity, the named entity resolution (or entity linking) task. This contrasts with rule-based or statistics-based parsers, where – having detected a mention – the 'entity linking' step needs to be performed separately.

However, there are also several disadvantages to this approach: coverage, the need for exact matches, and ambiguity, to which Rao et al. (2013) allude in their definition of entity-linking "aligning a textual mention of a named-entity to an appropriate entry in a knowledge base, which may or may not contain the entity". Research has attempted to overcome the challenges through the application of statistics and machine-learning techniques.

3.2.6.9 Beyond Gazetteers

Several authors have proposed using Wikipedia to learn how to identify mentions of named entities in text, often using hyperlinks as a training set for associating named entities (articles) with their mentions (often marked-up as hyperlinks). An early use of Wikipedia in NER was the work of Bunescu and Paca (2006). They show how Wikipedia can be useful for both the recognition and disambiguation tasks; detection is achieved by building an index of what they call 'display strings' used to refer to named entities, based on hyperlink text. The disambiguation task is approached as a ranking problem, using TF-IDF[13] scores to compute the similarity between the context of the mention, and the target article (to overcome sparsity, documents from parent categories are used as a boost). Article redirection and disambiguation pages are also considered, and a SVM kernel is used to trigger a special "out of wikipedia" entity.

A number of authors have adopted a similar approach: Milne and Witten (2008) also used Wikipedia[14], and focus on the disambiguation tasks. As in other previous work (Mihalcea and Csomai, 2007), instead of using similarity alone, they also consider the prior probability 'commonness' of the article. They investigate various strategies for balancing the importance of commonness with semantic similarity of the documents (considering more than just TF-IDF). They account for the effect of common words like 'the', which, as they often appear in the context act as a proxy for prior commonness, skewing the calculation.

More recently, Spitkovsky and Chang (2012) published a dataset of directed probabilistic mappings between short phrases and Wikipedia article URLs, the anchor text frequencies (based on Google crawl data) of hyperlinks directed at the articles, from inside and *outside* Wikipedia.

For a given article, the set of link anchor text is "potentially a useful source for mining synonyms" (Spitkovsky and Chang, 2012, p. 1), whilst the converse set of articles targeted by a given anchor text "exposes the ambiguity inherent in the string ... by distributing probability mass over several concepts" and "effectively disambiguates ... by concentrating most of its probability mass on a single entity" (Spitkovsky and Chang, 2012, p. 2).

A different strategy was used by Ritter et al. (2011) when developing their Twitter NLP pipeline (the Wikipedia articles arguably less suitable for entities in the Twitter domain, which can be more transient). They note the issues of using gazetteer lists for NER mentions of named entities frequently do not match the list because of mis-spellings, or because an acronym or abbreviation is used; alternatively the entity simply may not exist. Their goal

[13] *term frequencyinverse document frequency* - see: `https://en.wikipedia.org/wiki/TFIDF`

[14] The authors had different motivations: Benescu and Pasca were motivated by search result grouping, whilst Milne and Witten were interested in adding hyperlinks to documents

was to tag named entity mentions with an entity-kind label (the so-called 'named entity classification task'), they employ the Freebase dictionary lists [15] as a means of 'semantically bootstrapping' the machine-learning from their labelled training data, by using a technique called LabeledLDA, originally developed to model topics based on social bookmarking tags (Ramage et al., 2009).

In outline, their approach uses Latent Direchlet Allocation (LDA) (Blei and McAuliffe, 2007) model the topics associated with an entity mention (where the set of topics bijects with the set of freebase dictionary category). Furthermore, the set of topics associated with a particular entity string is further constrained to be the set of freebase dictionaries that contain that entity string[16]. The category and context words for a mention are predicted from the LDA models (i.e. a generative model). Training data is used to set the prior distributions, in which the context words associated with particular categories are learnt, further leveraged by the use of the freebase dictionaries. Overall, they achieve significantly better performance with this approach on their Twitter corpus, than using the Stanford tool.

3.2.6.10 Combining Gazetteers and CRF

Instead of developing what is essentially a gazetteer-based approach, an alternative strategy has been to incorporate gazetteer-based features into sequence model/CRF techniques, results have been rather mixed:

Smith and Osborne (2006) investigated including gazetteer alongside linguistic (word n-grams, POS-tags, orthographic) features for CRF-based NER taggers, based on a gazetteer including 27,265 places, with smaller numbers of other named entities. Note that their gazetteer is an order-of-magnitude smaller than web-scale gazetteers, such as those built from Wikipedia. They found a 1% improvement in accuracy, despite individual cases where the new tagger fails. The interpretation is that their model becomes overly reliant on gazetteer features, citing cases where performance drops for "sequences where no words are in any gazetteer", with similar difficulties when a token appears in both location and person name gazetteer lists. They attempt to obviate these issues, first by using separate priors for gazetteer and non-gazetteer features, before entirely separating the gazetteer and non-gazetteer features, and using a logarithmic opinion pool (Smith et al., 2005) to combine the results from both models.

Looking further back, other authors are critical of using gazetteers (Krupka and Hausman, 1998; Morgan et al., 1995), while others are more positive (Mikheev et al., 1999). However, as training data for NER tasks is so seldom available, it is difficult to experiment with the new features, and re-train classifiers.

3.2.7 Temporal Expressions

The *when?* facet is key to the event detection task (section 3.4), detecting and resolving temporal expressions is useful for identifying event commonality. Research into temporal expressions has generally focused on their detection and grammatical parsing (exceptions include the development of *grammar of time expressions* (Angeli et al., 2012)). The tempex task is beset with a number of complications: dealing with relative expressions ("next Tuesday"), a great variety of formats for representing dates and times, ambiguity (e.g. two-digit

[15] `http://www.freebase.com/`

[16] The example of 'Amazon' is given, either a company or a location.

years), and cultural-specific concerns (such as national holidays).

For these reasons, systems tend to use hand-coded rules to describe a formal grammar; resisting the wider trend towards statistical and machine-learnt approaches. Popular frameworks using such an approach include: HeidelTime (Strötgen and Gertz, 2010), GUTime (Mani and Wilson, 2000), with the more recent SUTime (Chang and Manning, 2012), part of the Stanford CoreNLP framework[17], considered to be a state-of-the-art system, as measured on the TempEval-2 dataset (Verhagen and Pustejovsky, 2008). Hence, the SUTime framework is a sensible choice for extracting meaning from temporal expressions found in the text associated with the social media documents – and can be readily incorporated along with other components from the Stanford CoreNLP framework.

As with the more general NER task; detection is not the only goal, temporal expressions need to be assigned a well-defined temporal meaning. The defacto representation for temporal expressions (and the one used in this study) is to represent the temporal expression as an interval: a day, a year. For less 'precise' temporal expressions such as 'Summer', a fixed time range needs to be chosen as the representation (in the case of Summer, SUTime uses the range between the equinoxes). A criticism of the existing work is that temporal expressions are represented as specific dates, or discrete intervals; hence, only expressions that can be represented in this way are considered within the scope of the task, an issue noted by Angeli et al. (2012, p. 9) "Futhermore, vague times (e.g. in the 1990s) represent a notable chunk of temporal expressions uttered. In contrast, NLP evaluations have generally not handled such vague time expressions".

The terminal set of the grammar is generally the months, dates, days of the week, religious festivals, public holidays, usually with an emphasis on the culture of the authors[18]. SUTime can be configured to use the JollyDay[19] library, which contains definitions of important dates for many cultures, but even in this case the definitions are restricted to the model of discrete intervals.

This approach is the most natural and simplest approach to the mathematical modelling of time, used throughout natural science. Work such as that of Allen (1981) is often cited as a philosophical underpinning of this model. Such an approach is useful for describing the physical world with mathematical precision, but is a poor means of describing the cultural definition of temporal language. In cases where it is difficult to assign specific date ranges, the advice is to leave alone:

> "Some expressions' meanings are understood in some fuzzy sense by the general population and not limited to specific fields of endeavor. However, the general rule is that no VAL is to be specified if they are culturally or historically defined, because there would be a high degree of disagreement over the exact value of VAL." (Lisa Ferro et al., 2005, p. 54)

An advantage of using this restricted, well-defined vocabulary is that it facilitates numerical evaluation of parsing accuracy, and performance can be compared with standard datasets, such as those from the TempEval series (Verhagen and Pustejovsky, 2008; UzZaman et al., 2013).

However, this emphasis on grammar has resulted in the research community overlooking the meaning of the terminals themselves. A recent development is the approach of Brucato et al. (2013), who, noting the maturity of tools developed for the traditional tempex task,

[17] `http://nlp.stanford.edu/software/corenlp.shtml`

[18] Often British.

[19] `http://jollyday.sourceforge.net/index.html`

widen the scope by to include so-called *named temporal expressions*, by creating a list from Wikipedia articles, merging the results with the JollyDay library. This list is used to train a CRF-based detector, which in turn was used to find completely new NTEs, such as sporting events.

However, they note the difficulties of learning definitions for the newly-discovered NTEs:

> "...it is difficult to automatically learn or infer the link between *New Year's Day* and *1st January*, or the associations between north/south hemisphere and which months fall in summer..." (Brucato et al., 2013, p. 6).

They resort to TIMEX3, a traditional, discrete interval representation. Secondly, it means that the scope of the temporal expression task is restricted to only those entities which can be represented as discrete intervals. In section 5.5 a novel alternative approach of *distributed approach* to representing the meaning of temporal expressions is presented – the usefulness of the technique when used in the overarching event clustering task is evaluated later in the thesis 6.11.

3.2.8 Pipelines and Systems

The preceding sections have given an overview of some existing work relating to the core tasks relevant to event detection and the calculation of event commonality, i.e. the extraction of the informational 'who/what/where/when' event aspects. The goal is ultimately to develop an integrated framework for organising social networking data into events. As has been alluded to, a number of frameworks for text processing already exist, integrating many of the techniques and approaches discussed here.

The GATE family of products and services, developed at the University of Sheffield, includes 'GATE Developer', an integrated development environment for language processing. The user is able to assemble his or her own pipeline of text processing components, and run it on a collection of documents. The resulting annotations can be inspected and edited – making the IDE a useful tool for creating gold-standard annotations.

A default pipeline, ANNIE, is provided, readily extensible with an enormous range of plugins and additional components. The underlying text processing pipelines are implemented as a Java library 'GATE embedded', so can be run independent of the GUI. GATE is released under the Lesser-GPL.

The Stanford University Natural Language Processing Group publishes a suite of tools for various tasks, integrated as the Stanford CoreNLP library, with NER and temporal expression components (already discussed) considered state-of-the-art. Again, these are implemented as Java libraries; but unlike GATE, there is no IDE environment, so it can be time consuming to inspect annotations and tweak the software, like GATE, a default pipeline is included.

Apache OpenNLP is a more lightweight library for text processing (again implemented in Java), with a focus on core tasks. The emphasis is perhaps towards simplicity and ease of use rather than state-of-the-art performance and configurability.

In summary, GATE an excellent framework for integrating different components into a single pipeline, with a well-designed user interface for inspecting (for debugging, configuration tweaking etc.) and creating document annotations, as well as extensive documentation.

CoreNLP is perhaps more state-of-the-art in terms of performance, but lacks such a user interface.

Existing work has actually been implemented as GATE plugins, the GATE TwitIE pipeline (Bontcheva et al., 2013)[20], the ANNIE-equivalent for Twitter text, influenced by the work of Ritter et al. (2011), and includes an instance of the Stanford POS-tagger. The extensibility, range of plugins, and the IDE make GATE a clear choice as the basis of the text-processing component of *SAESNEG*.

3.3 Images: Content Based Image Retrieval (CBIR)

Content-based image retrieval has been studied for more than 30 years (Sandhaus and Boll, 2011). A key research focus is the retrieval of images from a database; there are broadly two approaches: either through searching with keywords (or some other query), or retrieving images based on their similarity with some particular image, so-called content based image retrieval.

3.3.1 High-Level Techniques

Image content can be useful for the document event commonality task in a number of ways: some techniques extract high-level information, e.g. recognition of faces (Li et al., 2015), landmarks (Abbasi et al., 2009), scenes (Lazebnik et al., 2006), which can then be used for indirect comparison of photos (and indeed facilitating comparison of photos with types of data) and low-level techniques (colours, texture and structure) for more direct comparison. At a low level, images that are visually similar, i.e. they contain similar colours or have a similar layout, could be photographs of the same 'scene' and hence might be associated with the same underlying real world event. More elaborate techniques might be able to recognise objects in the photos e.g. recognising a Christmas tree might indicate that a photo is associated with Christmas, perhaps a famous landmark could be recognised indicating a precise location for the photo, whereas facial (and indeed clothing) recognition could indicate event attendees, when combined with clothing recognition objects could give clues to the kind of event, the time of year, time of day, facial (and clothing) recognition can determine event participants, whilst the recognition of famous landmarks can give clues to location.

Instead, a range of approaches have been developed to address this image comparison task, in a domain at a higher level of abstraction than the pixels. Such comparisons tend to be a two-step process: compute a summary of each image, the result of which represents the image within a domain that summarises the image. This is followed by a comparison calculation, based on a metric suitable for that domain, able to robustly handle small differences between images. The summary process, representation and similarity calculation are often designed to only represent one aspect of the image content, perhaps its colour or texture.

Such techniques are numerous, so a selection of common techniques which have proven effective are presented in outline; detailed technical descriptions are beyond the scope of this thesis.

[20] https://gate.ac.uk/wiki/twitie.html

3.3.2 Colour Histogram

The intuition of using the colour histogram is that similar images contain similar colours. A typical approach is to count the number of pixels of each hue, to form a histogram. Similarity can be computed by normalising the histogram, and then computing the Euclidean distance treating the bin values as a vector. This technique is robust and easy to understand, but lacks a spatial dimension.

3.3.3 Scalable Colour

An extension of the colour histogram is the scalable colour comparison metric. Histograms are created for Hue-Saturation-Value, which are then normalised, after which the Haar transform[21] is applied (similar to the Discrete Fourier Transform), comparison can be computed in the Haar domain, or by directly comparing the histograms. A standardised version of the metric is included in MPEG-7.

3.3.4 Colour Layout

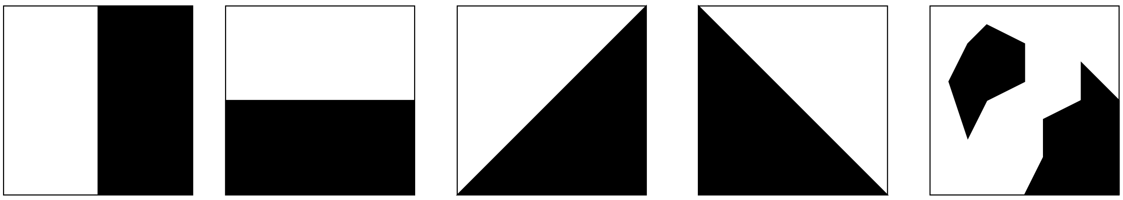
With these histogram-based approaches not capturing any layout information, MPEG-7 defines the 'color layout' standard to combine colour and layout information. Computation of the color layout representation can be done as follows:

1. The image is divided into an 8x8 grid.
2. A representative colour is chosen from each cell, typically by taking the average colour.
3. The color for each cell is represented in the YCbCr scheme (Luminance, red difference, blue difference).
4. Apply the Discrete Cosine Transformation to the 8x8 matrix (similar to the Discrete Fourier Transform, but with natural numbers only), to yield a 8x8 co-efficient matrix.
5. For each component, convert the matrix into a sequence using the zig-zag encoding (this puts the co-efficients for the low-frequency components together). The matching metric is the sum of the Euclidean distances between each of the 3 DCT co-efficient vectors, capturing the general similarity between the patterns of colours in the image.

3.3.5 Texture

Other than colour, a key aspect of images that are summarised by these techniques are the closely-related notions of texture and edges. Detecting and representing edges sounds complicated, but at their core the basic techniques are easy to understand.

[21] `http://www.whydomath.org/node/wavlets/hwt.html`



$$\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \quad \begin{pmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{2} \end{pmatrix} \quad \begin{pmatrix} 0 & \sqrt{2} \\ -\sqrt{2} & 0 \end{pmatrix} \quad \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$

Figure 3.1: Edge Detector Matrices.

3.3.6 Gabor Filtering

One popular approach to calculating and representing the texture in a particular region (assumed to be texturally homogenous) is to apply a set of Gabor filters[22]; the parameters describing the strength of the Gabor wave at a particular scale and orientation can be found by integrating whilst transformed by the Gabor function for a wave with a particular frequency and orientation. The resulting vector of parameters characterises the texture in that region of the image; this technique underpins the Texture Browsing Descriptor (TBD) and the Homogenous Texture Descriptor (HTD) defined by MPEG-7.

3.3.7 Edge Histogram

A simpler, more discrete approach to edge detection is to describe each region of the image with a histogram representing the different edges contained in that region.

MPEG-7 defines a standard procedure for computing an edge-histogram representation of an image (actually a set of edge histograms, describing the dominance of particular edge orientations in particular regions of the image):

1. The image is partitioned into a 4x4 grid of sub-images.
2. Each sub-image is itself sub-divided into a number of blocks.
3. Each block is converted into a 4-pixel grid (by averaging the brightness of the pixels).
4. 'Edge Strengths' are computed for the block by multiplying the strength of each pixel with its associated matrix value, and taking a sum. (The filter matrices are shown in figure 3.1).
5. If the maximum sum is greater than a threshold, the block is counted as having the associated texture.
6. Hence, a 5-bin histogram can be plotted for each of the 16 sub-images, after appropriate normalisation.

Histograms can be local, semi-local, or global. Similarity can be calculated by comparing histograms based on either the raw quantised values (using the L1 "taxi cab" metric) or the decoded values.

[22] For the 2D case, the Gabor wave is a sinusoidal wave in a plane where the magnitude is determined by a 2D Gaussian distribution.

Library	Language	Licence	URL
Caliph-Emir	Java	GPL	http://www.semanticmetadata.net/features/
img(Rummager)	C#	GPL	http://chatzichristofis.info/?page_id=213

Table 3.1: Open-Source MPEG-7 Image Descriptor Extraction Implementations

An extension of this idea is that filters of this kind can be used as convolution[23] matrices, to achieve image transformations, such as blurring, sharpening and a variety of other effects.

Various Software libraries have implemented these low-descriptors, according to the MPEG-7 standard, as shown in table 3.1.

3.3.8 Summary

This section has identified and reviewed a number of techniques that have proven effective for the extraction of low-level features for describing image content, as well as means of computing similarity between such measures. The de facto standard, MPEG-7, has been discussed, and potential implementations have been identified. Motivations, issues and concerns pertinent to working with image content and image content descriptors (colour representation, quantisation) have been identified, with the reader directed to further reading where appropriate. Finally, open-source frameworks have been noted for inclusion in the implementation of *SAESNEG*

3.4 Existing Systems

3.4.1 Introduction

Having reviewed a range of techniques for extracting information from images and techniques, this section revisits existing systems. In the previous chapter, a number of different systems were discussed as applications, considered as solutions. In this section, the focus is more technical. A range of systems are examined, with their common feature being that they extract or detect events, or perform event clustering. Some of the systems do not actually detect events, but are included because they have similar goals to parts of the pipeline being constructed. In the previous chapter, only systems which could be construed as OSN-lifelogging solutions were reviewed.

The goal of this section is to review techniques and approaches to event detection, including other applications. The focus is on techniques thought to be most useful for the implementation of an event-detection pipeline, for processing a user's social media footprint, with its range of different kinds of data. Work from other areas is reviewed, to examine whether it is relevant or useful for the system being implemented.

To facilitate analysis and comparison of this existing work, an attempt has been made to organise discussion around key stages in a theoretical event detection pipeline. These include:

[23] Wikipedia has a good explanation of the mathematical concept: http://en.wikipedia.org/wiki/Convolution

- The **source data** – one (or more) social networks, or somewhere else.
- The **kind** of data to be processed: photos, text, etc.
- How the data was collected – a user's personal data footprint, or gathered by searching or filtering public data from an online social network.
- **Extraction** of information from documents (event facets or otherwise)
- **Resolution** of that extracted information – such as named entities.
- **Event Representation**: how the notion of the event has been interpreted and adapted to the task.
- How systems **match** documents associated with the same underlying real-world event – the various algorithms and similarity metrics used.
- The **performance** of the system, perhaps in terms of some measure of accuracy against ground truth – if such ground truth exists.
- Finally, how the resulting data is presented in a **user interface**.

3.4.2 Event-Centric OSN Databases

The first set of existing systems in this review do not actually perform automated event clustering, they are included because they are arguably a foundation for the preliminary phase of the system planned for this thesis: systems that collect data from a variety of online social networks, and (whilst performing minimal information extraction) normalise the documents into the W*-like informational event facets. These systems are effectively implementations of the solutions discussed in the previous chapter – passive OSN backup. The systems reviewed are the work from Kobe University: Takahashi et al. (2014), Shimojo et al. (2010); and Aiken (2014), the closest predecessors of *SAESNEG*.

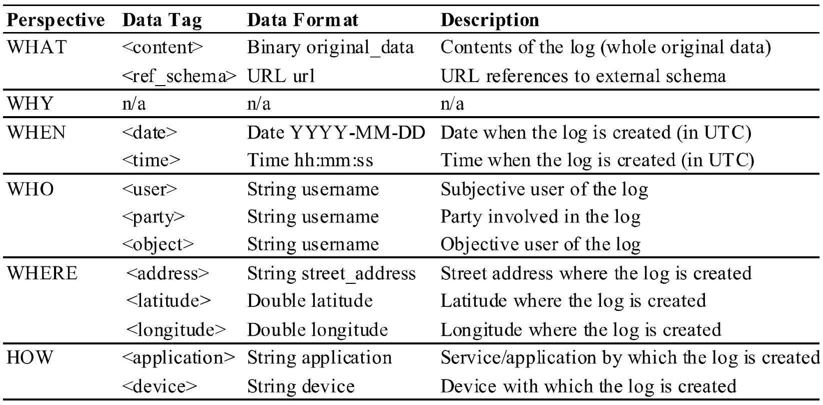
These systems are one approach to OSN archival: fetch all of the OSN documents associated with the user's various social networking accounts – similar to *SAESNEG* in the sense that they work with documents of different types, but unlike *SAESNEG*, they do not perform any event clustering.

Source Data: All three systems designed to fetch documents from more than one social network, through the appropriate web APIs, and are intended to gather data from the user's personal social media footprint. A variety of different kinds of data can be downloaded.

Extraction: Information is simply extracted from the fetched metadata and normalised – dates and times are converted to a standard format, for example.

Event Representation: These systems model events according to the usual w+ information event facets. Unusually, Takahashi (Takahashi et al., 2014) also includes a *how?* facet, and the *what?* is represented by a copy of the original data and a link its schema, as shown in table 3.2. Their paper also discusses the suitability of NoSQL databases, specifically mongoDB, for the preservation of this content.

Aiken 2014 does not use a W* representation explicitly, instead his event class just includes the properties that are displayed in the user interface (extracted directly from metadata) including a description, creation date, user, link to the content, icon, etc. Although some of the the W* facets are recognisable, listing 3.1 shows Aiken's 'event' source code.



Perspective	Data Tag	Data Format	Description
WHAT	<content>	Binary original_data	Contents of the log (whole original data)
	<ref_schema>	URL url	URL references to external schema
WHY	n/a	n/a	n/a
WHEN	<date>	Date YYYY-MM-DD	Date when the log is created (in UTC)
	<time>	Time hh:mm:ss	Time when the log is created (in UTC)
WHO	<user>	String username	Subjective user of the log
	<party>	String username	Party involved in the log
	<object>	String username	Objective user of the log
WHERE	<address>	String street_address	Street address where the log is created
	<latitude>	Double latitude	Latitude where the log is created
	<longitude>	Double longitude	Longitude where the log is created
HOW	<application>	String application	Service/application by which the log is created
	<device>	String device	Device with which the log is created

Table 3.2: "Common Data Schema of LLCDM". Reproduced from (Shimojo et al., 2010)

Listing 3.1: event.rb – Model Class for an Event. Reproduced from (Aiken, 2014)

```
module MyTimeline
  class Event < ActiveRecord::Base

    unless rails4?
      attr_accessible :description, :happened_on, :icon_name, :external_link, :original_id, :public, :importance
      attr_accessible :user, :linkable, :user_id, :linkable_type, :linkable_id
    end

    belongs_to :linkable, polymorphic: true, dependent: :delete
    belongs_to :user, class_name: MyTimeline.user_class.to_s

    validates :description, presence: true
    validates :happened_on, presence: true
    validates :importance, inclusion: {in: 1..10, allow_blank: true, message: "%{value} is not between 1-10." }

    scope :desc, order("my_timeline_events.happened_on DESC")
  end
end
```

Matching: Neither of these systems have any facility for automatic event clustering. The lifelog mashup API (LLAPI) (Takahashi et al., 2014) allows users to query the database of documents, in terms of any of the informational facets.

User Interface: Aiken's system displays a timeline, with the items grouped by day. Icons are used to distinguish items from different social networks. LLCDM does not provide a user interface.

These systems have shown how social network data from more multiple OSNs can be normalised into an event model, for subsequent processing – but without any event matching, and very little information extraction, the experience is limited.

3.4.3 Event-Clustering in GPS and *SenseCam* Data

The task of separating a sequence of (time, location) pairs into events represents an unembellished event clustering task, so is perhaps a good starting point for reviewing systems which perform event clustering. This is the task presented in the work of Gong et al. (2006): identify events in data from military reconnaissance patrols, based on GPS.

Source Data The dataset is a set of GPS documents, complete with timestamps – implying a natural ordering, with the implication that the event-detection task becomes a segmentation task.

Event Representation Events tend to have relatively short durations, such as "Getting out of humvee", and "the whole mission lasted 30 minutes and there are 1112 basic events". The scale of these events is much smaller than life story events, such as a holiday.

Matching Gong et al. (2006) propose a very simple metric for spatial-temporal clustering, similarity is binary: two documents are considered to be from the same event if both the distance and time difference are below their respective thresholds. These thresholds are computed for each dataset, and are simply a proportion of the maximum pairwise difference for the given event facet. A benefit of this simple approach is that it obviates any need for a second clustering step, as the clustering is completely determined by the document similarity metric. The dataset and the threshold-based clustering method mean that this system is arguably the closest approximation of a canonical event-detection system.

Another body of event-detection systems concern wearable cameras, specifically the Microsoft SenseCam. Worn round the neck, and equipped with sensors for GPS and light levels, which are used to trigger photographs, resulting in a large dataset of photos.

This presents a familiar research task: how to represent and organise such a large of body of content in a meaningful and navigable way? Organising this huge collection of photographs into events is the usual approach. The task eased by the availability of timestamps and GPS data; the task becomes one of segmentation and is relatively easy. But with so many events generated, the difficulty is in creating a summary of the event. Without text available, summarisation relies on the content of the image, so a key task is identifying the activity being undertaken based on the photographs; such high-level image processing is outside the scope of this thesis.

3.4.4 Photo Clustering

The previous chapter discussed the long history of photo management systems, with many of them being event-centric. The section examines the technical details of systems that perform this photo-clustering, with a focus on photos obtained from OSNs. Obtaining the photos from a social network means that metadata is different (perhaps more information) than in the case of collections of photos on a home computer. Hence, systems are able to both image content, and the metadata to perform event clustering.

Source Data As alluded to in the previous chapter, the choice of techniques (and their performance) must be considered in the context of the application and size of the source dataset and associated events: large/public or small/personal. Reuter and Cimiano (2012) obtain a large photo dataset from Flickr, using music events from *last.fm* as ground truth; large music events (such as concerts and festivals) are likely to have substantial content, uploaded by more than one user. Conversely, Rabbath et al. (2012) focus on personal events: their photo dataset is obtained from user's Facebook profiles, whereas Suh and

Bederson (2007) asked users to provide their own personal photo collections for the study. The number of photos per event would be expected to be different in each case – users might select the best photos for distribution on social networks, yet combining photos uploaded by many individuals would result in a greater number of photos.

3.4.5 Event Representation

Whilst the notion of the event is central to these systems, sometimes it is unclear how the implementation details relate to facets or aspects of events. The notion of the event is used throughout Suh et al.'s work (Suh and Bederson, 2007), examples given include a family birthday (lasting one day) and a camping trip (lasting five days) – users are left to choose their own ground-truth events, without the notion being explicitly defined (the work pre-dates Westerman and Jain's (2007) examination of the concept), but they do consider events as a hierarchy: the birthday event is comprised of two sub-events: a party and a family meal. Becker et al. adopt an event definition used in an earlier study on broadcast news:

> "An event is something that occurs in a certain place at a certain time." Becker et al., 2010

Yet, in this work (and subsequently (Reuter and Cimiano, 2012)), there is little mention of the representation of the event facets, with discussion instead centred around the metrics for comparing the documents (temporal, geographic, textual). Whilst temporal and geographic similarity metrics relate to specific event facets, textual similarity may encompass several.

The series of papers from Rabbath, Sandhaus and Boll clearly distinguish the event with its defining when/what/where/who informational facets, from the range of similarity metrics employed in their various experiments. In their work, an event is characterised specifically by the what/where/when facets (Rabbath et al., 2010, p. 2). Interestingly, they consider an event to be a special case of a more general 'social media story' class – where only one of these facets is defined, a more generalised representation of what is represented in a printed photo book, the original motivation for their study.

3.4.5.1 Extraction

Rabbath et al. (2012) explain their features in most detail. They use a number of global image features, adapting the approach of Redi and Merialdo (2011) combining 'coarsely localised information' to form a 'global, low-dimensional image signature' to represent the salient features of the image, alongside traditional representations such as colour and edge histograoms (section 3.3). Tags (i.e. people's faces), and other information is extracted from the metadata. After studying the face tags (by performing automatic facial recognition) they investigated the accuracy of the existing tags and the performance of facial recognition, finding that "84% of the tags matching some people are placed accurately". They also suggest that clothing can be used for event commonality:

> "For instance, if several people have photos with each of them wearing the same clothes in each photo, this considerably increases the probability that these photos belong to the same event, and the more people you have the higher the probability is." (Rabbath et al., 2012, p. 4).

They do this by taking a window around the tagged area and computing colour, edge, and texture histograms (section 3.3) for that area. Usernames of people tagged in albums, and attendance and temporal information from Facebook events are extracted from the metadata for using the matching metrics. Facial and clothing recognition is also used by Suh and Bederson (2007).

Whilst Rabbath's and Suh's research task centred around personal events, with a few photos, Reuter and Cimiano (2012) studies more public events with larger numbers of photos. This difference has led them to adopt different features and metrics for computing event similarity; much 'simpler' features can obtain high performance. Reuter et al. (Reuter and Cimiano, 2012) are able to achieve state-of-the-art performance without actually using any features based on image content (similar to Becker et al. (2012)). Instead they use features easily extracted from the metadata: the timestamp of the document, geographic latitude and longitude, and textual information (represented as a TF-IDF vector).

3.4.5.2 Resolution

NLP is usually not the primary focus of photo-clustering systems, despite the fact that textual information can play a more important role than image content in the task (Reuter and Cimiano, 2012). A typical approach would be to perform tokenisation (section 3.2.6.1) and then compute TF-IDF. The application of more elaborate techniques, such as the resolution of named entities, are hard to find in this kind of system. Rabbath et al. do not use geographic features[24], whilst Reuter only uses geographic features in metadata (ignoring what named entities might be available in the text). Suh et al. focus on people and events because Rodden and Wood (2003) found that "event" and "person" are the most frequently used metadata as well as chronological order of photos. When working with large collections of documents, one is able to select only the content which includes the requisite metadata; this is in contrast to the situation with personal photo collections, for which this theis believes it is important to extract the maximum amount of information. Herein lies one of the key motivations for the thesis – how can state-of-the-art photo event-clustering techniques be combined with state-of-the-art text event-clustering techniques?

3.4.5.3 Matching

In Rabbath's work, image content similarity is calculated by computing the Euclidean distance for the colour and edge histograms, with a specific algorithm used for their other global image features. They adopt an IDF-style metric for computing the similarity based on people.

Another feature relates to Facebook events: if two people mark themselves as attending (or perhaps attending) a Facebook event, who then both upload photos within five days, then photos may represent the same event, and the number of commonly-tagged people is used as a feature. It is unclear why they choose such a high-level feature – the temporal and person-similarity between the photo and the event could have perhaps been better represented by measuring the actual similarity in these dimensions, especially since the person-similarity is already represented by another event. Facebook photos do not include EXIF metadata, and the photo creation time is not available from the Facebook API, instead the upload time is used to partition Facebook albums into 'implicit events'; this assumption that if I upload two photos to an album in rapid succession seems to be somewhat dubious, as it is later used to construct the ground-truth for evaluation. Their overall similarity metric is

[24] Although place names are used in the construction of their ground-truth data

computed by binning values for each of the features, and then using a Bayesian approach considering the feature-bin labels independently for each feature – photos from within the 'implicit event groups' are used to compute the marginal probabilities.

Reuter and Cimiano (2012) uses Support Vector Machines instead of a Bayesian approach. The two key advantages of this approach is that there is no independence assumption among the features, and that unlike the discretised bin model used by Rabbath, the feature values are used directly by the geometric SVM approach.

Their equation conveniently normalises the temporal difference to the unit interval (recommended for SVMs) but the decision to compute the logatihrm is a little unjustified, and how it would interact with the choice of kernel function (which is not discussed). Great-circle distance is computed using the Haversine formula[25], and cosine similarity used for the various TF-IDF vectors. An SVM classifier is used to calculate the event commonality, configured to compute probability estimates (Chang and Lin, 2011, sec. 8). Event assignment is construed as a ranking problem: for a given document, candidates events are ranked by their match probability, the document is then assigned to the top-ranked event. However, if the match probability is below a threshold, a new event is created. The value of the threshold hyperparameter is itself computed by a SVM, based on features derived from the top-ranked match probabilities.

The temporal event aspect is key to the document event clustering task, various studies have seen how the timestamp alone (if accurate) can be enough to compute event commonality, without any additional information.

In Reuter and Cimiano's (2012), using time as the only feature yields an accuracy of 83%, boosted to more than 97% by the addition of geographic information or tag similarity. This is not unique to such large datasets either, Suh's study (on personal photo collections) uses temporal information as the basis for event clustering:

> "When there are interesting things around a user with a camera, the user tends to take a relatively large number of photos in a short period of time. Then, there is often a relatively long pause, followed by another burst of activity. Based on this characteristic, event boundaries are identified by detecting relatively long temporal gaps in a photo collection." (Suh and Bederson, 2007, p. 529) [26]

This is not the whole story: a key feature of photos is that their creation time will perfectly match the time of the event (which is not always the case with other media). This clearly is not the case for other types of document (such as Tweets). Even for photos, the creation time is not necessarily available. In the case of Flickr, the creation time is available in the metadata, effectively meaning the exact event time is available as a classification feature for all the content, whereas in Facebook (as in Rabbath's study) it is not, with drastic implications for performance.

3.4.5.4 Evaluation

Where large datasets are available, performance is high, with Reuter and Cimiano (2012) acheiving 97% on their ground-truth events, bootstrapped using *last.fm*. Rabbath's evaluation is based on using album names to construct ground-truth events (i.e. two friends

[25] The great-circle between two points on the surface of a sphere is measured along the shortest path on the surface of the sphere.

[26] Their study was actually investigating how a semi-automated annotation could assist users; it is included in this section because the techniques are still relevant.

with albums named after the same place, uploaded at a similar time), filtered depending on annotations that are available making it “hard to measure the exact precision and recall”. It is difficult to understand how well the testing scenario and data is representative of an event-clustering task, it seems that photos are only considered in the evaluation if they were taken at a location outside the user’s home country; to avoid such issues in *SAESNEG* ground truth events are collected directly from users (section 4.10). Their focus on holidays and events that are ‘album sized’ is appropriate for their photo-book application. It is unclear how well their approach could be adapted to different levels of granularity.

3.4.5.5 User Interface

The studies discussed in this section are demonstrating photo-clustering techniques, rather than focusing on the user interface and application. Rabbath’s study is used to drive a Facebook application with a virtual-photobook style interface. Suh and Bederson (2007) built a desktop application for managing personal photo collections.

3.4.6 Event Detection in Microblogging

This section examines examples of systems which detect events in micro-blogging services (such as Twitter). Clearly, they differ from photo-clustering applications in that the focus is on using NLP techniques to detect the existence of an event, and/or compute event commonality. Detecting events from the public Twitter ‘firehose’ requires handling a huge volume of data, and perhaps a requirement to work with the stream of data in real-time. This is a very different research task than trying to cluster a static dataset of a few thousand photographs. A key challenge of micro-blogging is that, unlike photos, the content creation time does not necessarily match the time of the related event; the temporal aspect, shown to be incredibly useful for photographs, is not available in the same way.

3.4.6.1 Source Data

Focusing on ‘public’ OSNs means that more data tends to be available, so as long as techniques have good precision, they are likely to produce (visually) impressive results. Conversely, in the context of the personal social networking footprint, especially with ‘average’ users (rather than enthusiasts), there might be less data available, a high precision and recall is necessary – this distinction is easily overlooked, but is hugely important as it greatly affects the suitability of techniques and their performance. This phenomenon is visible in the vocabulary used, discussions of ‘background noise’, ‘social signals’, with signal processing terminology and techniques being employed. Such techniques will clearly be unsuitable for a situation where a user makes a few status messages each month.

Indeed, micro-blogging services, public by nature (and hence unencumbered with privacy issues, making lots of data available) have been a popular medium for event-related systems. Twitris (Nagarajan et al., 2009) is a widely-known example of such a system, with Tweets considered as “citizen observations” with the rationale of “spatial, temporal and thematic integration” of those observations, to see changes over location and time.

In such systems, fetching source data is a typically a two-step process. The scope of input data is either specified manually by the user, or determined automatically in a separate step. This two-step process is closely tied to the event representation. Twitris (Nagarajan et al., 2009) uses seed search terms, bootstrapped using Google Trends[27] (formerly known as Google Insights), which are manually verified for suitability, before being used to search the Twitter API. Nichols et al. (2012) began by using the Twitter API to search for Tweets relating to the 2010 world cup, the 'event detection' was actually about detecting events occurring within that context, such as goals, which are arguably sub-events of the World Cup event.

Steiner et al. (2012) automate and generalise this procedure (they developed a Chrome web application called NiteOutMag for the collection of public social media data relevant to nightlife events, presented in the form of a virtual magazine). In their study, a number of 'event search APIs' are searched for recent event occurring within 5km; the results from this search are used to search other social networks for content relating to those events.

3.4.6.2 Detection

Systems for detecting events via Twitter commonly look for 'spikes' in activity (i.e. changes in the rate of Tweeting for a particular topic); TwitInfo (Marcus et al., 2011) for example uses a TCP congestion algorithm, and is able to operate in real-time, whereas Nichols simply uses the gradient of the Tweet rate, based on earlier work (Shamma et al., 2009).

'Noise' is often mentioned as an issue: Tweets that either do not relate to the intended topic, or that they do not relate to the event. TwitInfo uses a TF-IDF based approach to exclude content relating to globally popular topics (e.g. Justin Bieber), to prevent them from swamping the target context. Such high-volume detection algorithms are inappropriate for the task of document event clustering on a much smaller personal social media footprint, where the volume is much smaller, events may have only a few related documents.

3.4.6.3 Representation

As many of the systems detect events according to volume alone, the representation of the event is concerned with how to summarise the content: the evolution tracking representation of Lee et al. (2013), or the phrase-graph technique used by Nichols et al. (2012) to generate a summary sentence. More relevant to this thesis is the work of Ritter et al. (2012), with their 'open domain' event extraction pipeline represent events as a 4-tuple: named entity (e.g. 'iPhone'), event phrase ('announcement'), a date, and an event-type. The techniques used to extract these representations is discussed in the next paragraph. Steiner et al. (2012) adopts an event model that is more familiar – the event defined at `Schema.org`.

3.4.6.4 Extraction

For systems which use 'spikes' to detect events, information extraction is more concerned with representing and summarising the event than actually detecting it. The NLP pipeline of Nichols et al. (2012) can be outlined as follows:

1. Use Apache Nutch to detect language (to filter non-English Tweets) (the authors note that Tweets marked as English in their metadata are often not actually English)
2. Use heuristics to extract sentences

[27] `http://www.google.com/trends/`

3. Handling of repeated characters
4. Stop Word Removal
5. Keyword-Based Spam Filtering
6. Word Stemming using the Porter Stemmer
7. Building a 'phrase graph' (Sharifi et al., 2010) to summarise the event.

To track event evolution, Lee et al. (2013), as a compromise between a bag-of-words model and a running NER, (where the person/location/organisation categories are seen as too limited), and with the unnatural language, instead use all the nouns POS-tagged as NNS or NNPS (plural nouns), which are then stemmed. The rest of their pipeline is based on these tokens, citing previous work demonstrating that nouns are more closely related to topic than other parts of speech (Liu et al., 2003).

Because of the abundance of content in these 'public' datasets, systems can be selective with content, whilst not being overly concerned with extraction. In Twitris, the location in the profile is taken to be the location of the Tweet, because of "the lack of geo-coding information in the Tweets". Temporal information can be directly extracted from the metadata (hence, where a Tweet refers to an event at an earlier or later time, this would be excluded) – instead of extracting named entities from the Tweets themselves.

Ritter et al. (2012) take a different approach in their 'open domain' event extraction: they construct an NLP pipeline to extract events directly from Tweets. Having noted issues with unnatural language (section 3.2.4), their system extracts a 4-tuple representing the event, with a pipeline that includes:

1. POS Tagging – A 'Twitter-tuned' POS tagger.
2. Extraction of Named Entities – using a Twitter-specific NER-tagger from their previous work.
3. Extraction of Event Phrases – using a CRF-based sequence-labelling approach.
4. Extraction of Temporal Expressions – using TempEx (Mani and Wilson, 2000).

The classes of 'event types' are then determined by Latent Dirichlet Allocation, using LinkLDA (Liu et al., 2009).

3.4.6.5 Resolution

Steiner et al. geo-code the places mentioned in the first phase of their system, and use the latitude and longitude co-ordinates for the second search phase (rather than detecting named entities).

3.4.6.6 Matching

Systems which detect events as 'spikes' in the Tweet rate do not appear to perform document event matching in the sense that the photo-clustering systems do. However, a more salient point is that they re-enforce the key importance of the temporal event aspect. Such systems rely almost exclusively on temporal similarity for event detection, as was the case in the large-scale photo clustering systems, such as the study of Reuter and Cimiano (2012) (this work is discussed section 3.4.4). Harnessing the power of the temporal aspect in the context of the more restricted personal social media footprint will need to be a key focus of *SAESNEG*

In Lee et al.'s (2013) study, post similarity is dependent on topic similarity (in this case, nouns), and temporal similarity. They decide to combine these two measures by calculating a quotient dividing a set-based similarity measure (e.g. Jaccard Similarity) over the unit interval, with a monotonically increasing function of the time displacement (e.g. the exponent). They call this quotient the 'fading similarity'. Edges exist in their post graph when the similarity is greater than some threshold.

Clustering is achieved by identifying core posts (a cluster is the neighbourhood of a core post), thresholding on a measure of graph centrality, computed with a similar quotient to that used edge weights, where the denominator is instead a function of the age of the post, rather than the time displacement along the edge. The events are the core posts and their neighbourhood, with posts not in any such neighbourhood considered 'noise'. This approach allowing the evolution of the event clusters to be described in terms of operations to the graph, with proofs given for various results concerning this.

3.4.6.7 User Interface and Evaluation

Many of the systems studied in this section are intended to summarise events, and it is the summaries that are evaluated. Nichols et al. (2012), in their study of the 2010 football World Cup matches, evaluate the generated summary sentences alongside human-generated descriptions (using various ROUGE metrics (Lin, 2004)), and ask users to rate the quality of the generated text. They also evaluated the recall of events themselves; the system performed well with events such as goals (which generated a large number of Tweets), but recall was lower for other 'events' such the awarding of a yellow card (which perhaps generated less Twitter interest).

The nature of these systems with 'events' being 'discovered', can make evaluation difficult, as no obvious ground-truth exists. Nagarajan et al. (2009) present example output, rather than any formal evaluation. Marcus et al. (2011) present graphs of the event 'spikes' detected, with various charts to summarise the results from sentiment analysis. Steiner et al. (2012) get user's to evaluate the experience of their user interface (and the gathered content presented by it), finding that it varied according to the availability of the source data.

3.5 Conclusions

The majority of methods surveyed here have been developed and evaluated only on one kind of social media (e.g. Twitter or blog posts). Cross-media linking, going beyond connecting Tweets to news articles, is a crucial open issue, due to the fact that increasingly users are adopting more than one social media platform, often for different purposes (e.g. personal vs. professional use). In addition, as people's lives are becoming increasingly digital, this work will also provide a partial answer to the challenge of inter-linking our personal collections (e.g. emails, photos) with our social media online identities. The challenge is to build computational models of cross-media content merging, analysis, and visualisation and embed these into algorithms capable of dealing with the large-scale, contradictory and multi-purpose nature of multi-platform social media streams. For example, further work is needed on algorithms for cross-media content clustering, cross-media identity tracking, modelling contradictions between different sources, and inferring change in interests and attitudes over time. (Bontcheva and Rout, 2014, p. 4)

Chapter 4

Methodology

> “Instead of having a limited number of significant items left by the deceased, and perhaps a few stories from them or from the people that know them, we are likely to have far too many details of their life, the mundane mixed up with the exceptional....Search engines do an incredible job pulling items from billions of webpages based on short search terms. The kind of ingenious thinking that makes this possible could be brought to bear on personal content; some breakthrough piece of software could do the work of extracting the significant from the everyday. **This is not a trivial computer science problem, however**.” (Banks, 2011, p. 76)

4.1 Introduction

This chapter introduces *SAESNEG* both as an experiment and as a framework for data-mining and clustering of social media documents. To recap, the system takes as input a set of social networking documents of various types (photos, check-ins, status messages), and needs outputs a partitioning of that input set into event clusters (the so-called *life-story* of the user, denoted $L_C(u)$ in section 1.4). As set out in the formal problem statement (section 1.4) the events are sets (or “clusters”) of social media documents. For each event cluster in the output, the intention is that all documents contained within it relate to the same underlying real-world event, and thereby serve as a basis for navigation in user interface. The overarching experiment investigates whether the system (*SAESNEG*) can be trained to successfully perform this event clustering task, using the collected ground truth events as training data, and do so within the novel corpus of the personal social media footprint. It is argued that the personal social media footprint is a heterogeneous dataset, with its private events, creates novel challenges for the study (to summarise section 1.5):

- It is argued that the source data is *differently heterogeneous* (section 1.6): documents contain a mixture of text, image, and metadata (i.e. heterogeneous). However, the different types of documents have different schemas, and mix the kinds of data in different ways, the data having different semantics depending on the enclosing type (see section 1-the social media footprint).
- Private events tend to be sparse (there may be only one or two documents in an event).
- Indeed, using private data means that less data is available overall than might be the case with a public dataset – participants need to be individually recruited.

- The nature of private events mean that there is little or no external knowledge that can be cross-referenced or used as a ground-truth – hence, less training data is available for machine-learning algorithms.

SAESNEG has been developed as a response to these challenges – affecting its architecture and algorithms, especially in NLP and machine-learning. Ultimately, we will show that *SAESNEG* meets the challenges presented by this dataset (for the event clustering task); it is the overarching technical contribution of the thesis, and is returned to throughout the remaining chapters.

This chapter will explain in detail the experiment that was undertaken, the use of ground truth event clusters for training and evaluation, and how the *SAESNEG* was implemented to perform the experiment. In the next section, the experimental setup is introduced with a workflow to present a simplified, logical view of the core pipeline (with its two distinct phases), with a description of salient features, and the logical flow of data between the various steps. This is followed by a discussion of methodological issues of participant selection and collection and analysis of source data for the study (section 4.3).

After a detailed physical architecture (figure 4.5) for the entire system (including ancillary components to support the experimentation) is presented with discussion in section 4.5; the remainder of the chapter is organised according to the components and sub-systems. The techniques for extracting event-related information; and the strategies for comparing the information to estimate event commonality, are performed to support the overarching event-clustering experiment.

Chapters 5 and 6 focus on Phase A (extraction) and Phase B (clustering) respectively, including algorithms, implementation, and associated experimental results. There is no separate results chapter – the overall clustering performance is evaluated in chapter 6, and an extenaive sample of social media data is imcluded in (appendix B). Instead, this chapter focuses on the overall architecture, and components not discussed in chapters 5 and 6: notably the ground truth event web interface (sections 4.10 and 4.4), storage, and ancillary components.

4.2 Experiment Overview

4.2.1 Introduction

This section introduces the experimental pipeline at the core of *SAESNEG* and the salient features of the architecture. The aim is to convey how the social networking data logically moves through the main sub-systems of *SAESNEG* to perform the experiment, and give an overview of how both the architecture and components address the unique challenges created by the source dataset.

Figure 4.1 shows a flow chart with the key components of the core pipeline comprising the experimental setup. The core pipeline that performs this clustering is shown on the left hand side of the diagram, and is divided into two phases: extraction of annotations from individual documents, and comparison of documents' annotations to perform event clustering.

Ground truth event clusters (i.e. sets of documents as defined in section 1.4) are created by the users by partitioning their own source data. These event clusters are ultimately used for training the ML-classifier used in the clustering step or evaluating the performance

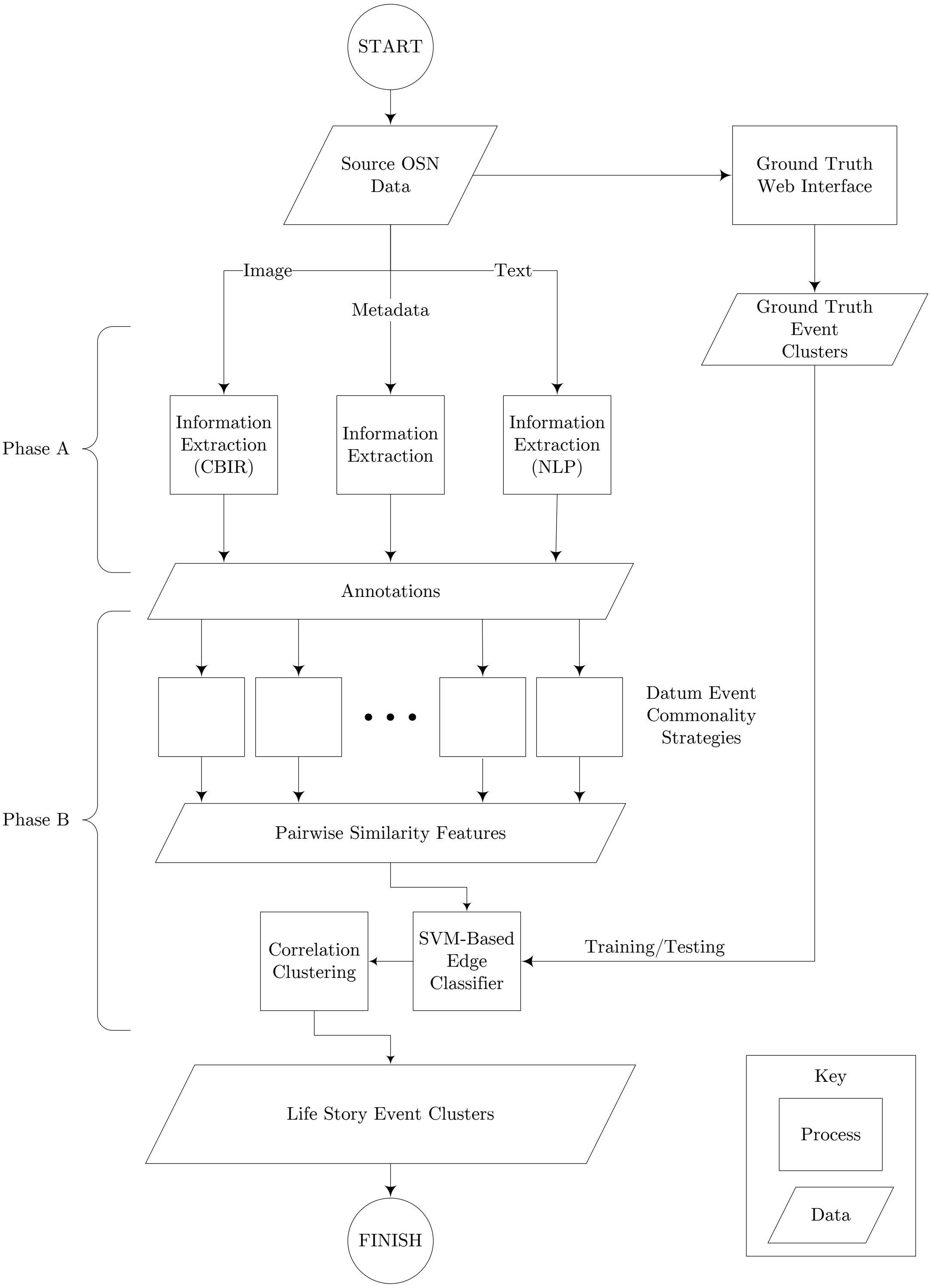

START
Source OSN Data
Ground Truth Web Interface
Image
Metadata
Text
Ground Truth Event Clusters
Phase A
Information Extraction (CBIR)
Information Extraction
Information Extraction (NLP)
Annotations
Datum Event Commonality Strategies
Phase B
Pairwise Similarity Features
Correlation Clustering
SVM-Based Edge Classifier
Training/Testing
Life Story Event Clusters
FINISH
Key
Process
Data

Figure 4.1: *SAESNEG*: workflow illustrating the core pipeline

of the pipeline, according to a cross-validation scheme typically used in machine-learning experimentation.

Due to the complexity of the complete system, a more detailed physical architectural diagram (figure 4.5) is presented separately, in section 4.5, showing numerous additional components.

4.2.2 The Experiment

As encapsulated in the formal problem statement (section 1.4), the system is intended to partition a set of social network documents (a mixture of photos, check-ins, status messages comprising a user's social networking footprint), into event clusters. This overall clustering task is evaluated by measuring the clustering performance – to the author's knowledge, this is the first study to investigate the document event clustering task on this corpus. The experiment investigates whether a system organized in this way can be trained to perform this event-clustering task. The system performs the clustering in two stages: extracting features from the source documents (Phase A) and machine-learning based clustering (Phase B); as explained below.

The collection of ground truth event clusters (and analysis of the resulting data), is a contribution in its own right, supporting the hypothesis of sparsity for some private events, and that events can contain multiple data types. This is evidenced through simple statistical analysis and is presented in section 4.4. The ground truth web interface also serves as an experiment into how ground truth of this nature can be collected, and whether the novel user interface developed for this study is a successful approach.

Having established the baseline performance of the pipeline, various information extraction techniques have been evaluated by measuring their effect on the performance of the pipeline (Phase A, chapter 5). Similarly, the comparison strategy and clustering algorithms in Phase B (chapter 6) can be evaluated for their contribution to the document event clustering task. With a multitude of tools for inspecting data relating to the various components, *SAESNEG* is used a framework for the development of techniques and algorithms, as well as their experimental evaluation.

4.2.3 The Two-Phase Architecture

The rationale for this architecture is broadly as follows: as discussed, the task is to cluster social media documents into sets representing real-world events. To achieve this, the system needs to compare documents to see if they indeed relate to the same event. The approach adopted is to compare pairs of documents to compute pairwise similarity, before computing the event clusters. To compute this similarity, a number of pairwise strategies are proposed (discussed extensively in chapter 6). These strategies rely on information extracted from the documents themselves. For simplicity, this extraction phase (A), can be performed on documents individually – a range of techniques for data mining the text, images and metadata are discussed in chapter 5. The interface between these layers is a system of annotations, which allow the event commmonality strategies in Phase B to reason about event commonality independent of the documents and data from where the annotations were extracted. These annotations allows documents of different types (e.g. photos, status messages, events) containing different kinds of data (images, text, metadata), relating to different event facets (temporal, spatial, social, etc.) to be compared freely. This abstraction thereby overcomes the challenge of the *differently heterogenous* (see section 1.6) characterising the source data for this study. The sample of users social networking footprint (see appendix B); organised into events by the users can then be used both to train and evaluate the system, annotations can be extracted from the sample documents, with the ground truth event clusters inducing a labelling on the document pairs (i.e. intra- and inter-event document pairs) – which can be used to train and evaluate the system's ability to compute the event clusters itself.

This two-phase architecture is also used to organize the next two chapters in the thesis:

Phase A – the extraction of annotations, operating on each document independently (chapter 5). Initially, text and image content are extracted from the metadata. For each of these kinds of data, appropriate information extraction techniques (as reviewed in chapter 3: NLP, CBIR) are employed to extract annotations broadly corresponding to who/what/where/when event facets, including image content descriptors. These annotations inform document event commonality decision making in Phase B (i.e. whether documents relate to the same real-world event, sharing that event "in common").

Phase B – partitioning into event clusters, by comparison of the annotations (chapter 6) extracted in Phase A. A set of independent algorithms, known as the *document event commonality strategies* examine each pair of documents – to see if the annotations indicate event commonality. For example, the spatial strategy (section 6.4) examines location annotations to determine whether there is geographical proximity (which may indicate event commonality). Each of these strategies output a vector of real-valued features for each pair of documents (concatenated into a single vector), with machine-learning algorithms used to make the final judgements regarding event clustering.

Chapter 6 describes the detail of Phase B, including how the ground truth event clusters are used both to train the machine classifiers for clustering, as well as the evaluation of performance (with discussion of the various issues and approaches to measuring the 'accuracy' of a clustering).

4.2.4 *SAESNEG* and the Challenges

As discussed in the introduction section (section 4.1), how the challenges of the source data are met by *SAESNEG* is a crucial argument in this thesis, and some of the ways this is achieved in different parts of the framework are introduced below.

In Phase A, because of the low number of documents in each event (one aspect of sparsity), to make informed judgements about event commonality, the system attempts to extract as much useful data as possible from each social networking document. This entails combining different techniques to match different kinds of data: NLP and CBIR for information extraction, together with business logic for extraction of annotations from metadata, as shown in the detailed architecture diagram shows (figure 4.5).

In a departure from existing studies, these two phases are purposefully decoupled, by means of a carefully designed set of abstract annotations. It is argued that this layer of abstraction makes best use of the extracted information being independent from the underlying kinds of data and types of document, allowing strategies for event clustering decision-making to be applied broadly. For example, the location annotations may have been derived from metadata (i.e. geotags) or textual information (mentions of a place by name). Without such abstraction, annotations would be incredibly sparse, and comparison between documents (especially those of different kinds) would have been problematic.

A related implication of decoupling strategies from extraction is that dimensionality of the feature set generated by the document comparison strategies (in Phase B) used for final clustering are thereby reduced, thus obviating sparsity for machine-learning and avoiding the so-called *curse of dimensionality.* These layers of abstraction have been manually designed (unlike other studies; where dimensionality is reduced using techniques such as LDA to map feature vector spaces to their latent subspaces); with the advantage that features retain semantic meaning, allowing further engineering of semantically high-level features, based on common-sense intuition.

Additionally, because there is no external source of ground truth for event clusters, the architecture needs to include a component that allows users to create them from the source data; such ground truth is crucial for training the machine-learning components, and evaluating the pipeline.

4.3 Participants and Source Data

4.3.1 Introduction

This section discusses the selection of participants and provision of source data.

It discusses how source data was collected from online social networks (OSNs), for a sample of OSN users. This data was used for the development of the pipeline, the individual extraction techniques (chapter 5), and the overall document event clustering experiment (chapter 6). Discussion includes the selection of participants, and key decisions about what social networking data is used in the experiment.

The scope of OSN data collected from each person in the sample was a key concern, the primary application of the developed algorithms being to facilitate an enhanced user experience with an easily-navigable overview of an archive of a user's social media footprint – hence, it was important that the social networking data used in this experiment is a representative sample of that data, and one that exhibits similar characteristics.

There were three three decisions:

1. Who will participate in the study?
2. Which social network(s) to use as sources of data for the study?
3. What data should be collected from the OSN for the study? (see section 4.8)

When making these decisions (and in the wider experiment), working with such personal data meant that ethical concerns were paramount.

4.3.2 Participant Criteria

For an ethically and methodologically sound study, there were a number of criteria for participation:

- The participant must have been a user of online social networks.
- The participant must have been able to communicate sufficiently well in English to give informed consent to the study, and to participate in questionnaires, interviews etc. as necessary for subsequent research objectives.

Study	N$\bar{o}$ of Participants
Wang (2012, p. 4)	13
Zhao et al. (2013)	13
Gouveia and Karapanos (2013)	12
Mei et al. (2006)	10

Table 4.1: Number of participants in similar studies.

- The participant must have been an adult (simplifies various ethical issues, and ensures participants are capable of understanding the study).
- The participant must have given informed consent for the study.
- The majority of the participant's social networking content must have been in English.

4.3.3 Number of Participants

Ultimately, the quantity of data available to the experiment depends not only on the number of participants, but also the amount of data associated with their social networking profiles, and the extent to which they contribute data to the study i.e. the extent to which participants are willing (and able) to use the ground truth interface, little of which was known when designing the study. Similar existing studies offer an indication of what is considered a reasonable number of participants – table 4.1 shows the number of participants in similar studies.

A key argument in this thesis is to investigate strategies for overcoming the sparsity of the source data, and is one of the distinguishing features of this work. This thesis argues that research in this area is dominated by the use of large, public datasets with pre-existing external ground truth. Research tasks where large datasets are not available; or where no independent ground truth is available can be overlooked. The thesis thus serves as a case study for the application of machine-learning to a data-mining problem when there is *not* an overwhelming volume of data available; and how the techniques and algorithms can compensate for this.

4.3.4 Recruitment of Participants

A decision was made to recruit Cardiff Metropolitan University computing undergraduates as participants; the rationale being:

- They were likely to be active social networking users.
- The study could be explained in person, demonstrate the use of the ground truth web interface in class, and be on-hand to answer any questions, or deal with bugs.
- They were adults, and good English speakers, so able to give informed consent.

Recruiting participants online, for example, would not have had these advantages.

OSN	Main Document Type(s)	Kind(s) of Data	Number of Users[1]
Facebook	Various: e.g. Photo, Status Message, Check-In, Event, extensible through OpenGraph[2]	Varying mixture of image, text, and metadata.	1,280,000,000[3]
Twitter	Tweet	Text, metadata, sometimes image.	255,000,000[4]
Foursquare	Check-in	Metadata, some text, optional image.	50,000,000[5]
Flickr	Photo	Image, metadata, some text.	32,000,000[6]
LinkedIn	User's profile is divided into "fields": work experience, qualifications, skills, etc.[7]	Text / metadata	300,000,000[8]
Google+	Various and extensible through http://schema.org markup[9].	Varying mixture of image, text, and metadata.	540,000,000[10]

Table 4.2: Examples of popular online social networks and associated data for the user's social media footprint

4.3.5 Choice of Social Network: Facebook

The second key choice was to decide which social network(s) to use for the collection of source data for the study. There is great variety in the kinds of data, and type of documents available in different social networks, as table 4.2 shows.

From table 4.2, it is clear that the typical social networking footprint (if a user is a member of several of these networks) is a differently heterogeneous dataset (section 1.6), but which ones could be chosen to for use in this study. Key concerns were:

- The popularity of the OSN – what proportion of the participants were (active) users of it?
- What types of document/kinds of data were available – the aim being to create a differently heterogeneous (section 1.6) sample of the footprint for each participant.

[1] Note that so-called 'active' users will be significantly less than the number of registered users, hence the table is unordered.

[2] https://developers.facebook.com/docs/opengraph

[3] Monthly active. March 2014. http://newsroom.fb.com/company-info/

[4] Monthly active. July 2014. https://about.twitter.com/company

[5] Registered. May, 2014. https://foursquare.com/about

[6] Registered. May 2009. https://www.flickr.com/help/forum/en-us/97258/

[7] https://developer.linkedin.com/documents/profile-fields

[8] Registered. July 2014. http://press.linkedin.com/about

[9] https://developers.google.com/+/api/moment-types/

[10] Monthly Active. October 2013. http://googleblog.blogspot.co.uk/2013/10/google-hangouts-and-photos-save-some.html

- Practical concerns: is there an API? is it well-documented? Is the OSN highly likely to remain active for the duration of the study?

After consideration, the decision was to initially use Facebook as the exclusive data source for the study. The rationale for this decision is was as follows:

- Facebook is popular (see table 4.2), it was highly likely that most would-be participants would be members of the OSN – and that a suitably large sample of data would be available.
- The Facebook API is well-documented, stable, and has good support community.
- Facebook was not likely to be shut-down or go bankrupt during the study!
- Facebook has a broad mixture of different types of data – by supporting a single OSN, it would be possible to gather a sample of a range of different types of document, creating (it is argued) a representative sample of the various types of document's found in a user's social media footprint, e.g. status messeages perhaps representing the user's Tweets.
- It was felt that the mixture of types in Facebook (five types of document were fetched from Facebook for this study, see table 4.4) was sufficient to demonstrate the application of *SAESNEG* to differently heterogeneous data (section 1.6) – it was found later that many of the ground truth event clusters did indeed comprise multiple types of document.
- Support for additional networks could be added later if needed.

If the use of Facebook was somehow unsuccessful, other OSNs could easily be integrated into the system, the system is highly extensible, and could easily be re-targeted to a new OSN. The system is designed from the ground up to support multiple social networks – dependency on the particular Facebook types is limited to where it is necessary: the fetcher, and the initial stages of Phase B – otherwise, the system is not dependent on any particular social network.

4.3.6 Summary

This section has discussed issues relating to the selection of participants, and the choice of using Facebook as the exclusive source of social networking data for this study.

In short, using Facebook exclusively is sufficient to demonstrate *SAESNEG* for the purposes of supporting the thesis (event clustering on a differently heterogeneous social networking data (section 1.6)) – whilst the system itself was been designed with support for additional networks from the beginning. Event clustering across multiple social networks would be relatively easy to implement, and, subject to selection of suitable participants, is a key opportunity for future work.

Note that ethical issues are discussed together with the electronic consent form in section 4.6, whilst the collection of ground truth is discussed separately in section 4.10. The next section introduces a detailed architectural diagram, serving as the means of organising the remainder of this chapter.

4.4 Analysis of Collected Events

This section analyzes the collected ground truth event clusters (the set of events for each user u, respectively comprising the ground-truth life-story $L_{GT}(u)$, for each user, section 1.4). Furthermore, it seeks to evaluate the ground truth web interface developed in this study, through analysis of the data collected from it – finding it to be broadly successful. The results of analysing the ground truth event clusters serve to confirm earlier hypotheses regarding the characteristics of the source dataset – its differently heterogeneous nature (section 1.6), the heterogeneity and sparsity of the events found in the personal social media footprint (section 1.7). Hence; these findings support the key motivation for the thesis – the need for a novel solution to perform document event clustering on the social media footprint, and calls for greater attention to be given to this research task.

4.4.1 Evaluation of User Participation

As shown in table 4.3, a total of 15 participants successfully used the interface to create ground truth event clusters, whom generated enough ground truth data for the experiments (this figure is similar to other studies – see table 4.1). However, the percentage of users who registered for the study, and who went on to create ground truth event clusters was 21% (and indeed, most participants did not create more than a few clusters (see figure 4.3). The reasons for this are unclear, since the system was also being used for other studies, and some of the users appearing to be spam accounts.

Table 4.3: Summary of participants' ground truth data.

Participants	82
...who created ground truth event clusters	15
Total ground truth event clusters	166
Mean clusters per user	2.02
Total documents in ground truth clusters	583
Mean documents per cluster	3.51

A possible explanation is that users did not want to sit through the instructional video, but there needs to be a balance between the quality and quantity of training data – users creating event clusters which were incorrect (having misunderstood the event clustering goal) could have been highly disruptive to the experiment[11]. Future work would seek feedback from both participating and non-participating users, and improve the service accordingly. Consideration could also be given to making a new instructional video of a shorter duration.

4.4.2 User Creation of Clusters

Even with a carefully designed user interface, further statistics relating to data collection highlight the effort required to create ground truth data of this kind. Focusing on the users who did participate in the study, the mean number of ground truth clusters per user was 2.02 (see table 4.3). Looking at figure 4.2 it seems many users created a low number of events. Potential explanations are either that the sample of 50 documents contained a small number of recognisable event clusters; and that the remaining documents either did not

[11] Skewing the study towards more technically competent users is another potential pitfall.

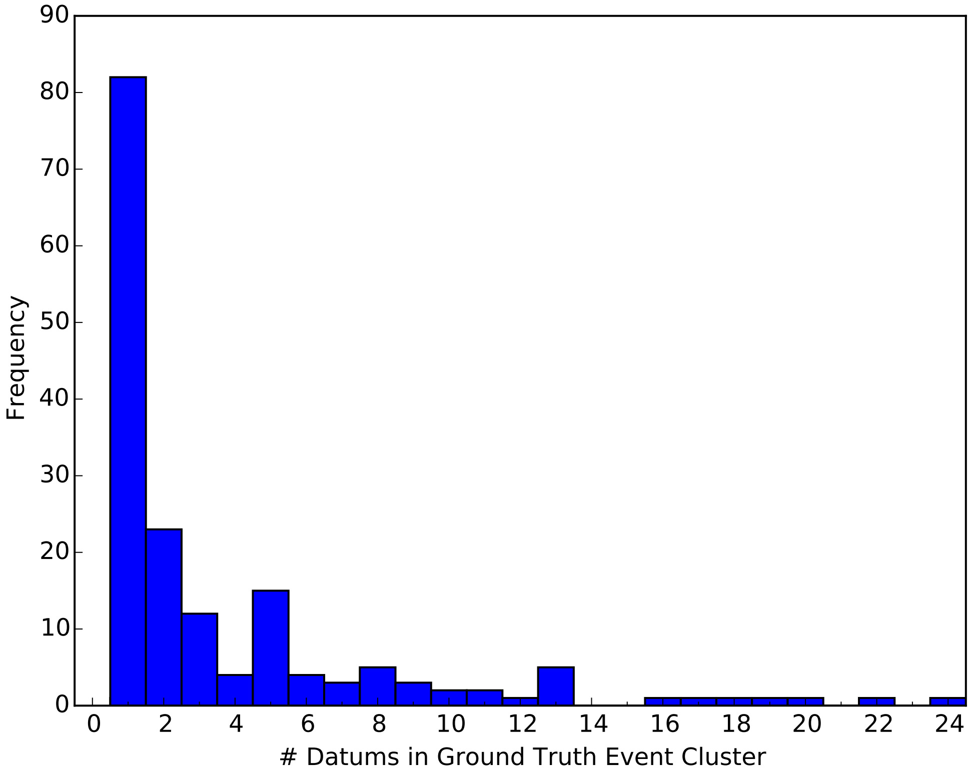


Figure 4.2: Sizes of Ground Truth Document Event Clusters

relate to an event (or were not thought of as doing so by the participant), or that there was no further event commonality. Problems of recollection may have been another issue – and users may simply have got frustrated from using the interface and given up after a creating one or two events.

Despite these issues, the data collected was sufficient for the study: the total number of documents included in the created event clusters was 583 – a sufficient sample to demonstrate that Phase B operates correctly on a range of source data, and for Phase B, a total of 128,142 edges between documents (counting both intra- and inter-document edges) can be used for training and evaluation, as shown in table 4.3, for the 166 ground truth clusters created by the participants.

4.4.2.1 Size of Clusters

Figure 4.2 shows that the size of clusters (the number of documents they contain) tends to be small, with a mean of 3.51 (table 4.3), this finding has a number of implications:

- It could indicate a sampling issue – events are actually larger, but the scope of source data was too small to allow the creation of representative event clusters. Given that a sample of 50 was used, with documents created at similar times, this is unlikely – the largest event contained 25 documents.
- The data confirms the sparsity of events in this context – that a significant proportion of the events contained only one or two documents, is a stark contrast to studies on flickr data, and highlights the unique challenge of this study.

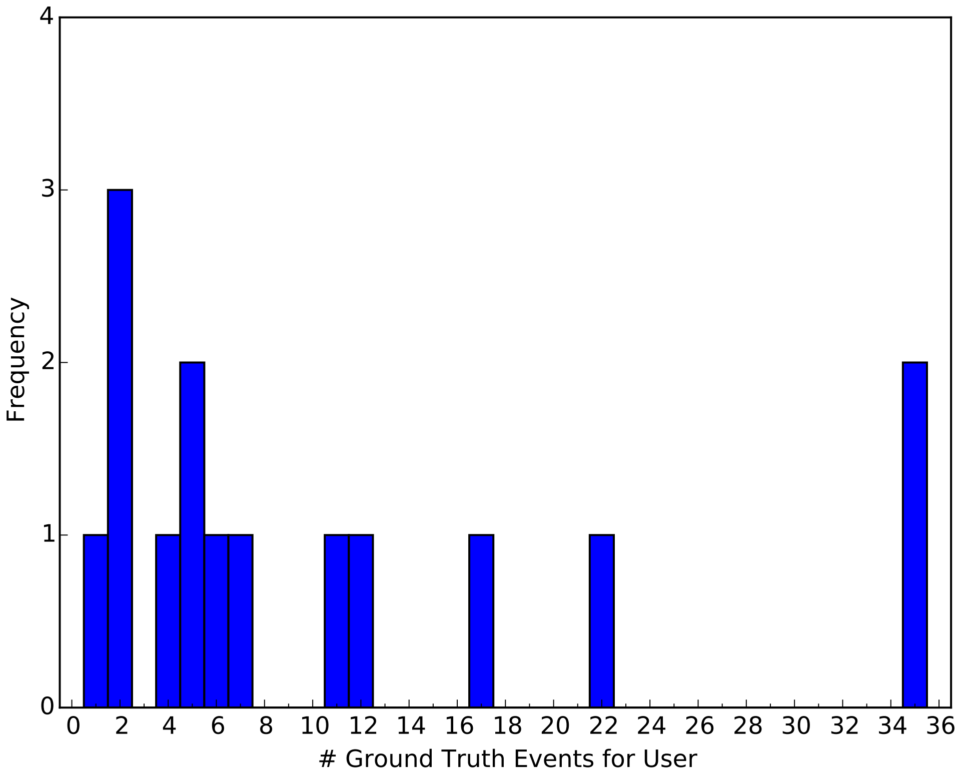


Figure 4.3: Number of Ground Truth Events Created by Users

- At first glance, the presence of many small events (only a few documents) may appear to undermine the key argument that an event-centric organisation of data. Mathematically, even clustering the documents into small events dramatically reduces the number of items on the timeline – events with two documents would half the number of points on the timeline. Even a handful of larger (e.g. 5, 10 or even 25 documents) events would significantly ease navigation by "de-cluttering" the user's timeline. When applied to many years (and ultimately, a lifetime) of social networking data – the resulting reduction would be significant.

Future investigations could build on this study:

- Integration with additional social networks (and perhaps other sources of data) may result in larger event clusters – as more apps and services (whether they are ostensibly relate to lifelogging or not) are used to any ever greater extent, the likelihood of document event commonality would likely increase.
- A future iteration of the user interface should consider this in the design, and perhaps implement a command "all remaining documents are relate to different events", or a one click "assign to singleton event" button.

4.4.2.2 Cluster Heterogeneity

The results also confirm that event clusters contain various different kinds of documents. Table 4.4 shows the variety of different types of documents that were included in the ground truth event clusters (indeed all of those fetched as primary documents featured in the clus-

Table 4.4: Frequency of Document by Type

Type	Frequency
Status Message	27
Photo	438
Check-In	24
Facebook Event	94
Total	583

Table 4.5: Ground Truth Cluster Documents by Type

Type(s) in Event Cluster	Frequency	
Photo	78	46.99%
Facebook Event	54	32.53%
Check-In	15	9.04%
(Mixed)	12	7.23%
Status Message	7	4.22%
Total	166	

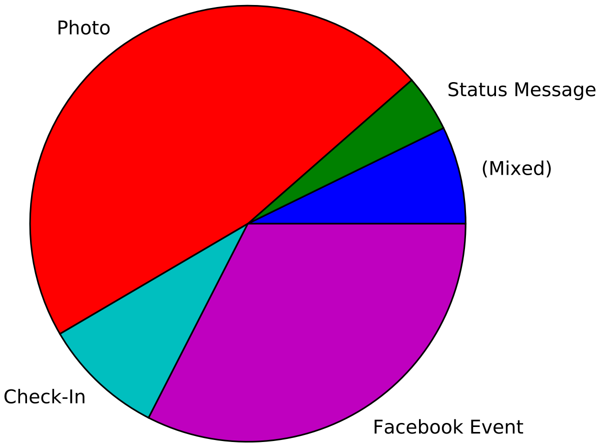


Figure 4.4: Ground Truth Event Document Composition by Type

ters). This result demonstrates that photos feature heavily in events, this could be due to a combination of factors, as photos may have been:

- perceived as being more valuable, or representative of events.
- more likely to relate to an event (or be perceived as such – the relationship between events and other types of document may be less obvious to users, especially because they can be created at different times – this view is confirmed by the popularity of the *Facebook event* type in the event clusters.)
- more common in the footprint.
- easier to recollect; and thus cluster correctly.

Although photos commonly featured in event clusters, they represented less than half the documents incorporated into events, other types of document deserve research attention (notwithstanding work on clustering of Tweets). Indeed, in terms of the composition of the event clusters, table 4.5 and figure 4.4 shows that many of the events contained a mixture of different types of documents, and demonstrates the importance of cross-type document event clustering – even with a small sample of the social media footprint, based on a single social network, nearly 8% of the events contained more than one type of document.

For comprehensive clustering of the personal social media footprint, research cannot overlook the importance of non-photo document types, and whilst research continues on Tweet/status message event clustering (Nichols et al., 2012; Marcus et al., 2011) the importance of wider classes of documents, and mixed-type event clusters cannot be overlooked.

Event clustering in the context of the wider social media footprint (undoubtedly a valuable lifelogging and reminiscence resource) cannot overlook the challenges of this heterogeneity, and its consequences for information extraction and approaches to clustering – as is alluded to by Bontcheva and Rout (2014, p. 24):

> “The majority of methods surveyed here have been developed and evaluated only on one kind of social media (e.g. Twitter or blog posts). Cross-media linking, going beyond connecting Tweets to news articles, is a crucial open issue, due to the fact that increasingly users are adopting more than one social media platform, often for different purposes (e.g. personal vs professional use). In addition, as people’s lives are becoming increasingly digital, this work will also provide a partial answer to the challenge of inter-linking our personal collections (e.g. emails, photos) with our social media online identities.
>
> The challenge is to build computational models of cross-media content merging, analysis, and visualisation and embed these into algorithms capable of dealing with the large-scale, contradictory and multi-purpose nature of multi-platform social media streams. For example, *further work is needed on algorithms for cross-media content clustering*, cross-media identity tracking, modelling contradictions between different sources, and inferring change in interests and attitudes over time.”

This finding confirms the motivations given earlier in the thesis regarding these challenges – to which *SAESNEG* is presented as an effort towards a solution.

4.5 *SAESNEG* System Architecture

A detailed diagram illustrating the various components in *SAESNEG* is shown in figure 4.5. These components include:

Participant Information Sheet / Consent Form (section 4.6) – the participant information sheet and consent form were provided in electronic form.

Facebook OAuth (also section 4.6) – integrates *SAESNEG* with Facebook to facilitate data collection.

Ground Truth Web Interface (section 4.10) – used to collect ground truth document event clustering from users for training/evaluation of pipeline.

Fetcher (including service, daemon) (section 4.8) – fetches source OSN dafa for the study.

Pre-Processing, Source Data Serialiser (section 4.9) – some pre-processing occurs prior to the data being serlised and persisted to disk.

User Database (section 4.7) – stores basic user information, and ground truth event clusters – used by many of the other components.

Phase A Sub-System (section 4.11 & chapter 5) – extracts annotations from the documents, using a variety of techniques.

Phase B Sub-System (section 4.12 & chapter 6) – organises documents into event clusters, based on annotations extracted in Phase A, uses ground truth document clusters for training and evaluation purposes.

Note this is an expanded view of the core experimental pipeline shown in figure 4.1.

4.6 Ethics and OAuth

4.6.1 Introduction

This section seeks to demonstrate that the study was ethically sound, conducted after full consideration of a range of potential ethical issues, and that University regulations regarding ethics were followed correctly.

Ethical and technical issues are tightly linked: the ethics-related functionality is included in the system architecture (section 4.5), in recognition of the primary importance of ethical issues in a study of this kind; dealing with personal information. Features grouped together as a single component include: an electronic participant information sheet, consent form and integration with Facebook OAuth – implemented as a Java web servlet[12]. This section is a discussion of these workings comprising this component of the architecture, and how they relate to ethical issues.

[12] Running under Apache Tomcat.

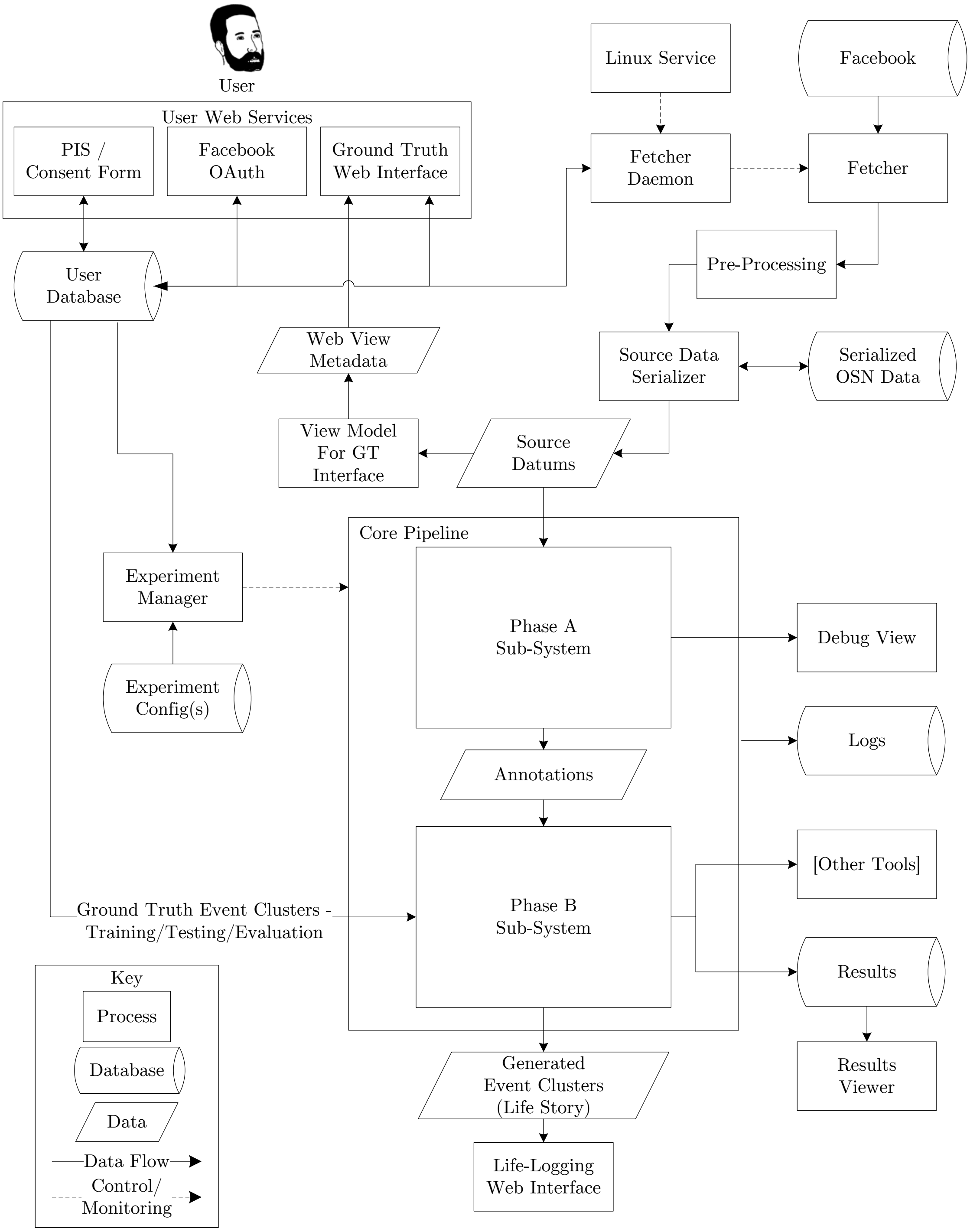


Figure 4.5: Detailed *SAESNEG* Architecture

4.6.2 University Ethical Approval

In accordance with University regulations, ethical approval was sought (and granted) for the research, and discusses:

- Issues regarding the participants, and their *capacity* to give informed consent (more detail in section 4.3).
- Risks, including hacking/security of data, inclusion of personal data in publications, distress/discomfort associated with review of past memories/events: issues relating to thanatosensitivity (Massimi and Charise, 2009; Massimi and Baecker, 2010), i.e. friends/family members who have died), and relationship break-up (Sas and Whittaker, 2013). Appropriate measures could be taken to minimise all risks identified.
- The ethics application also outlined the use of electronic participant information sheet/consent form, and the role of Facebook OAuth, discussed below.

4.6.3 Electronic Participant Information Sheet and Consent Form

Ethical guidance stipulates that a study should have a participant information sheet so that participants are informed of the particulars of the study (what they would be expected to do, how their data would be used, etc.), and a consent form where users formally give their consent.

For convenience, a combined participant information sheet and consent form was used, forming the first step in the sign-up process. Checkboxes are used to indicate consent; JavaScript code checks that all three are checked before allowing the user to participate in the study. The consent is recorded as a flag associated with the users details in the user database (section 4.7), which is checked at login[13].

Having consented, users are directed to login through the Facebook OAuth mechanism, discussed in the next section. Subsequent logins do not require repeat of the consent process.

4.6.4 Facebook OAuth

OAuth is the mechanism that controls access to the Facebook Graph API. Through OAuth, the user is able to grant *SAESNEG* access to their data, this permission is represented by an oauth access token. The RFC memo (Internet Engineering Task Force, 2012) summarises how OAuth can be used to grant access to this data (in this case, Facebook):

> "The OAuth 2.0 authorisation framework enables a third-party application [*SAESNEG* to obtain limited access to an HTTP service [Facebook]...on behalf of a resource owner [the Facebook user] by orchestrating an approval interaction between the resource owner and the HTTP service."

[13] This serves as a record that consent has been given, and used as a safeguard against circumvention.

The process is implemented by a series of redirects between the third party (*SAESNEG*) and Facebook, and is summarised in a sequence diagram shown in figure 4.6. Detailed documentation on implementing server-side OAuth with Facebook is available online[14].

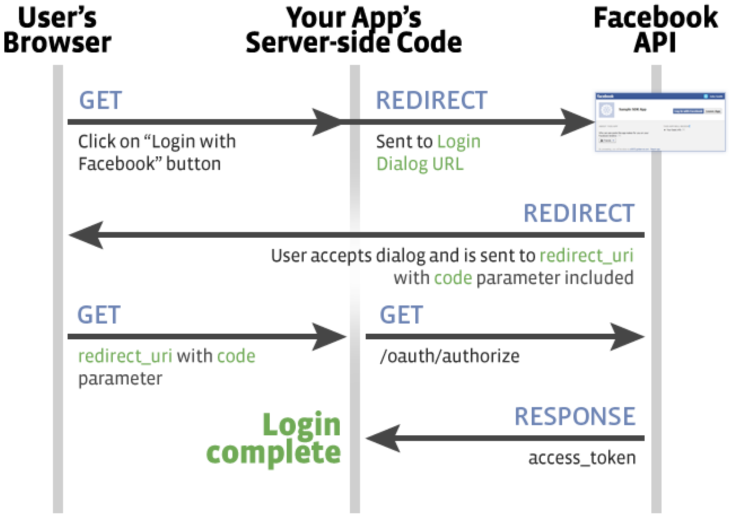

User's Browser
Your App's Server-side Code
Facebook API
GET
Click on "Login with Facebook" button
REDIRECT
Sent to Login Dialog URL
REDIRECT
User accepts dialog and is sent to redirect_uri with code parameter included
GET
redirect_uri with code parameter
GET
/oauth/authorize
Login complete
RESPONSE
access_token

Figure 4.6: Sequence Diagram showing Facebook OAuth Server-Side Flow. Reproduced from: http://webstersprodigy.net

Initially, the app (i.e. *SAESNEG*) is registered with Facebook, and issued with an ID and private key. The user is redirected to Facebook (the app ID and app's redirect URL are included as parameters), and after logging in, if the user is yet to grant permission to *SAESNEG* they are shown a dialog asking them to confirm the request (if the permission has already been granted, the process continues without prompt).

If the user approves, the user is redirected to the specified URL belonging to the app (verified against the apps' details), with a temporary code representing the permission. Upon receiving this request from the user's browser, the app then makes a server-side web request including the code, and the other app details (including the private key, which is not exposed publicly) in exchange for an access token.

Additionally, a state parameter is used to guard against cross-site request forgery attack[15] – although the scheme can still be attacked in various ways (Lundeen, 2013).

Having obtained the API token, basic user information is fetched from Facebook and stored in the user database (section 4.7) along with the API token. The Facebook user ID serves as the primary key, and is used to uniquely identify the user for subsequent visits.

Meanwhile, having effectively 'logged on' to *SAESNEG* using Facebook, users are directed to the homepage, with a hyperlink to begin assembly of ground truth event clusters (section 4.10), other links shown in the screenshot relate to other experiments outside the scope of this thesis.

SAESNEG s use of OAuth serves three purposes:

1. *SAESNEG* uses Facebook to verify the identity of the participants, allowing them

[14] https://developers.facebook.com/docs/facebook-login/manually-build-a-login-flow/v2.0

[15] http://en.wikipedia.org/wiki/Cross-site_request_forgery

to login securely, obviating any need for separate usernames/passwords. Users' of Facebook are dissuaded from having multiple accounts, hence ensuring that each participant has at most one profile in *SAESNEG*

2. Users are able to grant *SAESNEG* access to their Facebook data, with granular control over what data *SAESNEG* is able to access.

3. The confirmatory message shown to users serves as an additional consent mechanism – users are free to deny *SAESNEG* access to their data, or particular classes of objects.

At the time of PhD submission, the system is offline pending a re-review from Facebook, hopefully a system will be available online here: `http://www.benblamey.com/`. As the primary purpose of the proposed system is to create a new user experience for backed-up social networking data, the system cannot function without support from Facebook. Ironically, this further re-enforces the rationale behind the system: social networking data held in the cloud by a third party will always be vulnerable to loss, deletion or other interference.

4.7 User Database

At the core of the *SAESNEG* architecture lies the mongoDB user database, used by many of the other components. MongoDB[16] is a document-oriented database, where documents are represented as JSON objects, and is consequently considered a NoSQL database. The rationale for choosing such a database were the advantages of convenience, flexibility and simplicity in not having a formally defined database schema to maintain – the 'schema in code' design pattern (Wolff, 2013).

JSON is a text-based, ubiquitous data exchange format, frequently favoured over XML for its simplicity (Crockford, 2006). Using mongoDB means that instances of Java classes can be serialised into JSON and persisted in the database very easily. The Java drivers for mongoDB were used for communication with the database, and a PHP management interface, RockMongo[17], was used for debugging.

The user database plays a key role in the fetching of data, and in combination with the algorithm used to control the fetcher (section 4.8), forms a rudimentary queuing system for the fetching of source OSN data. The source social networking documents were not stored in the user database, but instead were persisted as files on disk (see section 4.9), with the metadata recorded.

Listing 4.1 shows an example database entry for a user in JSON, which includes the OAuth access token, basic profile information, details of fetched source OSN data, and ground truth document event clusters generated in the ground truth web interface (the entry has been edited for brevity and security). This is a abridged example of an entry for a user, and is included to evidence the implementation and operation of the system.

[16] `http://www.mongodb.org`

[17] `http://rockmongo.com/`

Listing 4.1: Example User Database Entry (JSON, abridged)

```
{  "FACEBOOK_BIRTHDAY":"04\/15\/1906",
   "FACEBOOK_EMAIL":"blamey.ben@gmail.com",
   "FACEBOOK_FIRST_NAME":"Ben",
   "FACEBOOK_LAST_NAME":"Blamey",
   "FACEBOOK_OAUTH_TOKEN":"CAAF7...3hSkZD",
   "FACEBOOK_OAUTH_TOKEN_EXPIRES_UNIX":1407845261,
   "FACEBOOK_OAUTH_TOKEN_UPDATED_UNIX":1402661263,
   "FACEBOOK_USER_ID":"728995201",
   "GROUND_TRUTH_EVENTS":{
      "lifeStoryFileName":"728995201_BenBlamey_1393247705.xml",
      "timestamp":1394804866822,
      "events":[
         { "datums":[
               { "title":"Ben Blamey worked so hard he was asleep
                  standing up, album:Cardiff and Newports Trip to
                  the Nautilus 2013",
                 "id":"10202360286585788" },
               { "title":"Paul struggling to eat all his meal at the
                   Pepley inn Sheffield at our training session
                  before the nautilus tomorrow, album:Timeline
                  Photos",
                 "id":"637616102947769" }
            ] },
         { "datums":[
               { "title":"The Welsh National Championships 2013",
                 "id":"10202565431794290" },
               { "title":"The winning Team Cardiff and Newport still
                   at the pool., album:The Welsh National
                  Championships 2013",
                 "id":"10202565438234451" }
            ] } ]
   },
   "HAS_DONE_ETHICS":true,
   "LIFE_STORY_INFOS":[
      {  "CREATED":1393247705,
         "CREATED_PRETTY":"2014-02-24T13:15:05.996Z",
         "FILENAME":"728995201_BenBlamey_1393247705.xml",
         "SOURCE":"FETCHER_DAEMON",
         "SUCCESS":true,
         "VERSION":2 },
      {  "CREATED":1394849750,
         "CREATED_PRETTY":"2014-03-15T02:15:50.838Z",
         "FILENAME":"728995201_BenBlamey_1394849750.xml",
         "SOURCE":"FETCHER_DAEMON",
         "SUCCESS":true,
         "VERSION":3 } ],
   "REGISTRATION_DATE":"2014\/02\/24 12:54:44" }
```

4.8 The Fetcher

4.8.1 Introduction

The fetcher component is responsible for retrieving a sample of the user's social networking footprint from the source social networks, the data upon which the remainder of the study is built.

Social networking data is often available via a HTTP API. Access to these APIs, is typically controlled through a scheme of access tokens associated with users of that OSN according to the OAuth mechanism. Section 4.6 contains a general description of the OAuth mechanism, its use in this system, and the related ethical issues – this section focuses on the retrieval of that data using the Facebook OAuth access token already obtained. The choice of source OSN has been discussed in section 4.3.

4.8.2 Scope

Whilst the rationale for using Facebook data exclusively has been discussed in section 4.3, another key decision to be made regarding the scope of source data was the time period for source social networking documents. Rather than simply downloading all available data, a decision was made to only download data created subsequent to the beginning of 2012, for the following reasons:

- Using data from too far in the past may affect people's ability to remember the related events, which may affect the accuracy of the ground truth data.
- However, this span was considered long enough (approximately 1.5–2 years) to increase the probability that people have documents created at different times that relate to the same event (perhaps before/during/after) – e.g. status messages like: "looking forward to my holiday", "at the airport", "thinking about what a great holiday I had last summer".

4.8.3 The Daemon

The execution of the fetcher component on the server was controlled by means of a daemon, so that the fetching was triggered and scheduled to run appropriately. A shell script (running as a Linux service) repeatedly runs a Java program – `DaemonMain.java`.

This scheme was chosen for robustness – if any problem occurs during fetching, the JVM terminates (with the error logged) and after a short pause, the program is restarted[18]. The service is set to start automatically when the server boots, and can be manually stopped and started for maintenance purpose [19].

The JVM is run with a low CPU priority[20], so that performance of other web services are not adversely affected.

Each time the fetcher is executed, it checks whether it should fetch data for any of the users who have signed up (details of signed-up users are held in the user database, section 4.7), if so data is fetched for one user, after which the application terminates.

4.8.4 Facebook Graph API

This section concerns the fetching of data itself. Facebook's Graph API was chosen for obtaining data. A short overview of accessing data through the API is given below, more detailed information is available online[21]. Data is retrieved with HTTP GET requests, with URLs formed according to simple patterns, as exemplified in table 4.6 (`me` is a pseudo-id representing the user ID of the owner of the Facebook OAuth access token used in the request):

The API token (obtained through the OAuth process – see section 4.6 is specified by means of a field-value pair in the URL query string (omitted from the examples). Because of the

[18] c.f. with the "let it crash" design pattern `http://c2.com/cgi/wiki?LetItCrash`, and the philosophy of Erlang `http://www.erlang.se/doc/programming_rules.shtml#HDR11`

[19] However, the system is robust enough that new `.class` files can be uploaded at any time – it does not matter if this causes the fetcher to crash, it will simply execute again after the delay!

[20] Using the `nice` utility (`http://linux.die.net/man/1/nice`)

[21] `https://developers.facebook.com/docs/graph-api/using-graph-api/v2.0`

URL	Response
`http://graph.facebook.com/<someid>`	A JSON representation of object with the specified ID.
`http://graph.facebook.com/me`	A JSON representation of the basic profile information for the user.
`http://graph.facebook.com/<someid>/<conn>`	a JSON array of objects which have the specified "connection" relationship with the object specified by the ID.
`http://graph.facebook.com/me/photos`	A JSON array of photos in which the user is tagged.

Table 4.6: Example URLs for Facebook Requests

large number of objects involved, results are paginated – see Facebook documentation for details[22].

4.8.5 Classes of Objects Fetched

As set out in the formal problem statement (section 1.4), the scope of this study is the personal social media footprint, so the objects of primary interest are those obtained with URLs of the form `/me/<connection>`. The set of connections for the user class roughly correspond to the different classes of objects found in the social network, objects of interest can be retrieved according to the "connection-based" URL scheme described above. The Facebook API documentation[23] shows the complete list of connections to the 'user' object class, stating whether or not the associated objects were retrieved as part of the social networking footprint, with a rationale for the decisions.

In making these judgments, the goal was to focus on a range of common object classes, making up the bulk of the social media footprint, with sufficiently broad range of data types to fulfil the novel aim of handling differently heterogeneous data (section 1.6), whilst avoiding distraction of low-volume data types. Objects that are always private (e.g. direct messages) are excluded on ethical grounds. Other information also fetched included the user's profile and list of friends, for use later in the experiment.

4.9 Representation and Persistence of Life Stories

The Facebook Graph API returns a representation of the objects in JSON format, an example of a photo is shown in listing 4.2. The example illustrates the nature of the source data: structured metadata fields (such as dates), free-text fields, which can contain information about the event (in this case, a music event in Manchester), and the URL of the image file itself (the 'link' field). An extensive sample of source data is shown in appendix B.

[22] `https://developers.facebook.com/docs/graph-api/using-graph-api/v2.0#paging`

[23] `https://developers.facebook.com/docs/graph-api/reference/user/`

Listing 4.2: Example Facebook Representation of a Photo (JSON, abridged)

```
{ "id": "10152272417078217",
  "from": {
    "id": "[removed]",
    "name": "Tamsin [surname removed]"
  },
  "name": "Radio 6 Festival, Manchester!",
  "picture": "<url removed>",
  "source": "<url removed>",
  "height": 540,
  "width": 720,
  "images": [
    { "height": 720,
      "width": 960,
      "source": "<url removed>" }
  ],
  "link": "<url removed>",
  "icon": "https://static.xx.fbcdn.net/rsrc.php/v2/yz/r/StEh3RhPvjk.gif",
  "created_time": "2014-03-02T15:46:46+0000",
  "updated_time": "2014-03-02T15:46:47+0000",
  "tags": {
    "data": [
      { "id": "728995201",
        "name": "Ben Blamey",
        "created_time": "2014-03-02T15:46:47+0000",
        "x": 40.927261352539,
        "y": 48.102165222168 },
      { "id": "...",
        "name": "Tamsin [surname removed]",
        "created_time": "2014-03-02T15:46:47+0000",
        "x": 66.696281433105,
        "y": 41.715492248535 }
    ],
    "paging": {
        "cursors": {
        "before": "NzI4OTk1NjAx",
        "after": "NjE0ODExHjE2"
      }
    }
  }
}
```

Whilst JSON lacks built-in support for schemas, the representation of Facebook objects is stable, and libraries have been created to map the JSON into static types. *SAESNEG* uses such a Java library, RestFB[24], which is able to fetch individual items from the Facebook API, and parse the JSON objects; converting them into instances of Java classes – mapping JSON attributes to properties on the Java objects automatically. This yields the benefits of static-typing: the properties are easily discoverable when developing through the IDE assist mechanism, and references to Facebook objects and their properties are automatically checked by the Java compiler.

Other libraries are available for other social networks, so there is no issue adding support for them in the future, alternatively, the documents could have been represented as raw XML or JSON objects – RestFB was used purely for convenience. The resulting Java objects are then each stored inside *SAESNEG* wrapper classes.The resultant set of Java wrapper objects, representing the newly fetched documents needed to be serialised to disk so that the experiment could be run repeatedly without fetching the information each time. XML serialisation was chosen over binary for portability, readability and debug-ability. The XStream[25] framework was chosen.

[24] `http://restfb.com/`

[25] `http://xstream.codehaus.org/`

4.10 Ground Truth Web Interface

This section describes the design and implementation of the ground truth web interface component of *SAESNEG* The interface and its supporting functionality was designed for fully automated collection of ground truth document event clusters for use in the training of the machine-learning based clustering, and evaluation of the pipeline on the on the document event clustering task. Chapter 6 explains exactly how the data was used for these purposes (an overview is presented in section 4.2), whilst the formal problem statement can be found in section 1.4. Evaluation of the ground truth web interface itself i.e. how successful it was at collecting ground truth event clusters – is discussed in section 4.4.

4.10.1 Rationale for User-Created Ground Truth

The direct creation of ground truth data can be highly time-consuming (Huang and Dom, 1995, p. 55), and many existing studies take existing data (perhaps applying some heuristic[26]) as a source of ground truth. Clearly, such an approach is prudent as machine-learning techniques can be highly effective when large quantities of training data is available. However, such approaches can be approximations of ground truth (depending on the situation), could be criticised as having what is arguably a distorted evaluation, whilst the heuristic can restrict the scope of investigation to a specific class of cases (e.g. place names mentioned in album titles (Rabbath et al., 2012)). Instead, this study lets users create a variety of event clusters themselves, judgements which *SAESNEG* then attempts to replicate.

Another reason for not using existing source of ground truth was that no such resource could be found. Existing studies use public datasets that typically consist of a single type of document (e.g. Becker et al. (2010)) – the defining characteristic of this study is the differently heterogeneous nature of the source data (section 1.6): a mixture of image, text and metadata, with various types of documents, constituting the personal social media footprint, and containing many private events – no source dataset with these characteristics, and ground truth annotations, could be found.

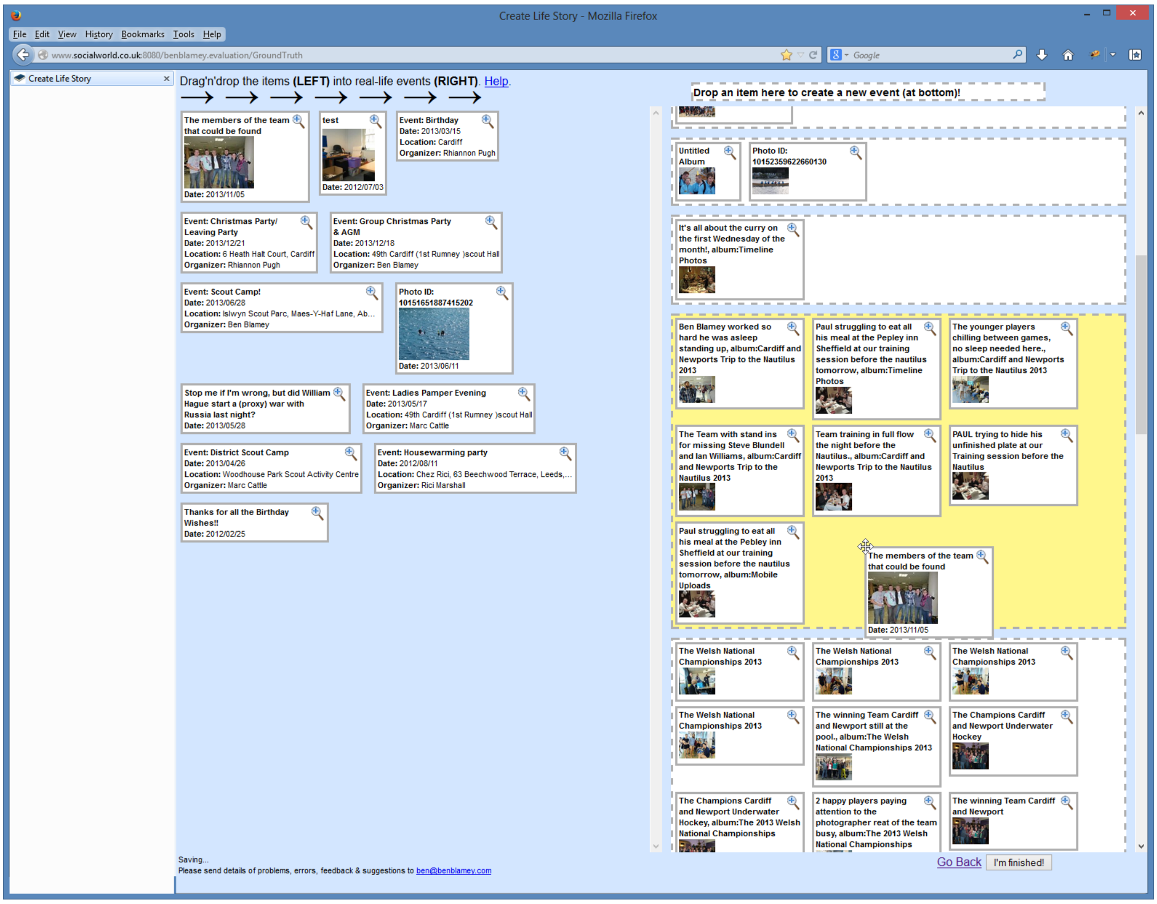
Hence, it was decided that ground truth would be gathered from users, so that the artificial intelligence challenge was to directly replicate human judgement, rather than replicate a heuristic – it is expected that this choice of a "real world" problem would create difficulties, and possibly result in lower performance.

4.10.2 Layout Overview

The ground truth web interface is shown in figure 4.7, the key features of the layout are:

- The area on the **left hand side** shows the pool of unallocated documents from the user's social media footprint.
- The area on the **right hand side** contains the document event clusters.
- The **header** gives simple instructions, with a link to the FAQ, with "new event" area – a document dropped here is allocated to a newly created event.
- The **footer** contains a message indicating the status of the auto-save, contact email, and other commands.

[26] See section 6.8 for discussion of using albums as a source of ground truth events.


Create Life Story - Mozilla Firefox
File Edit View History Bookmarks Tools Help
www.socialworld.co.uk:8080/benblamey.evaluation/GroundTruth
Google
Create Life Story
Help
Go Back
I'm finished!

Drag'n'drop the items **(LEFT)** into real-life events **(RIGHT)**. Help.

The members of the team that could be found
Date: 2013/11/05

test
Date: 2012/07/03

Event: Birthday
Date: 2013/03/15
Location: Cardiff
Organizer: Rhiannon Pugh

Event: Christmas Party/ Leaving Party
Date: 2013/12/21
Location: 6 Heath Halt Court, Cardiff
Organizer: Rhiannon Pugh

Event: Group Christmas Party & AGM
Date: 2013/12/18
Location: 49th Cardiff (1st Rumney)scout Hall
Organizer: Ben Blamey

Event: Scout Camp!
Date: 2013/06/28
Location: Islwyn Scout Parc, Maes-Y-Haf Lane, Ab...
Organizer: Ben Blamey

Photo ID: 10151651887415202
Date: 2013/06/11

Stop me if I'm wrong, but did William Hague start a (proxy) war with Russia last night?
Date: 2013/05/28

Event: Ladies Pamper Evening
Date: 2013/05/17
Location: 49th Cardiff (1st Rumney)scout Hall
Organizer: Marc Cattle

Event: District Scout Camp
Date: 2013/04/26
Location: Woodhouse Park Scout Activity Centre
Organizer: Marc Cattle

Event: Housewarming party
Date: 2012/08/11
Location: Chez Rici, 63 Beechwood Terrace, Leeds,...
Organizer: Rici Marshall

Thanks for all the Birthday Wishes!!
Date: 2012/02/25

Saving...
Please send details of problems, errors, feedback & suggestions to ben@benblamey.com

Drop an item here to create a new event (at bottom)!

Untitled Album

Photo ID: 10152359622660130

It's all about the curry on the first Wednesday of the month!, album:Timeline Photos

Ben Blamey worked so hard he was asleep standing up, album:Cardiff and Newports Trip to the Nautilus 2013

Paul struggling to eat all his meal at the Pepley inn Sheffield at our training session before the nautilus tomorrow, album:Timeline Photos

The younger players chilling between games, no sleep needed here., album:Cardiff and Newports Trip to the Nautilus 2013

The Team with stand ins for missing Steve Blundell and Ian Williams, album:Cardiff and Newports Trip to the Nautilus 2013

Team training in full flow the night before the Nautilus., album:Cardiff and Newports Trip to the Nautilus 2013

PAUL trying to hide his unfinished plate at our Training session before the Nautilus

Paul struggling to eat all his meal at the Pebley inn Sheffield at our training session before the nautilus tomorrow, album:Mobile Uploads

The members of the team that could be found
Date: 2013/11/05

The Welsh National Championships 2013

The Welsh National Championships 2013

The Welsh National Championships 2013

The Welsh National Championships 2013

The winning Team Cardiff and Newport still at the pool., album:The Welsh National Championships 2013

The Champions Cardiff and Newport Underwater Hockey

The Champions Cardiff and Newport Underwater Hockey, album:The 2013 Welsh National Championships

2 happy players paying attention to the photographer reat of the team busy, album:The 2013 Welsh National Championships

The winning Team Cardiff and Newport

Figure 4.7: Screenshot of Ground Truth Web Interface.

4.10.3 User Interface Design Considerations

Given the difficult and challenging task of ground truth data creation, design of the web interface to make it as easy as possible for users to pursue this goal, and to ensure the quality of the generated data. A number of design features of the interface have this in mind:

- Documents are presented summary form.
- Full details are available by opening the document in a pop-up.
- Filtering the documents so that only a manageable quantity are displayed, and certain documents are excluded.
- Persisting the state of the user interface between visits, automatically.
- A simple drag-and-drop interface, entirely mouse-driven.
- An interface which is enjoyable to use, displaying the users historic data and providing visual feedback.
- Documentation and Support.

The remainder of the section discusses the implementation of these features.

4.10.4 Presentation of Documents

The ground truth interface is implemented a Java EE Servlet, running in an Apache Tomcat server.

A number of plugins are used, as part of the jQuery[27] framework to implement the drag and drop functionality for documents: jQuery-UI draggable and jQuery-UI droppable[28]. A "zoom" button is available, which shows a more detailed view (i.e. more of the fields), implemented with the jquery-popup plugin – to inform user's clustering judgements.

Presenting the user with an excessive volume of data in the ground truth interface would likely make them feel overwhelmed and cause them to "give up"[29], so a maximum of 50 unallocated documents are shown on the left hand side of the interface. Subsequent visits or refreshing the page would result in the unallocated document pool on the left hand side being 'topped up' to the maximum (dependent on the underlying OSN sample). Facebook events that occur in the future are excluded, and excluded from further evaluation – the study tries to focus on events that occur in the past.

The documents are left as a pool and without additional structure – the idea is to let users organise their content without being their judgement being distorted by existing structures – allowing the relationship between document event commonality and user structures such as albums to be investigated objectively (see section 6.8).

4.10.5 Persistence of Ground Truth

Each time the web interface is loaded, the set of previously allocated ground truth event clusters is read from the user database (if any exists) and reconciled with documents contained within the persisted OSN source data – documents already allocated into event clusters are displayed on the right hand side, unallocated documents from the source OSN data on the left hand side.

An auto-save feature periodically saves the ground truth event clusters (the data is POSTed back as JSON)[30]. Hence, the state of the user ground truth interface is automatically persisted between visits, so that users can visit repeatedly to build additional ground truth. The core experimental pipeline uses the same database, so any new ground truth data is immediately available for experimentation (a snapshot was taken for the final comparisons) – this meant that development of the pipeline could continue whilst data collection was underway. A mechanism prevents complications if two instances of the interface are loaded at the same time, and cleanup is performed to ensure that documents are not duplicated into multiple ground truth clusters.

4.10.6 Help and Support

Automation is about more than the user interface itself: resources were provided to instruct and support users and integration with the user database and store of persisted source social networking data – the intention was that a significant number of participants could use the system with little direct support.

[27] `http://jquery.com/`

[28] `http://jqueryui.com/draggable/` `http://jqueryui.com/draggable/`

[29] This was a common observation when gathering feedback from users during early development.

[30] Hence, there is no 'save' button for the user to forget to click

- Prior to entering the ground truth interface, users are directed to an instructional screencast[31].
- An invite email is sent to inform users that the ground truth interface is ready for use.
- A list of hints and tips.
- A hyperlink to allow users to easily email feedback/queries to the author.

4.11 Phase A: Extraction

The Phase A sub-system is responsible for extracting information from the source documents likely to be useful for computing document event commonality. The input is the set of documents in their original representation from the source social network, and the output is a set of annotations associated with each document. Phase A operates on documents individually, although the text for all documents is actually processed as a single document for speed and simplicity. Figure 5.1 shows the implementation of Phase A in more detail, the various components shown are discussed in this section.

The different kinds of data (image, text, metadata) require different approaches. These existing techniques have been reviewed extensively in chapter 3 – for example, for textual data; natural language processing techniques are used, such as named entity extraction and parsing of temporal expressions, whilst CBIR techniques are used on image content data[32].

The document wrapper types are responsible for marshalling the information in the respective source document class they represent into the appropriate processing pipeline. Various configuration options for the processing components are controlled separately as part of the experiment harness. Phase A is discussed in more detail in chapter 5.

4.11.1 Information Extraction from Metadata

For the purposes of this discussion, *metadata* is any source data which is neither a free-text field nor image content – information such as timestamps, IDs of related objects, and short text-fields that do not require full text parsing, such text fields known to be locations or names[33]. This component is effectively a mapping from Facebook-specific representation to an OSN-neutral annotation representation of information for the computation of document event similarity. Other fields (such as titles, descriptions) are handled separately in the text processing sub-system.

However, the mapping is non-trivial, and depending on the type of document, different fields may have different semantic relationships with the underlying real-world event – so some degree of what is effectively *business logic* is required.

[31] `http://player.vimeo.com/video/74226104`

[32] Image Content != Photo. Image content is often present in a variety of documents, not just photos. Conversely, photo documents contain a mixture of image, text and metadata.

[33] The boundary between these kinds of metadata can be somewhat blurred: consider this example, "25/06/2014" may appear in an XML document `<event date="25/06/2014" />` – metadata, yet the same string can appear in text: "The Scout Fete will be held on 25/06/2014." A similar issue occurs with locations, and the names of people, consequently, the values of such fields are looked up appropriately using the same resources as in the text sub-system; but are not given full text processing – because full tokenisation, named entity detection etc. is not necessary.

4.11.2 Image Processing Sub-System

CBIR techniques are used extract to image content features (such as Edge & Color[34] Histograms). The two key components in the implementation are the **Image Feature Extractor** (a wrapper for Caliph-Emir[35] (Lux, 2009)), incorporating a feature cache (keyed according to the URLs of source images) for performance. The annotations extracted were some of the MPEG-7 image content descriptors. For further details of image feature extraction in Phase A – see section 5.9, these annotations are used by the scene strategy in Phase B (see section 6.7).

4.11.3 Text Processing Sub-System

NLP is a key topic for this thesis, with several contributions relating to novel text-processing techniques developed in this study. One of the important rationales for this project was to apply state-of-the-art techniques from the field of NLP into the event-clustering task, where the multimedia research community tend to focus on images and image metadata. A goal of this thesis is to investigate whether state-of-the-art text processing (especially named entity resolution) is useful in the context of event clustering, especially as an approach to overcoming the sparsity private events.

Section 3.2.8 discusses the strengths of the Stanford coreNLP system and GATE – a decision was made to simply include support both frameworks, allowing their performance on the corpus to be compared, and their various features to be used together. The text-processing system is implemented as follows:

Text Extraction is performed for each document, within the wrapper type. Text from the documents is concatenated into a single GATE document for subsequent processing.

The GATE's ANNIE pipeline is then executed on the data, with its various processing components.

The Stanford coreNLP pipeline is executed on the same document; the temporal expression parser was modified, and a novel social event parser was also developed (see chapter 5).

The GATE Document was then serialised to disk as GATE XML. Annotations from both Stanford coreNLP and ANNIE are included for viewing in the GATE GUI for debugging, analysis and comparison. Copies of the documents were manually gold-annotated, allowing evaluation of the generated annotations. Documents were sorted (by ID) so that the generated GATE document would be consistent with the gold annotated document.

In the pipeline, the various detected named expressions (and social events) are processed into annotations for use in Phase B (shown as post pipeline text processing).

4.11.4 Annotations

These annotations are the output of Phase B. As discussed in section 4.2, the design of the annotations as a layer of abstraction between the phases is a crucial part of the strategy to make greatest use of the available source data and overcome sparsity. The annotations are independent from the kind of underlying data (where possible), for example, a temporal

[34] American English has been used throughout for consistency with MPEG-7

[35] `http://www.semanticmetadata.net/features/`

annotation extracted directly from a timestamp field (i.e. metadata) is represented in the same way as an annotation generated by the text processing sub-system, such as from a temporal expression.

Crucially, the annotation layer represents facets of events, rather than directly representing the information found in the source events. Hence, the information can require some interpretation based on the surrounding context – for example, a timestamp on a status message which is geo-tagged is likely to directly indicate the time of the event. However, the upload timestamp of a photo indicates that the photo was taken (and thus the event occurred) at some point *before* that time. Such interpretation is often dependent on the semantics of the particular document type, and is implemented differently in each wrapper class.

A more abstract way of looking at Phase B is that it maps between two different semantic spaces – information relating to *documents*, to information relating to *events*. More details of the specific annotations generated are to be found in chapter 5, with chapter 6 detailing how they are used by the various document similarity strategies.

4.12 Phase B: Clustering

The purpose of Phase B is to partition the source documents into event clusters. The input is the source documents, with the annotations extracted during Phase A; and the output being the partitioning of that set into document event clusters; constituting the life story of the user. Whereas Phase A operates on documents individually, Phase B compares the annotations of different documents in the social media footprint to perform the clustering. Phase B is itself broken into two stages.

In summary, every document pair (representing an edge in the graph of documents) is compared by a set of commonality strategies (spatial, temporal, etc.), each of these strategies outputs one or more features (with real values in [-1,1]). These output features are disjoint, but input annotations may be used by more than one commonality strategy. Next, an SVM classifier is trained to distinguish between those edges where the documents are in different events (i.e. inter-event edges) and those where the documents are in the same event (intra-event edges). These edge classifications are used by a clustering algorithm to create the final partitioning into sets of documents representing events. This process is repeated separately for each participant.

The strategies for judging document event similarity are based on intuitions about what constitutes positive or negative evidence for document event commonality, and are a mixture of existing and novel approaches. To re-iterate, the decoupling achieved through the use of the abstract annotation layer between phases A and B mean that the strategies are dependent neither on particular kinds of data nor types of document (where possible) – hence Phase B is concerned only with abstract, informational event facets.

This decoupling helps to overcome the inherent sparsity of the private events in the source dataset, by effectively aggregating information relating to the various informational event facets from the variety of source data. This allowing documents of different types to be compared in an abstract way, necessary when event clusters contain multiple types of document – without such abstraction, the dimensionality would be too great for machine-learning to be effective.

The ground truth data is used by Phase B in various ways:

- It serves as training data for the SVM classifier to learn to classify inter- and intra-event edges (see above).

- A typical cross-validation scheme can be used to evaluate the per-edge performance of the classifier (independent on the clustering algorithm).

- For the overall experiment, performance is evaluated mathematically by using the *overlapping normalised mutual information* measure (section 6.11) to compare the generated and ground-truth partitionings. Results are stored in a database for later analysis.

- Various other tools were developed for visual comparison between the ground-truth and generated clusters (see appendix D), a web 'matrix' view, outputting clusters in plain text – allowing source code file comparison 'diff' tools to be used to inspect generated clusters (see appendix C). This allows the investigator to examine events where clustering performance has been poor, perhaps discovering new ideas for additional document comparison strategies.

- Clearly, it was necessary to partition the set of ground truth clusters into training and testing sets for meaningful evaluation.

For the final clustering step, the performance of various algorithms was evaluated, as discussed in section 6.10.

Hence, the complete system (combing phases A and B), when trained, is able to take an unseen set of source OSN documents belonging to a user, and partition that set into event clusters, for presentation in a user interface – facilitating event-centric user navigation of a social media footprint, naturally containing a mixture of different kinds of documents. To the knowledge of the author, this is the first investigation into event-clustering on such a corpus. Phase B is described in more detail in chapter 6.

4.13 Summary

This chapter has introduced *SAESNEG*, and described its implementation. The core pipeline has been introduced, with its two-phase Architecture – as well as an overview of the primary experiment consistent with the formal problem statement 1.4. The other constituent parts of *SAESNEG* have been examined: the various tools and components that are crucial to the operation of the system.

The ground truth web interface has been discussed, and the collected data has been analysed – the resultant findings supporting earlier hypotheses regarding the unique characteristics of the social media footprint, and the challenges these create for the clustering task. Such findings call for more attention for the clustering of different types of documents.

Furthermore, the chapter has also developed the argument of how the framework has been designed to specifically address these challenges: the two-phase design, decoupled with the abstract annotation layer – in an effort to overcome sparsity, overcome the 'curse of dimensionality', and facilitate the comparison of different types of data. Phase B has been introduced, showing how it incorporates a variety of information extraction techniques to make best use of various kinds of source data (as reviewed in the previous chapter) – another strategy in overcoming event sparsity.

Wider issues have been discussed: the selection of participants (section 4.3), ethical issues (section 4.6), whilst discussion has been supported by a variety of figures, tables, listings and examples – giving a firm foundation for the remaining chapters, and the methodology of the study.

The next two chapters (5 and 6) respectively focus on the two phases of the core pipeline (phases A and B), including the design considerations, algorithms, implementation, results of experimental evaluation. The contributions are then summarised in chapter 7.

Chapter 5

Phase A: Information Extraction

5.1 Introduction

As discussed, the core *SAESNEG* pipeline is implemented according to a two-phase Architecture: Phase A (extracting information from individual documents) and Phase B (comparing extracted information to perform document event clustering). The overall system generates the user's life-story (denoted $L_C(u)$ in section 1.4), from the source documents. The scope of this chapter is Phase A; documenting its design decisions, choice of algorithms – both new and existing (drawing on the technical literature review in chapter 3), implementation, representation of output annotations, and empirical analysis.

Each of the documents in the target user's social media footprint (or a subset thereof) are processed sequentially through the components described in this chapter, yielding a set of annotations representing the extracted information associated with each document. Information associated with different documents is not compared until Phase B, where the annotations are used to determine event commonality. The extraction techniques are implemented in *SAESNEG* to support the event commonality strategies in Phase B (chapter 6), in turn to conduct the overarching event-clustering experiment[1]. All data examples shown are sourced (with permission) from users' social media documents.

Each of the documents in the target user's social media footprint (or a subset thereof) are processed sequentially through the components described in this chapter, yielding a set of annotations representing the extracted information associated with each document. Information associated with different documents is not compared until Phase B, where the annotations are used to determine event commonality. The extraction techniques are implemented in *SAESNEG* to support the event commonality strategies in Phase B (chapter 6), in turn to conduct the overarching event-clustering experiment. All data examples shown are sourced (with permission) from users' social media documents.

As discussed in chapter 1, a defining feature of the source dataset (and this study) is its so-called *differently heterogeneous* mixture of data: individual documents are a mixture of image, text and metadata – and additionally – the schema of the documents differ according to type, and may confer a different interpretation depending on that context. In addition to such properties of the data, given its sparsity, it is important that maximum information is extracted from source documents (section 1.6).

[1] see the 'key idea' highlighted paragraphs in chapter 6 for the intuitive rationale for extracting the annotations.

This chapter explains how the design and implementation of Phase A (and its components) addresses these challenges, and how the so-called issue of *curse of dimensionality* is overcome through the use of an appropriate annotation layer, carefully abstracting the useful information away from the type and kind of its originating context.

Classes for each type of document control the extraction and interpretation of image, text and metadata – whilst actual information extraction from image and textual data have their own respective pipelines. Note that each instance of an annotation may (theoretically) originate from image, text or metadata, from within any of the underlying document types – for this study, a selection of extraction components have been implemented as appropriate for the source data, with lots of opportunities for future expansion.

Given that each of the components in Phase A are independent, there is not a dedicated 'results' section, instead, results are located alongside discussion in section to which they relate. Clustering performance itself is not evaluated until the next chapter. Figure 5.1 shows the components which constitute the Phase A sub-system, whilst figure 4.5 illustrates its place within the *SAESNEG* architecture.

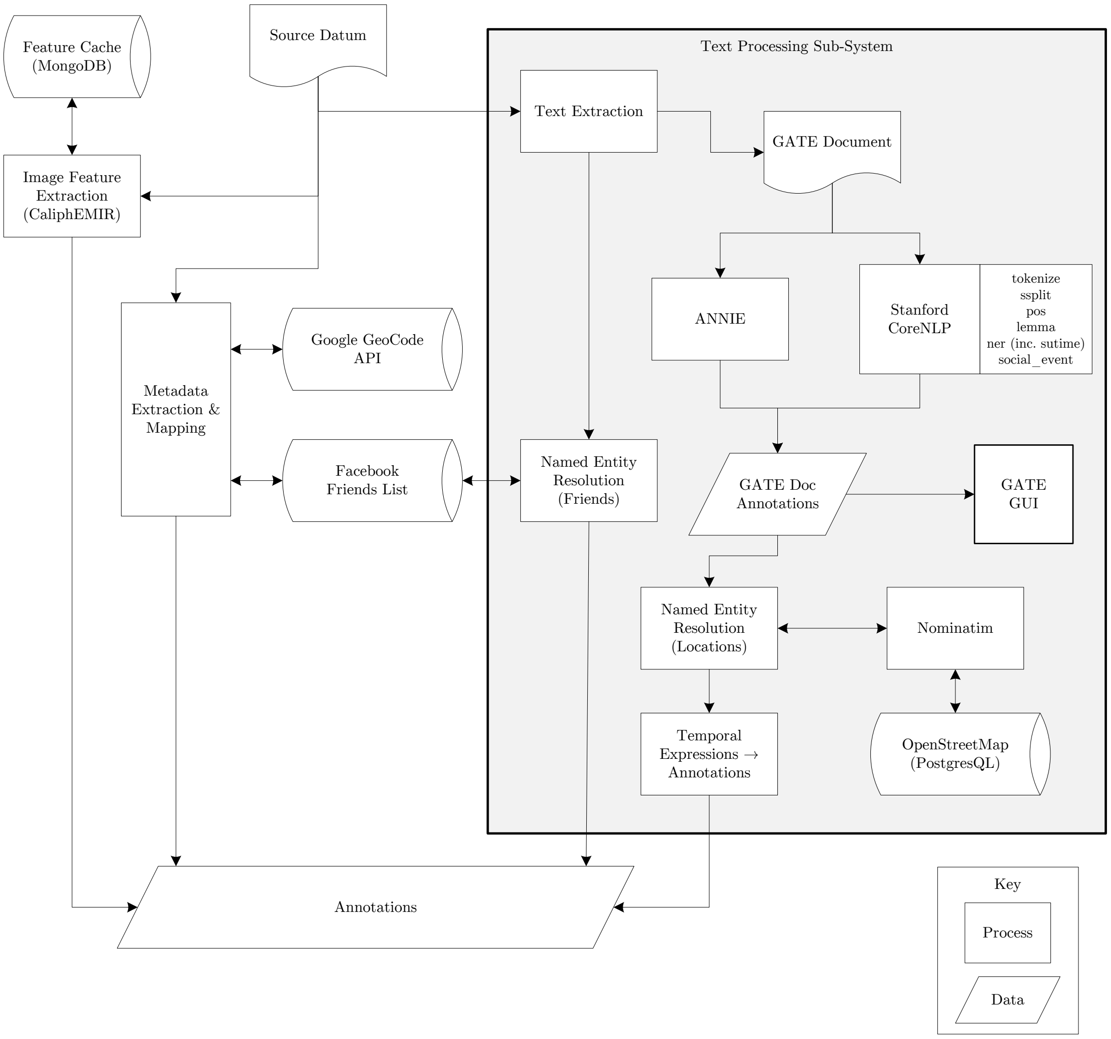


Figure 5.1: Physical Implementation of the Phase A Sub-System.

This chapter is structured according to the output annotations, hence, the section for each annotation, friends (section 5.3), locations (section 5.4), temporal information (section 5.5) and well-known events (section 5.6) may actually contain discussion of multiple components. The work on the temporal expressions in natural language is a **key contribution of this**

thesis, and has resulted in a peer-reviewed publication Blamey et al. (2013).

Section 5.11 outlines the contributions of this chapter, especially the more novel components relating to temporal information (section 5.5) and well-known social events (section 5.6). The chapter begins with an overview of the text pipeline (section 5.2) – text and natural language processing being a key focus of this study.

5.2 Text Pipeline

5.2.1 Introduction

This section gives an overview of the text processing functionality in Phase A of *SAESNEG* focusing on the pipeline itself rather than particular components, which are discussed in greater detail in later sections of this chapter.

The text processing pipeline served a dual purpose: primarily, the extraction of information useful to the document event commonality task – but also to serve as a platform for the evaluation of existing components in pursuit of this aim, and the development of new techniques and components.

A GATE document is used as the basis for processing all the text associated with all the documents belonging to a particular user, with other components interact with this GATE document in various ways. Listing 5.1 gives an overview of the text processing pipeline, focusing on the life-cycle of this GATE document, whilst figure 5.1 shows the components in Phase A, including those associated with text-processing.

This section serves to explain how the different components were integrated, please see other referenced sections for details of individual components. Please note the distinction between GATE annotations in text and annotations associated with documents used in Phase B.

1. An empty GATE document is created.
2. For each document:
    (a) For each natural language string field in the document (as determined by the java class corresponding to the document):
        i. Appends the text in natural language metadata field to the GATE document.
        ii. Each is added as a separate line, and all the fields corresponding to a given document correspond to a contigious section in the GATE document.
        iii. An annotation is added to the GATE document to identify each text field.

    (b) an annotation is added for the contiguous region associated with the document.

    (c) The start and end indices of the block associated with the document are stored inside the java object representing the document.
3. Optionally, the resultant GATE document is exported for manual annotation and inspection in the GATE GUI. Because the document is a consistent representation of the text associated with a social networking footprint, it can be replicated, and annotations from different analysis procedures can be compared on an offset basis, and so that gold standard annotations can be created.
4. The document processed with the Stanford CoreNLP pipeline and/or GATE ANNIE, customised with the various techniques described in chapter 5. This adds annotations for named entities, temporal expressions etc.
5. The annotations generated by the Stanford pipeline are converted from their native format and added to the GATE document.
6. Optionally, the document can be exported at this stage for review of annotations.
7. For each document:
    (a) For each annotation that occurs inside the contiguous region associated with the document (as determined according to the previously stored boundaries).
        i. Additional processing is performed, such as resolution of named entities with external resources.
        ii. The information in the annotation is converted into annotations for Phase B (alongside those entries extracted directly from the metadata and images).
8. Optionally, the document can be exported at this stage for review of annotations.
9. Optionally, the annotations in the document can be evaluated against a gold standard document created earlier, allowing the computation of annotation accuracy metrics as listed in chapter 5.

Listing 5.1: Phase A Text Processing Algorithm Overview

5.2.2 Choice and Configuration of Software

To summarise earlier discussion (section 3.2.8), because of the requirement for being able to inspect annotations for development and evaluation, GATE was a clear choice with its fantastic user interface – which allows the even large documents to be navigated easily, and annotations to be inspected and created. However, the system was designed so that

Document Type	Fields Included
Album	name, description
Check-in	message, comments
Event	name, description
Photo	name, comments
Status Message	message

Table 5.1: Fields included for Text Processing

other text analysis libraries, such as components from Stanford's NLP suite (especially its CRF-based NER tagger), which are perhaps more state-of-the-art, could be used as well – although they lack such a friendly user interface. Although many such libraries are converted into GATE plugins, a decision was made to use the libraries directly, so that they could be extended more easily, and not so tightly integrated into GATE.

Regarding the issue of unnatural language, much of the text, when inspected, appeared relatively 'normal', originating from Facebook, where there is not the 140-character limit of Twitter[2]. The demographics of the participants (i.e. University educated, mostly first-language English) may have been a factor. A decision was made to use standard tools initially, and then investigate OSN-specific alternatives (e.g. Ritter et al., 2011) if and when necessary. The literature review discussed the issue of unnatural language in greater depth, and expressed concerns about the term being used as a catch-all to describe subtle, perhaps corpus-specific features.

5.2.3 Extraction

During the initial extraction step, text from each document is appended to the GATE document (annotated to show the originating field). Not all the text from the documents was included, for example, comments were excluded from some document types because they tended to contain content not directly associated with the underlying real-life event (comment metadata was also excluded from 'friends' analysis – section 5.3), temporal expressions in particular often referred to other events. The text fields selected for analysis for each document type are listed in table 5.1.

This document was then serialised to disk, so that it could be opened in GATE for inspection – it was at this stage that a copy could be made, and annotated with gold labelling (using GATE), for evaluation of the location component (section 5.4).

5.2.4 Processing

The text processing itself is performed in two stages: in the 'processing' step the document itself is processed by the tools being evaluated, and the output annotations from that tool are converted to GATE annotations. Subsequent 'post-processing' is described in the next section. The initial processing step is performed on complete GATE document, although the GATE segment processing PR was used to separate processing of individual documents in ANNIE. The Stanford NLP suite was used as the basis for text processing because of its state-of-the-art NER components, and its TokensRegex[3] functionality for defining and

[2] Much of the literature relating to unnatural language of late has focused on Twitter, whereas Facebook is the corpus for this study.

[3] http://nlp.stanford.edu/software/tokensregex.shtml

extending grammars. The SUTime component was modified (section 5.5) and a newly-developed parser for social events (section 5.6) was also included. Mentions of friends were detected by a simple gazetteer (section 5.3), with GATE's used for sentence splitting, for detection of location 2-grams (section 5.4).

5.2.5 Post-Processing

All text annotations created during text processing are stored as GATE annotations (i.e. the annotations from Stanford CoreNLP are converted into GATE format). These annotations are then processed, so that the context (i.e. originating type, field) can be considered if appropriate, and so that external resources and heuristics can be used to generate the final annotations, adding to those extracted from images and metadata.

5.2.6 Evaluation

The final GATE document could also be exported, for inspection of final location annotations. GATE's *Annotation Diff Tool*[4]) was used to evaluate the performance in the case of the locations; see screenshot in appendix C. The following sections discuss the extraction of information, according to the output annotation (i.e. friends, location, temporal etc.).

[4] `https://gate.ac.uk/sale/tao/splitch10.html#sec:eval:annotationdiff`

5.3 Friends

5.3.1 Introduction

This section concerns the extraction of information related to people – this might be mentions of peoples' names in textual data (i.e. NER), occurrence of user-ids in metadata, or the recognition of individuals faces in image data, to facilitate the document event clustering task.

Regarding the document event commonality task, the rationale for extracting people-related information is that, supposing there are two photos, and a set of people appear in both, this is positive evidence that the photos may relate to the same event (dependent on other evidence). Appearance in photos is just one way that individuals can be associated with a social media document – they could be mentioned in some text field (such as a title or description) or the metadata could explicitly indicate a relationship (such as users 'tagged' in content).

In this section, the functionality discussed is that which extracts information regarding user event participation. Existing work is reviewed, the implementation is described, along with the representation of the extracted information as annotations – used in the so-called 'friends' strategy in Phase B (section 6.3).

5.3.2 Existing Work

Using information about event participants to compute event similarity is not a new idea: various existing studies have sought to use the information when it was available.

Section 3.4.5 discusses such existing work in detail, including ways in which image content (i.e. facial and clothing recognition) can indicate event participants, for use in the event clustering task. However, for this initial implementation of *SAESNEG* information extraction for the discovery of event participants was not directly applied to image content – Facebook was used as the source of social networking data for this study (section 4.3): Facebook has its own functionality for facial recognition, to assist people in *tagging* users appearing in the photo. Users are able to 'untag' themselves from photos that they do not wish to be associated with their profile. Because of this, it was decided that facial recognition would not be implemented in this iteration of *SAESNEG* many of the people in the photos would already be tagged, and those that were not perhaps did not want to be. Future work could easily implement facial recognition, which might be especially useful for other social networks where 'tagging' is not used, or is less common.

Instead, for this initial study, information extraction was applied to the text (i.e. named entity extraction – section 3.2.6.4) and metadata only. The next sections discuss the implementation of event-participant information extraction from metadata and text respectively.

5.3.3 Implementation: Metadata

Metadata can indicate various kinds of associations between documents and users, in Facebook, the users in such associations are unambiguously represented are as the pair (full name, user id), requiring no further processing – the question is which associations are interpreted as indicating event participation, this depends both on the type and field. The

Document Type	Meaning/Interpretation of other fields, interpretation for event semantics.	Meaning/Interpretation of 'owner'	Owner assumed to be event attendee?
Check-in	Users tagged in the check-in.[5].	User who 'checked in' to the location.	Yes.
Event	Event Attendees: 'Maybe' and 'Declined' invitees are excluded.	User who created the event.	Yes.
Photo	Users tagged in photo[6]	User who uploaded the photo is assumed to have taken the photo.	Yes.
Status Message		Creator of message.	Yes.
Album	N/A	Creator of Album.	No – not necessarily event attendee.

Table 5.2: Semantic interpretation of references to users in metadata, by document type.

representation of this metadata in its original form is shown in listing 4.2 – the creator is represented in the `from` field, and the tagged in the `tags` fields, (although it has been modified to protect personal information).

One such association found in the metadata is the *owner/creator* of the document – the semantic meaning of this field depends on the type of document, and is shown in table 5.2, this meaning is interpreted as to whether it indicates event attendance, based on common sense. This is true for photos, e.g. (Fred took a photo at the BBQ) $\implies$ (Fred attended the BBQ), but false for albums: users can add photos to shared albums belonging to other users, they are not assumed to relate to a single underlying event. The creator of the album was not necessarily present at the event(s) relating to the photo(s) within it (this relationship is explored in section 6.8).

Not all cases are trivial: for status messages, the owner (i.e. the person who wrote the message) is assumed to be an event attendee. This is almost certainly true if the status message is geo-tagged (if the creator feels it important to semantically tag a status message, the event is likely occurring "now"). In other cases (also assumed to be true) the status message may be some non-event related message, or referring to an event that the user did not physically attend – investigation is deferred to future work.

The interpretation of the metadata fields/type is summarised in table 5.2. Although it seems trivial, this mapping from document semantics to event semantics is a crucial: it allows the annotations to describe event-level semantics (rather than document-level semantics), and enables high-level comparison of documents for the estimation of event commonality – based on common sense reasoning about events, and agnostic of the particulars of the underlying data.

5.3.4 Implementation: Text

Similarly, mentions of people by name in text in documents might be indicators of event attendance (and thus useful for reasoning about event commonality, or other reasoning). As

[5] Charlie tagged in status message → Charlie was there at the time.

[6] Bob tagged in photo → Bob appears in photo → Bob attended the event

Document Type	Field(s) searched for named users.
DatumAlbum	
DatumCheckin	Check-in Message.
DatumEvent	
DatumPhoto	
DatumStatusMessage	Status Message.

Table 5.3: Text Fields used for Named Entity Recognition of People

with the case of metadata, it is necessary to use heuristics to apply assumptions about the interpretation of mentions: comments were excluded because they often contained conversations between users who were not necessarily event attendees.

A gazetteer-based approach (section 3.2.6.8) is used to recognise and resolve mentions of individuals in text. The list of friends is used as a gazetteer for detecting mentions of people, in the respective user's social media footprint, on selected text fields where mentions of people are thought to be strong indictors of event participation, (the selection is shown in table 5.3).

An implementation detail is that because this list is different for each user, the text is searched separately to the main text-processing pipeline (where all users are processed together, for performance reasons). For each attendee detected, an annotation is added to the document (described below).

5.3.5 Annotation Generation

For each attendee detected with the techniques described above, an annotation is created representing the attendance of an individual at the event related to the document – ready for use in Phase B. A class diagram of the `PersonAnnotation` class is shown in figure 5.2. Tables 5.4 and 5.5 show a summary of the number of `PeopleAnnotation`s found in the source data, and a breakdown of annotations by source document type and data kind. Overall, people-related annotations were found for around one third of documents – with the bulk of annotations originating from Facebook tags. Note that no annotations are created for the user themselves, they are assumed to have an association with all documents in their social networking footprint, hence this information is of no use in clustering.

Table 5.4: Summary of People Annotations.

Total Documents	583
Total People Annotations	1827
Documents with ≥ 1 annotation	559
Mean annotations / document	3.13

Table 5.5: People Annotations by Source Data Kind and Document Type.

Document Type	Text	Metadata	Image	Total Annotations
Status Message	0	27	0	27
Photo	0	1592	0	1592
Check-In	0	120	0	120
Facebook Event	0	88	0	88
Total	0	1827	0	1827

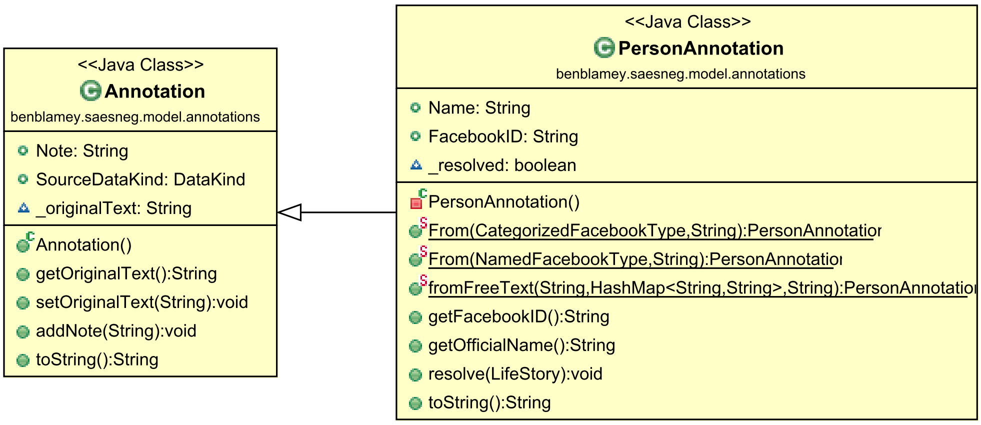


Figure 5.2: The `PersonAnnotation` Class.

5.4 Locations

5.4.1 Introduction

This section describes the implementation of the component of Phase A responsible for detecting associations between documents and geographic locations, and generating annotations representing this information, to support the *spatial* document event commonality strategy in Phase B (section 6.4). The strategy is based on geographic proximity, so as part of the Phase A implementation, any reference to a geographic location is resolved to geographic co-ordinates, so that the Phase B implementation can focus on the similarity calculation. In the initial implementation of *SAESNEG* these associations were generated from two sources:

- Mentions of locations in text fields needed to be detected and resolved.
- Metadata fields known to be locations, which only needed to be resolved.

For the text, as discussed in the literature review, detection of named entities (such as locations) can be achieved with broadly two methods – a gazetteer (i.e. a list of named entities), or a grammar (rules are computed statistically, perhaps with a conditional random field, rather than being coded by hand in state-of-the-art systems).

A decision was made to combine the advantages of both approaches, using a gazetteer in combination with the conditional random field based tagger: using a web-scale gazetteer alone resulted in far too many false-positives to be useful, whereas the tagger alone missed many instances of major cities that could have easily be detected with a gazetteer. As discussed in the literature review, previous attempts at combining gazetteers with taggers have produced mixed results – however, many such studies did not make use of the web-scale gazetteers available today. An alternative resource is used for the resolution of the location metadata fields, which is discussed separately.

5.4.2 Gazetteer Selection (Text)

Traditional gazetteers do not include every named place on earth – GATE, for example, has built-in gazetteer lists, but the longest has only 1989 cities included. There has perhaps been a bias (at least historically) towards the large cities typically mentioned in newswire articles. One could argue that this creates a similar situation to that with the temporal expressions: research focuses on tasks where accuracy can be clearly calculated, sometimes bypassing deeper discussion. The source data in this study is not newswire, it is data from peoples personal social media footprint and mirrors their everyday lives; with these sparse, private events (see section 4.1) it is necessary to extract the maximum amount of information from the source data – ignoring everything but mentions of capital cities is clearly not an appropriate strategy. To this end, the author sought to construct a web-scale gazetteer, of which a number were available:

- GeoNames
- OpenStreetMap
- Wikipedia

With profiles typically containing more than 13,000 words, using an online service as a gazetteer would be prohibitively slow. Wikipedia has been used in a number of experiments, and has good coverage of places – but the extra effort of downloading the large database dump, filtering and normalizing the articles was unnecessary. Both Geonames and OpenStreetMap were integrated into *SAESNEG* after some initial experimentation, OpenStreetMap was selected over GeoNames because of the detailed information it had about places.

OpenStreetMap has a number of advantages over traditional gazetteer lists:

- Exceptional coverage, the number of *ways* approaching 300 million in December 2014.
- Wikified – data is being amended and updated all the time. Mapping efforts are often organised in response to disasters, such as the 2014 Ebola Outbreak [7]. OpenStreetMap has 1.9 million registered users[8].
- Extensive metadata (e.g. boundaries, settlement size, etc.) – this creates an oppurtunity to use metadata to make judgements about whether to match with the text.
- Multilingual – OSN corpora may contain more than one language, an English speaker may use a place name in a local language.
- All geographical entities, buildings, parks – creating an opportunity to widen the scope of the gazetteer in the future.

5.4.3 Experimental Setup & Implementation (Text)

This section describes *SAESNEG*'s implementation of NER for textual data. One of the means by which source documents are annotated with location information. The implementation uses a combination of the gazetteer (backed by the OpenStreetMap resource) and the StanfordNER CRM-based tagger, as described below.

The OpenStreetMap resource contains three different kinds of elements: nodes, ways, and relations, which are used in combination, and then tagged to represent a huge variety of different geographic entities and information – there are tags indicating a car wash, sty, motorway, and tactile paving[9], with a range of metadata.

A dump of the OpenStreetMap world data – 'planet.osm'[10] was downloaded in PBF format[11][12], and then filtered using Osmosis[13] to remove a range of entities outside the scope of the gazetteer. This filtering reduced the size of the planet PBF from 22GB to 491MB. The data was then imported into PostgresQL using Nominatim, a software framework (written in PHP) which is able to import data from OpenStreetMap and store it in a PostgresQL database, from where it is searchable from a web-service wrapper.

After some initial experimentation, to achieve acceptable performance when gazetteering large text documents, *SAESNEG* was then integrated directly with the PostgresQL database (without the web wrapper) – the database was slightly modified some extra indices added,

[7] `http://wiki.osm.org/wiki/2014_West_Africa_Ebola_Response`)

[8] Jan 2015, `http://wiki.openstreetmap.org/wiki/Stats#Registered_users`

[9] `http://wiki.openstreetmap.org/wiki/Map_Features`

[10] `http://wiki.openstreetmap.org/wiki/Planet.osm`

[11] `http://wiki.openstreetmap.org/wiki/PBF_Format`

[12] Using uncompressed XML is prohibitive due to 500GB filesize!

[13] A processing tool for PDF files, see: `http://wiki.openstreetmap.org/wiki/Osmosis`

and some of the search features ported to Java. A cache was also added to *SAESNEG* to improve performance of the lookup. A CentOS virtual machine was used to serve the database (assigned 8GB RAM, 4 CPU Cores). As part of the text processing pipeline, an OpenStreetMap search was performed for every token (as identified by the tokenizer) – by itself, this creates an overwhelming volume of location annotations, due to the extensive and detailed global coverage of the resource. Hence, the text is also processed by the CRM-based StanfordNLP NER component, to identify tokens likely to represent named entities, this extra information can be utilised to support a simple heuristic for the filtering of the OSM results, as shown in a snippet of the `Datum` class.

Listing 5.2: Heuristic for Filtering OSM Locations in Datum (Document) class.

```
boolean passesFilter =
    stanfordEntityType.equals("LOCATION")
    || (
        trimmedTokenText.toLowerCase().equals(r.name.
           toLowerCase())
        && (r.osm_type != OpenStreetMapElementKind.PostCode)
        && (r.osm_sub_class.equals("state"))
       )
    || (
        (stanfordEntityType.equals("ORGANIZATION")
                || stanfordEntityType.equals("PERSON"))
            && (r.admin_level <= 10)
       )
    || (r.admin_level <= 6);
```

Those OSM results which pass the filter heuristic are added as Location annotations for further processing in Phase B. Note that the result includes the `admin_level` field, as the location strategy requires this information. Tables 5.7 and 5.6 show the coverage of the location annotations found in the source data.

Table 5.6: Summary of Location Annotations.

Total Documents	583
Total Location Annotations	10578
Documents with ≥ 1 annotation	322
Mean annotations / document	18.14

Table 5.7: Location Annotations by Source Data Kind and Document Type.

Document Type	Text	Metadata	Image	Total Annotations
Status Message	478	0	0	478
Photo	2494	110	0	2604
Check-In	266	24	0	290
Facebook Event	7158	48	0	7206
Total	10396	182	0	10578

5.4.4 Handling of Location Fields in Metadata

Many of the source documents contain metadata fields which are known to represent location information. For such fields, no detection is necessary – field values simply need to be converted into location annotations. In some cases, it is necessary to geocode street

Document Type	Location Fields	Notes
Album	Location	
Check-in	Place → Location	Check-in is to a ‘place’, place has a ‘location’.
Event	Location	
Link	N/A	
Photo	Place → Location	Check-in is to a ‘place’, place has a ‘location’.
Status Message	N/A	

Table 5.8: Location Metadata Fields in Source Data.

addresses found in these fields into latitude and longitude co-ordinates. Because the locations were in a street address format, rather than using a gazetteer approach, the Google GeoCoding API[14] was used to resolve the addresses. This information is then represented as a `LocationAnnotation`, and added to the document. For such street addresses, an `admin_level` of 11 (i.e. street) is used, according to the OSM scheme, for utilisation by the strategy in Phase B. Table 5.8 shows which fields from which document types were available for processing.

5.4.5 Conclusions

This section has briefly discussed the selection of a gazetteer source, demonstrated a successful and simple proof-of-concept system for combining a state-of-the-art probabilistic tagger (in this case, the Stanford NER CRF tagger) with a web-scale gazetteer, via means of a simple heuristic. These results originating from text are combined with those originating from metadata fields, all results being represented using instances of the Location Annotation for further processing in Phase B. Hence, location-related information was successfully extracted from a variety of document types, and kinds of source data (i.e. text and metadata) – the single `LocationAnnotation` representation facilitates comparison independent of these differing sources, thus helping to overcome the challenges associated with the differently heterogeneous social media footprint (section 1.6).

5.5 Temporal

This section concerns the extraction of temporal information from documents, for use by the temporal strategy in Phase B. The temporal dimension is key to the event clustering task: events always take place at a particular time, and documents contain an abundance of temporal information. As described in section 3.2.7, SUTime, from the StanfordNLP toolkit was selected as the component for extracting temporal expressions from the source social media documents. SUTime was configured to use UK public holidays.

Table 5.10 shows the number of temporal annotations extracted from the documents used in the study, and table 5.11 shows the breakdown by document type.

The temporal dimension is seen as being crucial to the event clustering task: events always take place at a particular time, and documents contain an abundance of temporal information; for this reason, this study has focused heavily on the temporal aspect. Consequently, the research presented in this section is cited among the theoretical contributions of the thesis (section 7.2), which will hopefully find wider applicability outside *SAESNEG*.

[14] `https://developers.google.com/maps/documentation/geocoding/`

The implementation of the temporal extraction components – and this section – hinge upon three requirements which are briefly outlined below. Temporal information can be found in OSN documents within both text and machine-readable metadata fields (e.g. content upload and/or creation timestamps). This means that the temporal component of Phase A must:

Requirement 1: Detect, parse, and ascribe meaning to temporal information found in text – i.e. temporal expressions (a well-studied research problem – section 3.2.7).

Requirement 2: Interpret and ascribe meaning to metadata fields containing temporal information.

After this extraction step, the set of extracted temporal data needs to be converted into a single representation suitable for Phase B (section 1.2) – i.e. the temporal component of Phase A must:

Requirement 3: Combine temporal information from various sources within the document, into a single representation of the temporal information associated with the underlying event, in a way that respects the semantics and context of that information.

Because of the sparsity of the private events, the key importance of the temporal event facet in clustering, and the relatice ambundance of temporal information, there is a motivation to make maximum use of all available temporal information – a key rationale for the techniques presented herein. Making full use of this information (and respecting the subtle semantics) is pivotal to successful event clustering, and lies at the core of this thesis[15].

At a deeper level, this section is really about addressing the challenge of mapping from document temporal semantics to event temporal semantics, i.e. what does temporal information found in a document (whether it be textual or metadata), tell us about the time that the underlying real world event took place?

The remainder of section 5.5 documents the research in addressing this research question, and fulfilling the three software requirements outlined above. Section 5.5.1 describes a novel approach to the representation of temporal information – using distributed semantics to represent temporal information. In existing approaches, temporal semantics are typically represented as discrete ranges or specific dates, and the task is restricted to text that conforms to this representation (section 3.2.7). This section proposes an alternate paradigm: that of *distributed temporal semantics* – where a probability density function models relative probabilities of the various interpretations. The rationale and advantages of this approach to capture more detailed semantics of the temporal information, enable different temporal information to be combined in a natural way via their PDF representations – and allowing a broader range of expressions to be considered temporal expressions (widening the scope of the associated research tasks). Experimental methodology, results, and software implementation are discussed in detail. This approach is crucial to the fulfilment of the three requirements.

Building on this approach, sections 5.5.4 and 5.5.6 describe how temporal information in metadata is interpreted in *SAESNEG* (taking Facebook photos as an example), and then

[15] Image data can also contain temporal information in various forms. EXIF metadata may contain a timestamp, but this is not present in Facebook photos. In some cases, it may also be possible to infer the time of day, or time of year associated with images, based on objects recognised or other image features. For example, a photograph of a sunset, for example, would occur in the evening (although this is ambiguous without knowing where the photograph was taken), whereas a photograph of Santa Claus has obvious temporal association. Because of the limited training data available for sparse private events, and the focus on NLP in this study, integration of support for such techniques is left to future work.

proposes a method for combining multiple instances of temporal information to produce a single temporal annotation (whilst representing the nuances of the semantics implied from the context). Section 5.5.5 briefly touches on deeper philosophical issues[16].

5.5.1 A Distributional Approach to Representing Temporal Information

Temporal expressions communicate more than points and intervals on the real axis of unix time – their true meaning is much more complex, intricately linked to the culture, and often difficult to define precisely. Extracting the temporal semantics of text is important in tasks such as event detection (Ritter et al., 2012). Section 5.5.2 presents a technique for leveraging big data to capture the *distributed temporal semantics* of various classes of temporal expressions (the term *distributed* began to appear in the context of automatic thesauri construction during the 1990s (Grefenstette, 1994)). The approach models the inherent ambiguity of traditional temporal expressions, as well as widening the task to infer semantics from quasi-temporal expressions not previously considered for this task. This work demonstrates how such data can be viewed as a distributed *definition* of the expressions, and that this definition can be incorporated into temporal expression software by modelling it as a probability density function (PDF).

A distributional approach is pursued for three reasons: firstly, a distributed definition can capture a more detailed cultural meaning. Examples from the study show that these common temporal expressions are often associated with instances outside their official, or historical definition. Distributions were found to have a range of skewness and variance, some with more complex patterns exhibiting cultural ambiguity; specific detailed examples are discussed in appendix A: capturing detailed semantics allows fulfilment of the first two key requirements listed in the previous section.

Secondly, the PDF representation naturally allows temporal information to be combined in the same way that events can be composed in Bayesian statistics – rich representations of temporal expressions can be combined in a way that retains the detailed semantics associated with the underlying expressions. This detail can be carried through for later processing, meeting the third requirement listed in the previous section.

Thirdly, the approach allows a much larger range of expressions to be considered as temporal expressions. Under the current paradigm, phrases need to be associated with specific intervals or instances in time. Religious festivals and public holidays can be resolved to their official meaning, but this is not possible for expressions where no single such definition exists. Indeed, there are many expressions that have consistent temporal meaning, without any universal official dates. See appendix A for a discussion of examples in this category, such as “Freshers’ Week” and “Last Day of School” (figure A.5) – this widens the scope of research tasks associated with temporal expressions.

Section 3.2.7 has discussed how existing work has overlooked the distributed semantics issue, section 5.5.2 describes a technique for mining a distributed definition from photo metadata downloaded from the photo-sharing service *Flickr*. Resulting definitions are shown in appendix A, with a discussion of cultural nuances that are found. In section 5.5.3 the changes to the state-of-the-art SUTime framework (part of StandfordNLP) are described. Section 5.5.4 shows how the system is integrated within *SAESNEG* for determining the creation time of Facebook photos – highlighting how the approach facilitates incorporation of a prior probability.

[16] i.e. how the approach relates to the point- and interval- based systems of time used in AI.

The theoretical model defines time $t \in \mathbb{R}$. A temporal expression S is represented by a function $f(t)$, which is a probability density function (PDF) for the continuous random variable T_s (section 5.5.5 for interpretation of this random variable). For the purposes here, a PDF (probability density function) $f(t)$ is defined simply as:

$$P(T_s \geq t_1, T_s \leq t_2) = \int_{t_1}^{t_2} f(t)\, dt \tag{5.1}$$

it follows that:

$$\int_{-\infty}^{\infty} f(t)\, dt = 1 \tag{5.2}$$

and

$$f(t) \geq 0 \, \forall \, t \tag{5.3}$$

In practice, the implementation actually used a smaller, finite date range, suitable for the context. For this paper, a single 'generic year' was used, and focus on handling temporal expressions with date-level granularity.

5.5.2 Mining Definitions

Photographs uploaded to the photo-sharing site Flickr[17], used in numerous other studies, have been used as the basis for the definitions. The Flickr API was used to search for all photos relating to each term uploaded in the year 2012[18]. Metadata was retrieved for each matching photo[19], the 'taken' attribute of the 'dates' element is the photo creation timestamp (Flickr extracts this from the EXIF metadata, if it exists).

The aim for using an online social network as a data source was to build culturally accurate definitions. A photo-sharing service was used because the semantics of the photo metadata would be more closely associated with the timestamp of the photo than would be the case for a status message. Tweets such as "getting fit for the summer", "excited about the summer", "miss the summer", etc, do not reveal a specific definition of the word or phrase in question. Conversely, a photo labelled "Summer", "Graduation" or such like, indicates a clear association between the term and the time the photo was taken. Measuring this association on a large scale yields a distributed definition – literally a statistical model of how society defines the term. Examples for a selection of temporal expressions are shown in appendix

For the initial system, only temporal expressions for which the patter was expected to repeat on an annual basis were studied. To some extent, this obviates some of the error in the photo timestamps, inevitably originating from inaccurate camera clocks, and timezone issues. A similar approach should be suitable for creating definitions at other scales of time.

Having collected a list of timestamps for each term, there was a need to find a probability density function to provide a convenient representation, and smooth the data appropriately. An added complication is that mapping time into the interval of a single year creates an issue when trying to fit, say, a normal distribution to the data. The concentration of probability density may lie very close to one end of the interval (e.g. "New Year's Eve"), which means contributions from peaks of probability density that lie in neighbouring years in such cases cannot be neglected.

[17] http://www.flickr.com

[18] Using the endpoint described at: http://www.flickr.com/services/api/flickr.photos.search.html

[19] Using the endpoint described at: http://www.flickr.com/services/api/flickr.photos.getInfo.html

Generally, the timestamps were arranged in distinct clusters, so frequencies were computed for 24-hour intervals, and then attempted to fit mixture models to the data, using the expectation-maximisation process, implemented in the Accord.NET scientific computing framework (Souza, 2012), shown as pseudocode in listing 5.3. Initial attempts used a mixture of von Moses distributions, a close approximation to the *wrapped normal distribution*, the result of wrapping the normal distribution around the unit circle. There were difficulties reaching a satisfactory fit with this model, so instead a mixture of normal distributions was used, adapted to work under modulo arithmetic (using the so-called *mean of circular quantities*). Hence, the probability density greater than ±6 months away from the mean is neglected, for each normal distribution in the mixture. With standard deviations typically in the region of a few days, this is reasonable.

Listing 5.3: Maximum Entropy Algorithm (pseudo code) – as Implemented in Accord.NET

```
// Define functions and constants.
function computeProb(observation, model)
    { return probability }
function computeModel(observations, model)
    { return [ (weight,mean,variance), ... ] }
const threshold, iterlimit

// Initialize mixture model and observed frequencies.
mixture ← [ (weight,mean,variance), ... ]
observations ← [f1, ..., fn]

DO {
    // EXPECTATION STEP (model → data).
    foreach (frequency in observations) { // increment i
        foreach (model in mixture) { // increment j
            γ[i][j] ← weight * computeProb(observation)
        }
    }

    // MAXIMIZATION STEP (model ← data).
    foreach (model in mixture) { // increment i
        // Estimate new model mean, variance and weight from
           the
        //  γ values for that model.
        (weight, mean, variance) ← computeModel(γ[i][], model)

        // (Models that made little overall contribution
        //  have their weight reduced accordingly.)
    }

    // CHECK FOR CONVERGENCE
    loglikelihood ← log(mixture)
} WHILE ((#iterations < iterlimit) AND (Δloglikelihood >
   threshold))
```

After mixed results using k-means clustering to initialise the model, a uniform arrangement of normal distributions was used. A uniform distribution was also included to model the background activity level – without this, because the normal distributions had standard deviations of just a few days, fitting was disrupted by the presence of many outliers.

After fitting, normal distributions with a mixing coefficient of less than 0.001 are pruned from the model.

5.5.3 Modifications to SUTime

When modifying the SUTime framework (Chang and Manning, 2012), the aim was to preserve the existing functionality, as well as implement the distributed approach. A number of Java classes are used to represent the parsed temporal information, the key modification was to augment these classes so that they stored a representation of a probability distribution alongside their other fields. Modifications were then made to the grammar definitions to ensure that instances of these classes were associated with the appropriate probability distributions upon creation, and updated appropriately during grammatical composition.

To explain the core temporal classes in more detail: where appropriate, a class field was added which could optionally hold an object representing the associated probability distribution. Effectively, this object is a tree whose nodes are instances of various new classes: `AnnualNormalDistribution`, `AnnualUniformDistribution`, (as the leaves of the tree), and those representing either a Sum or Intersection (i.e. multiplication) as the internal nodes. When no distributed definition was available (e.g. when parsing “2012”), a representation of the appropriate discrete interval is used as a leaf[20]. These classes implemented a method to return an expression string suitable for use in gnuplot (visible in the online demo[21]) – with the two internal nodes algebraically composing the expressions returned by their children in the obvious way. Generation of an alternative syntax, or support for numerical integration could be implemented as additional methods. A new temporal class was introduced, to represent a temporal expression which does not have a non-distributed definition (such as “Last Day of School”), for which composition is possible under the distributed paradigm, but which uses a dummy implementation under the traditional paradigm.

Secondly, it was necessary to make a number of changes to the grammar definitions – these files control how instances of temporal classes are created from the input text, and also how the instances of these classes are combined and manipulated based on the underlying text. After fitting the mixture models to the Flickr data (section 5.5.2), definitions were generated in the syntax used by SUTime. Rules defining the initial detection of these expressions were updated so that the probability distributions were included, and modified to allow misspellings and repeated characters that were found in Facebook photo album names (common to online social networks (Brody and Diakopoulos, 2011)). Rules were introduced to detect the new temporal expressions, and assign their distributed definitions. Other modifications were made to adapt the grammar to the domain of photo album names, relating to British English date conventions, and to support temporal expressions of the form ’YY. The rules for temporal composition were largely unchanged, as they are expressed in terms of the temporal operators defined separately.

SUTime defines 17 algebraic operators for temporal instances (e.g. THIS, NEXT, UNION, INTERSECT, IN). Facebook photo album names tended to contain mostly absolute temporal expressions (none of the form “2 months”, or “next week”), and it was only necessary to modify the INTERSECT operator. In the distributed paradigm, intersecting two temporal expressions such as “Xmas” and “2012” is simply a case of multiplying their respective probability density functions. The existing implementation of the operator is unaffected. Adaptation of ‘discrete’ operators such as PREV and NEXT into the distributed paradigm presents an interesting problem, and is left for future work. All that remained was to include an expression for the final probability density functions in the TimeML output[22].

[20] $tm\_year(x)$ and related gnuplot functions were useful for this (Williams and Kelley, 2013, p. 27).

[21] `http://benblamey.com/tempex`

[22] An ‘X-GNUPlot-Function’ attribute was inserted into the TIMEX3 element for this purpose

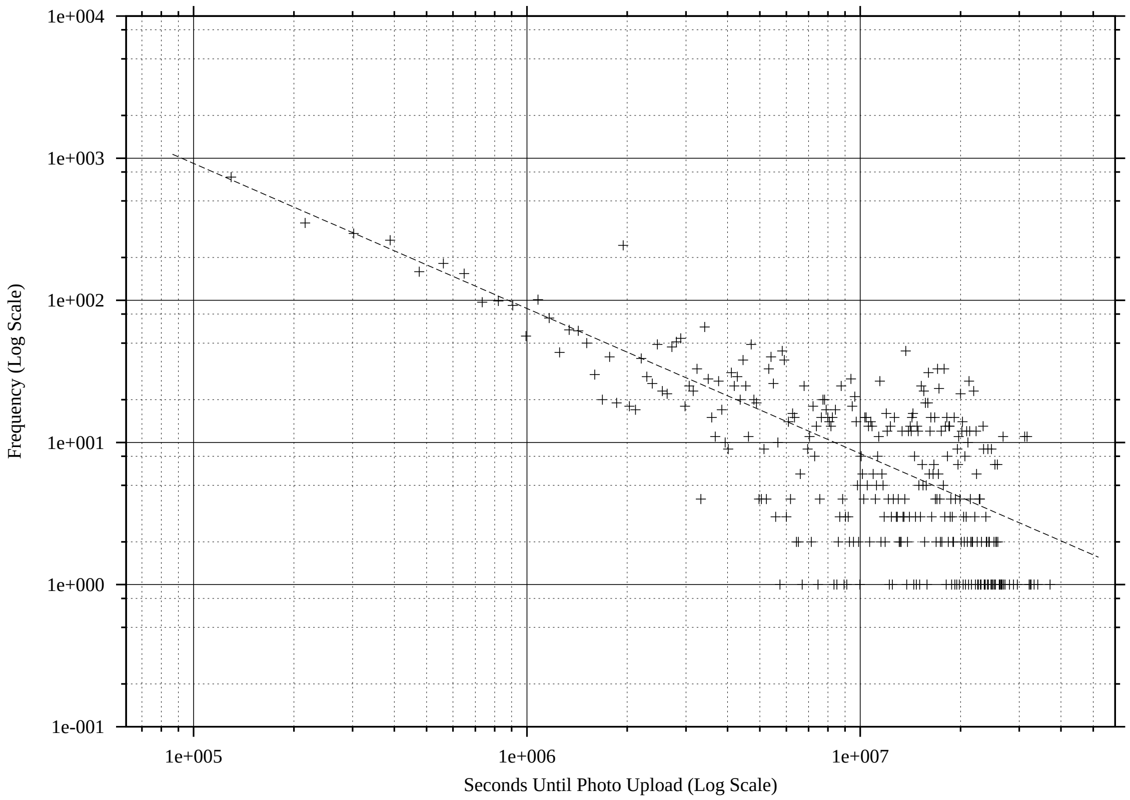


Figure 5.3: Distribution of the *Upload Delay*, estimated from Flickr photo metadata.

5.5.4 Temporal Information in Documents: A Worked Example

In tasks such as event detection, it is useful to know the time that a photograph was taken. In Facebook, the EXIF metadata is removed for privacy-related reasons (James, 2011), and the API does not publish the photo creation time (as is the case with Flickr). In Facebook, photo album titles tend to be rich in temporal expressions, and an album title such as "Halloweeeeennnn!" should indicate the date the photo was originally taken, even if it was not uploaded to Facebook until later. The usual technique would be to parse the temporal expression and resolve it to its 'official' meaning; in this case, October 31st.

Figure A.3 shows that some of the probability density for "Halloween" actually lies before this date – peaking around the 29th (although the effect is greater with "Christmas", figure A.2). Having represented the temporal expression as a probability density function, it can be combined with a prior probability distribution, computed as follows. The photo metadata collected from Flickr (section 5.5.2), contains an upload timestamp[23] in addition to the photo creation timestamp. The *upload delay* is defined to be the time difference between the user taking the photo, and uploading it to the web. Figure 5.3 shows the distribution (tabulated into frequencies for 24-hour bins), plotted using a log-log scale. Taking y as the frequency, and x as the upload delay in seconds, the line of best fit was computed (with gnuplot's implementation of the Levenberg-Marquardt[24] algorithm) as:

$$log(y) = a\ log(x) + b \tag{5.4}$$

[23] The time when the photo was uploaded to the web, shown as the 'posted' attribute of the 'dates' element, see: `http://www.flickr.com/services/api/flickr.photos.getInfo.html`

[24] `https://en.wikipedia.org/wiki/LevenbergŰMarquardt_algorithm`

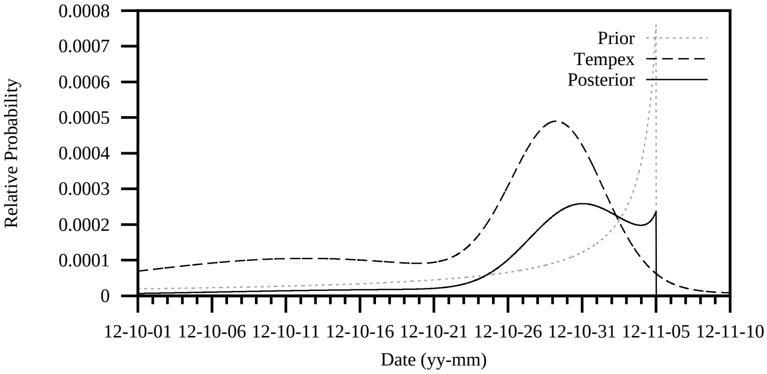


Figure 5.4: Computation of the posterior probability distribution for the creation time of the photo, from the prior probability, and the distribution associated the temporal expression "Halloweeeeennnn!".

with:

$$a = -1.0204 \tag{5.5}$$

$$b = 18.5702 \tag{5.6}$$

This equation can then be used as a prior distribution for the *creation* timestamp of the photo, by working backwards from the *upload* timestamp, which is available from Facebook[25]. Figure 5.4 shows this prior probability, the distribution for "Halloweeeeennnn!", and the prior distribution obtained by multiplying them together, respectively scaled for clarity. The resulting posterior distribution has much greater variance than what would have resulted from simply parsing the official definition of October 31st, accounting for events being held on the surrounding days, whilst the application of the prior probability has resulted in a cut-off and a much thinner tail for earlier in the month.

5.5.5 Deeper Issues with Time: Interpreting $S \sim T_s(t)$

Section 5.5.1 discussed the association (denoted by $\sim$) between the temporal expression S and the continuous random variable $T_s(t)$. Detailed discussion is beyond the scope of this work, but a few interpretations are briefly outlined:

1. S represents some unknown instant: $S \sim t_s$. $T_s(t)$ models $P(t = t_s)$.
2. S represents some unknown interval: $S \sim I_s = [t_a, t_b]$. $T_s(t)$ models $P(t \in I_s)$.
3. The meaning of S is precisely $S \sim T_s(t)$, and only by combining T_s with additional information can anything further be inferred. Particular instances or time intervals may have cultural or historical associations with S, it may be possible to recognise their effect on T_s. But T_s itself is the pragmatic interpretation of S.

There is extensive discussion relating to the models underlying (1) and (2) in the literature, and one can construct various thought experiments to create paradoxes in either paradigm. By constructing the probability distribution by modelling time as $\mathbb{R}$ which means the approach is undeniably using the classical point-based model of time, rather than the

[25] See the 'created_time' field at: https://developers.facebook.com/docs/reference/api/photo/

interval-based model (Allen, 1981). However, the situation is a little more subtle: employing a probability density function only allows computation of the probability associated with an arbitrary *interval*. For a continuous random variable, the probability of any particular instance is zero by definition; something which is arguably more akin to an interval-based interpretation of time. So, associating the temporal expression with a PDF means that the theoretical basis of the point-based system of time is retained, whilst the mathematics restricts us to working only with intervals. Whether this dual-nature obviates the dividing instant problem (Ma and Knight, 2003), requires a more rigorous argument, and is left to future work.

The thrust of the contribution is to suggest that temporal expressions in isolation are intrinsically ambiguous[26] (interpretation (3)). It is argued that that such expressions cannot be resolved to discrete intervals or instants (without loss of information), and attempts to do so are perhaps unnecessary or misguided. In some cases, it may be desirable to defer resolution, perhaps to apply a prior probability (as in section 5.5.4).

5.5.6 Integration into *SAESNEG* (Non-)Definitive Temporal Information

Having described the modifications to the SUTime NLP framework to incorporate the distributed approach to representing temoral expressions, discussion moves to how the framework with its modifications can be incorporated into SUTime. The first issue is that of content selection: what input data is subject to temporal analysis. For the initial implementation of *SAESNEG* text was processed through the usual text pipeline – which included an SUTime component (the details of which are described above). Metadata was inspected and temporal information was extracted and converted into a representation according to the temporal approach. No temporal information was extracted from image content data in the initial implementation of *SAESNEG* although the future work section (7.6) discusses ways this might be done.

The remaining discussion in this section concerns (a) the details of the representation of temporal information – the output of Phase A, namely the `TemporalAnnotation` class, (b) the handling temporal information found in metadata. Section 6.5 explains how the sets of temporal annotations associated with documents are combined and compared in order to compute document event similarity. Concerning representation, the `TemporalAnnotation` class includes the usual annotation information such as the source data kind, the TIMEX annotation originating from SUTime (if any), the original text (if any) and so forth. The core of the temporal representation is the `TimeDensityFunction`, which represents the temporal information in two ways:

1. A expression representing the PDF which is used to plot the PDF in gnuplot – this is then used to generate plots of the functions for display in web interfaces for debugging and engineering purposes.
2. A floating point array containing samples of the PDF on a day-by-day basis, within an arbitrary defined sample range (Jan 1st 2005 to Dec 31st 2014), normalised so that it sums to 1. Annotations can then be combined and compared (see section 6.5) using this array of samples. Implementing the normalisation and similarity calculation of PDFs through the means of numerical integration was thought to be simpler than the book-keeping involved in performing algebraic integration of the combinations of Gaussian mixture modules under modulo arithmetic.

[26] The "weekend", and *precisely* when it starts, is a good example of this. Readers will be able to imagine many different possible interpretations of the word.

Source Data	Definitive?	Distribution
[Text]	No	Various.
Photo Uploaded	Yes	Photo Upload Distribution.
Mobile Photo Upload	Yes	Live – Single Day.
Check-in Timestamp	Yes	Live – Single Day.
Event Timestamp	Yes	Live – Single Day.
Status Message Timestamp	No	Live – Single Day.

Table 5.9: Interpretation of Temporal Information based on Context

The remaining detail of the representation of the extracted temporal information is an `isDefinitive` flag indicating is its relationship to the time of the underlying real-world event – i.e. whether or not time in the annotation is known to indicate the time of the event. Whether temporal information indicates definitive information about the time of the underlying event is a matter of interpretation (it is argued), based on the context of the originating document type, and the source field. For example, the timestamp associated with a `CheckinDatum` can be interpreted as definitive – the user would have been at the physical location of the check-in (at that time) in order to create it. However, this is not necessarily the case for temporal expressions discovered in text, as this example of a school reunion shows:

HOWELL'S HALF A DECADE REUNION LEAVERS OF 2006

Whilst 2006 is recognised (correctly) as a temporal expression, the school reunion itself did not occur in 2006. The handling of these definitive and non-definitive expressions is discussed in the chapter 6. Table 5.9 shows the whether text and metadata fields are interpreted as being definitive. Whether temporal information is definitive is subtly related to the notion of a 'live document': a live document is one where document is created during the event – and hence the timestamp can be used to unequivocally indicate the event time (an example of definitive temporal information). However, the distribution associated with the photo upload timestamp is definitive, because it models the probability distribution of the underlying event – but a photo is a not 'live' document because it may be uploaded after the event. Non-definitive temporal information may have no association at all with the time of the event, and refer to something else entirely – it is clearly difficult for the system to tell the difference between such temporal expressions.

Table 5.10: Summary of Temporal Annotations.

Total Documents	583
Total Temporal Annotations	918
Documents with ≥ 1 annotation	560
Mean annotations / document	1.57

5.5.7 Conclusions

This section has described the work related to the extraction of temporal information from source documents in Phase A of *SAESNEG*. By critiquing the existing paradigm in the

Table 5.11: Temporal Annotations by Source Data Kind and Document Type.

Document Type	Text	Metadata	Image	Total Annotations
Status Message	5	27	0	32
Photo	52	497	0	549
Check-In	4	24	0	28
Facebook Event	220	89	0	309
Total	281	637	0	918

literature review, this section set out alterative paradigm of distributed representation of temporal expressions. A methodology has been described for the creation of such definitions based on social networking data, and the approach has been incorporated into the StandfordNLP text pipeline – this section forms a substantial contribution of the thesis, making a number of contributions:

- The main contribution is a proposal for an alternative *distributed* paradigm for parsing temporal expressions (section 5.5.1). The approach has several advantages:
  - It is able to provide definitions for a wider class of temporal expressions, supporting expressions where there is no single official definition.
  - It captures greater cultural richness and ambiguity – arguably a more accurate definition.
  - It facilitates further processing, such as the consideration of a prior probability, as demonstrated with an example in section 5.5.4.
- A technique for mining definitions from a large dataset has been demonstrated, and statistically modelling the results to create a distributed definition of a temporal expression (section 5.5.2).
- A state-of-the-art temporal expression software framework has been adapted to incorporate the distributed paradigm, allowing some of the temporal algebraic operators to be implemented as algebraic operators (section 5.5.3).
- This work has then been integrated into the *SAESNEG* pipeline to support the wider goal of performing event clustering on an individual's social media footprint. To this aim, the approach has also been used to represent temporal information derived from non-text sources, such as modelling the likely photo creation time based on the upload time.

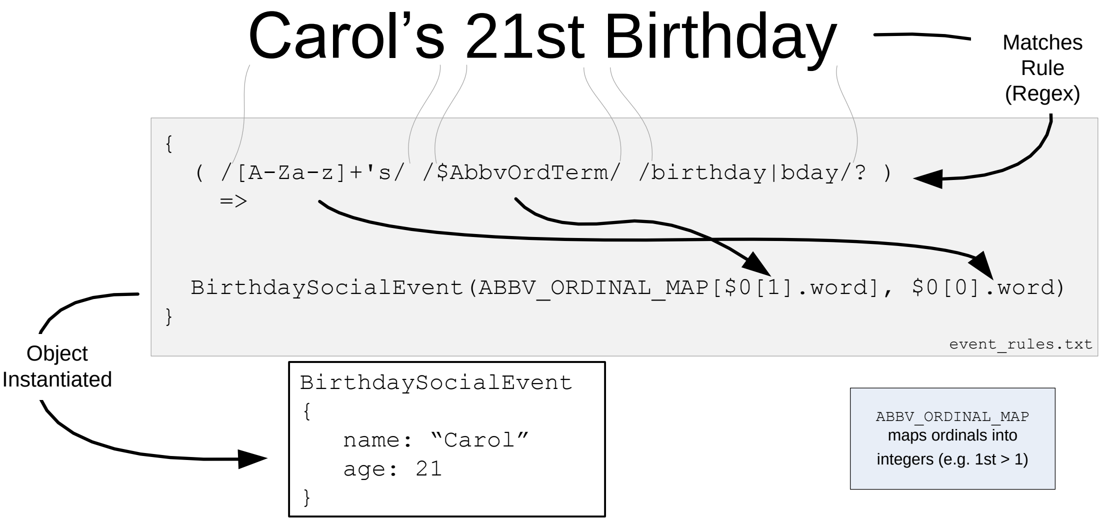
5.6 Social Events

5.6.1 Introduction

This section discusses a novel social events annotator developed for this study, which is able to detect mentions of *well-known* social events (such as birthdays, weddings) in test, and capture details of their semantics. Many social networking documents mention such social events[27], and the information can be extremely useful for computing document event commonality later in the pipeline (especially when social event details are available) 6.6, as well as being useful for the presentation and summarisation in the user interface. To the author's knowledge, there has been no research investigation on this task to date.

5.6.2 Implementation

The social event extraction is applied only to textual data, as is implemented as a new annotator for the Stanford CoreNLP pipeline – mentions of social events were often found in event descriptions, and photo album titles, etc. Parsing rules were written in the *Stanford TokensRegex*[28] representation (Chang and Manning, 2014) – a powerful and flexible language for describing recognition and extraction rules. This textual representation is converted into a parser at runtime, through which the text from the documents is parsed as part of the wider text-processing pipeline (section 5.2), figure 5.5 shows an example of a parsing rule.



```
{
  ( /[A-Za-z]+'s/ /$AbbvOrdTerm/ /birthday|bday/? )
    =>

  BirthdaySocialEvent(ABBV_ORDINAL_MAP[$0[1].word], $0[0].word)
}
```

```
BirthdaySocialEvent
{
   name: “Carol”
   age: 21
}
```

Figure 5.5: Example Parsing Rule

The annotations are exported to GATE for inspection and debugging. In *SAESNEG* the social event annotations are read from the text, and converted into social event annotations for use by the social events strategy in Phase B, section 6.6. As a proof of concept, support for two kinds of well-known events was implemented (these were common in the source data of the participants, to the demographics of the study participants). The rules used to

[27] The term "social event' is used to refer specifically to particular well-known event classes, such as weddings and birthdays – and their mentions in text (as well as detection, parsing and representation thereof). This is to differentiate from the term 'event' – used throughout the thesis to refer to document clusters, and the associated underlying real world events. Clearly, many 'events' are indeed 'social events' – but the latter refers only to explicit mentions in text.

[28] `http://nlp.stanford.edu/software/tokensregex.shtml`

detect mentions of these events is shown in listing 5.6.2. The same approach is used for the SUTime temporal expression parser used in section 5.5.

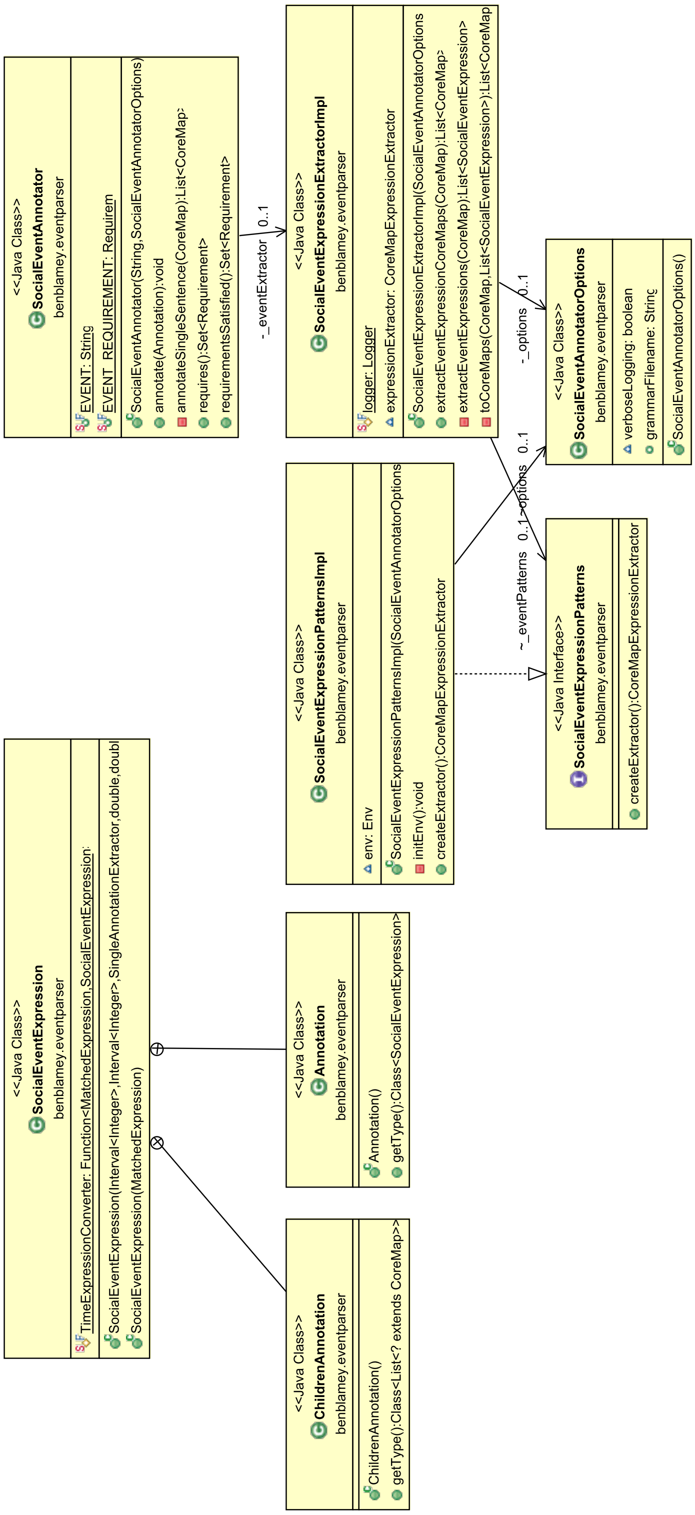
Listing 5.4: event_rules.txt – Social Event Rules

```
############### BIRTHDAYS ###############
BirthdaySocialEvent = { type: "CLASS", value: "benblamey.
   saesneg.model.annotations.socialevents.
   BirthdaySocialEventAnnotation" }

# Remember - regexes are case sensitive!
# It doesn’t seem to be possible to bind to groups inside
   tokens - binding groups are intra-tokens.
# $X refers to the xth intra-token group.
# For Birthdays ctor - first param is age, subs. params are
   birthday owners.

{ ( /[A-Za-z]+’s/ /$AbbvOrdTerm/ /birthday|bday/? )
                   => BirthdaySocialEvent(ABBV_ORDINAL_MAP[$0
                      [1].word], $0[0].word) }
                       { ( /$AbbvOrdTerm/ /birthday|bday/ )
                   => BirthdaySocialEvent(ABBV_ORDINAL_MAP[$0
                      [0].word]) }
{ ( /[a-zA-Z]+’s/ /birthday|bday/ )
               => BirthdaySocialEvent(-1, $0[0].word) }
{ ( /Birthday/ ) => BirthdaySocialEvent() }
{ ( /my/ /$AbbvOrdTerm/ /(birthday|bday).*/? )
               => BirthdaySocialEvent(ABBV_ORDINAL_MAP[$0[1].
                  word], "WHO_OBJECT_CREATOR") }
{ (  /my/ /($AbbvOrdTerm)/ ) => BirthdaySocialEvent(
   ABBV_ORDINAL_MAP[$1[1].word], "WHO_OBJECT_CREATOR") }

############### WEDDINGS ###############
WeddingSocialEvent = { type: "CLASS", value: "benblamey.saesneg
   .model.annotations.socialevents.WeddingSocialEventAnnotation
   " }
{ ( /[A-Za-z]+/ /and/ /[A-Za-z]+’s/ /wedding/ )
   => WeddingSocialEvent($0[0].word, $0[2].word) }
{ ( /[A-Za-z]+/ /(and|\+|\&)/ /[A-Za-z]+’s/ /wedding/ )
   => WeddingSocialEvent($0[0].word, $0[2].word) }
{ ( /wedding/ /of/ /[A-Za-z]+/ /(and|\+|\&)/ /[A-Za-z]+/ )
   => WeddingSocialEvent($0[2].word, $0[4].word) }
{ ( /wedding/ /ceremony/ /of/ /[A-Za-z]+/ /(and|\+|\&)/ /[A-Za-
   z]+/ )
   => WeddingSocialEvent($0[3].word, $0[5].word) }
{ ( /[A-Za-z]+’s/ /wedding/ /ceremony/? )
   => WeddingSocialEvent($0[0].word) }
```

The various classes created for the annotator are shown in table 5.12, consistent with the design of other annotators – see the associated class diagram shown in figure 5.6. The annotations generated by the annotator are shown in figure 5.7, which are used subsequently in Phase B.

5.7 Results and Discussion

Rich information was detected from the text, listing 5.6.2 shows multiple rules for the semantics of particular classes of social event. For example, rules for extracting birthday related information included the age and the name of the person having the birthday, whilst weddings included the names of the people getting married. Such information is not always clear from the mention (in the simplest case, the mention may simply consist of a single

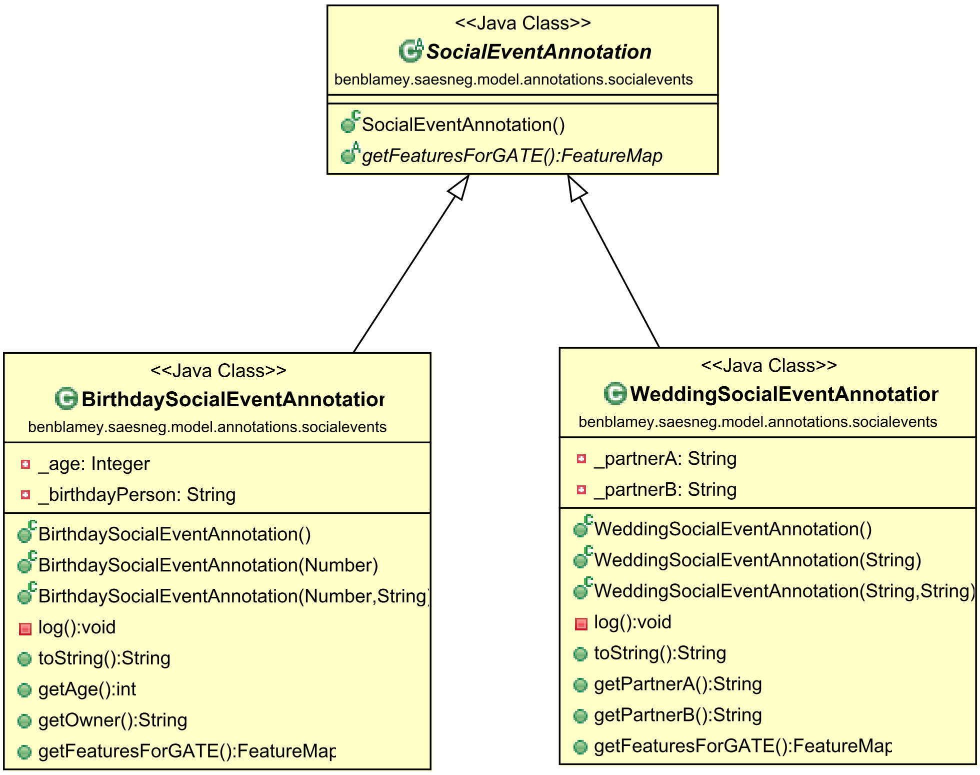


<<Java Class>>
SocialEventAnnotator
benblamey.eventparser
EVENT: String
EVENT_REQUIREMENT: Requirem
SocialEventAnnotator(String,SocialEventAnnotatorOptions)
annotate(Annotation):void
annotateSingleSentence(CoreMap):List<CoreMap>
requires():Set<Requirement>
requirementsSatisfied():Set<Requirement>
-_eventExtractor 0..1
<<Java Class>>
SocialEventExpressionExtractorImpl
benblamey.eventparser
logger: Logger
expressionExtractor: CoreMapExpressionExtractor
SocialEventExpressionExtractorImpl(SocialEventAnnotatorOptions
extractEventExpressionCoreMaps(CoreMap):List<CoreMap>
extractEventExpressions(CoreMap):List<SocialEventExpression>
toCoreMaps(CoreMap,List<SocialEventExpression>):List<CoreMap
-_options 0..1
<<Java Class>>
SocialEventAnnotatorOptions
benblamey.eventparser
verboseLogging: boolean
grammarFilename: String
SocialEventAnnotatorOptions()
~_eventPatterns 0..1
~options 0..1
<<Java Class>>
SocialEventExpressionPatternsImpl
benblamey.eventparser
env: Env
SocialEventExpressionPatternsImpl(SocialEventAnnotatorOptions
initEnv():void
createExtractor():CoreMapExpressionExtractor
<<Java Interface>>
SocialEventExpressionPatterns
benblamey.eventparser
createExtractor():CoreMapExpressionExtractor
<<Java Class>>
SocialEventExpression
benblamey.eventparser
TimeExpressionConverter: Function<MatchedExpression,SocialEventExpression>
SocialEventExpression(Interval<Integer>,Interval<Integer>,SingleAnnotationExtractor,double,doubl
SocialEventExpression(MatchedExpression)
<<Java Class>>
Annotation
benblamey.eventparser
Annotation()
getType():Class<SocialEventExpression>
<<Java Class>>
ChildrenAnnotation
benblamey.eventparser
ChildrenAnnotation()
getType():Class<List<? extends CoreMap>>

Figure 5.6: Class diagram of the *Social Event Annotator*.

Class	Description
`SocialEventAnnotator`	Added to the pipeline to parse for social events.
`SocialEventAnnotatorOptions`	Options passed to the annotator.
`SocialEventExpression`	Internal representation of mentions of social events in text.
`SocialEventExpressionExtractorImpl`	Implements core functionality.
`SocialEventExpressionPatterns`	Factory – creates extractor from grammar file. (Interface).
`SocialEventExpressionPatternsImpl`	Factory – creates extractor from grammar file. (Implementation).

Table 5.12: Classes comprising the Social Event Annotator.



<<Java Class>>
SocialEventAnnotation
benblamey.saesneg.model.annotations.socialevents
SocialEventAnnotation()
getFeaturesForGATE():FeatureMap
<<Java Class>>
BirthdaySocialEventAnnotatior
benblamey.saesneg.model.annotations.socialevents
_age: Integer
_birthdayPerson: String
BirthdaySocialEventAnnotation()
BirthdaySocialEventAnnotation(Number)
BirthdaySocialEventAnnotation(Number,String
log():void
toString():String
getAge():int
getOwner():String
getFeaturesForGATE():FeatureMap
<<Java Class>>
WeddingSocialEventAnnotatior
benblamey.saesneg.model.annotations.socialevents
_partnerA: String
_partnerB: String
WeddingSocialEventAnnotation()
WeddingSocialEventAnnotation(String)
WeddingSocialEventAnnotation(String,String)
log():void
toString():String
getPartnerA():String
getPartnerB():String
getFeaturesForGATE():FeatureMap

Figure 5.7: *SAESNEG* Annotations generated by the Social Event Annotator

word such as 'birthday').

The advantage of extracting this extra information is that it facilitates more conclusive comparison for the purposes of computing document event commonality. For example, if two documents mention a birthday where the mention includes the name of the person and the age – and the information matches – then it is fairly clear the two documents refer to the same event, unless there is strong evidence to the contrary. Conversely, birthdays associated with a different age are certain to be different events.

Even in cases where such detailed information is not available, detecting the kind of event can often be sufficient to conclude that the documents relate to different events i.e. a real world event is never simultaneously a wedding and a birthday[29]. Regardless of whether cases are conclusive, during pairwise comparison in case B, evidence from different strategies are combined before the final pairwise judgements are made i.e. two social events of class 'birthday' attended by the user at a similar place and time would likely be the same event. In light of this, the despite the low proportion of documents for which the annotator was able to detect social event annotations, (around 4%, see table 5.13) the annotator can theoretically enable document-event-commonality judgement decisions to made with a high degree of certainty in these cases. Future work to expand the ruleset and classes of detected social events would hopefully improve this coverage. As table 5.14 shows, mentions of social events can be detected in various document types.

Regarding future work, it is easy to see how the cases of supported events and their specific semantics could be extended for other kind classes of events, based on existing schemas[30] – this task is left for future work. A potential avenue for expansion would be to build the set of rules automatically based on machine readable schema information.

As mentioned, for the initial *SAESNEG* implementation, detection of well-known event semantics is limited to text. Indeed, where metadata is available, appropriate schema alignment could allow for comparison of event related information originating from metadata as well. Similarly, image content based techniques could theoretically be trained to recognise specific event types based on objects appearing in the photo.

The approach is an example of the general approach taken to data-mining to support the document event commonality research task addressed in this thesis – extracting maximum detail from the source data to achieve conclusive comparison, overcoming the data sparseness and the differently heterogeneous nature of the data. High level comparison between social events is performed independently of the underlying type of document and source data kind (even though the implementation of component is based exclusively on extracting text information; future work has been discussed to extend this to other data kinds such as image and metadata).

Table 5.13: Summary of Social Event Annotations.

Total Documents	583
Total Social Event Annotations	51
Documents with ≥ 1 annotation	31
Mean annotations / document	0.09

[29] Notwithstanding the low probability of peculiar edge cases.

[30] http://schema.org/Event

Table 5.14: Social Event Annotations by Source Data Kind and Document Type.

Document Type	Text	Metadata	Image	Total Annotations
Status Message	0	0	0	0
Photo	11	0	0	11
Check-In	0	0	0	0
Facebook Event	40	0	0	40
Total	51	0	0	51

5.8 Summary

This section has outlined proof of concept of a novel component for the detection of mentions of social events in text, with the extraction of metadata associated with those events. This component provides instances of `SocialEventAnnotation` to support the Phase B counterpart (see section 6.6) for the comparison of documents based on available social event related information.

5.9 Image Content

5.9.1 Introduction

Documents (such as photos) often contain image information, (45% of documents selected as ground truth by participants were photos – see table 4.4) and these photos can potentially contain a variety of information that may indicate or refute document event commonality. At a low level, photos with similar image content (e.g. colours, layout) are more likely to be photographs of the same scene, so are perhaps more likely to relate to the same real world event. Whilst at a high level, it may be possible to recognise people, landmarks, or other features in the photos which may give clues for the computation of document event commonality in this application. This chapter focuses on Phase A, the extraction of annotations from individual documents – this section concerns the extraction of annotations from image content specifically. The subsequent comparison of annotations from different documents for the computation of document event commonality is performed within Phase B, and is discussed in chapter 6 accordingly (for the use of image features in Phase B, see section 6.7). This section discusses the selection of image feature extraction techniques, their implementation and integration into the SAESEG architecture.

5.9.2 Choice of Extraction Techniques

Examples of some high- and low-level extraction techniques used for similar applications are discussed in the literature review (section 3.3). For initial implementation of SAESEG, a decision was made to focus exclusively on low-level image similarity comparison: given the choice of source data from Facebook, and the nature of the associated events, it was felt that there was more scope for extracting useful information from text and metadata than perhaps would be the case with image-based extraction techniques, and that instead, text would be a greater focus for the study.

For example, in the case of facial recognition (section 3.3.1), given that Facebook photos are typically 'tagged' i.e. metadata accompanies the image within the photo document describing those people appearing in the photo, there is little motivation for integrating a facial recognition system[31].

As discussed (section 3.4.5), some studies approach the event clustering task by using a large existing dataset; such as photos of famous landmarks, as source of groud truth. Given a typical social media footprint, what proportion of typical photos actually include such landmarks remains to be determined, but was not high based on personal experience. It is possible to find many thousands of photos online of the Eiffel Tower (and hence train for image recognition accordingly), but managing to get a machine classifier to recognise the interior of a given nightclub; a particular suburban backyard; or my next door neighbour's cat is perhaps optimistic, given the much lower availability of annotated public source data for such examples.

This thesis focuses on techniques appropriate for the typical social media footprint, and is hence critical when selecting techniques demonstrated on datasets which are not a representative sample thereof. Given the broad scope of the project, and the focus on text and metadata-based extraction techniques, it was felt that the inclusion of low-level image similarity techniques would be sufficient to demonstrate the use of information derived from image content for the document event commonality task. The architecture of *SAESNEG*

[31] Facebook actually use such a system to suggest these 'tags'.

has been designed to allow future incorporation of components to perform such extraction. In light of this decision, a number of low-level image extraction techniques were selected, choosing implementations from the popular MPEG-7 standard, as discussed in the literature review: dominant color[32], color histogram, scalable color, color layout, and edge histogram.

5.9.3 Implementation

In the future, as support is added for a wider variety of social networks and document types, documents containing image data can be selected for processing, For the initial implementation of *SAESNEG* image content was selected from the documents of type Photo, the full-size image was used.

A benefit of using the MPEG-7 standard for the purposes of image comparison, is that there exist various library implementations of the languages involved. The *Caliph-Emir* library was selected because it is open source and written in Java, making its integration into *SAESNEG*'s primarily-Java code-base relatively easy. Images were downloaded from the Facebook API for the `PhotoDatums`, and the Caliph-Emir library was used to extract the chosen image descriptors. These were then added to the documents in the form of *SAESNEG* annotations, simple wrappers around Caliph-Emirs implementation of the MPEG-7 descriptors. A class diagram showing the output annotations is shown in figure 5.8. Hence, the annotations associated with different document images can be compared during Phase B. For performance reasons, a cache (backed by mongoDB) was used to store extracted image descriptors, so that these did not need to be extracted each time during repeated executions of the experiment. Annotations were extracted from all the images associated with the photo documents in the source data, as shown in table 5.15.

Table 5.15: Summary of Image Content Annotations.

Total Documents	583
Total Image Content Annotations	429
Documents with ≥ 1 annotation	415
Mean annotations / document	0.74

[32] Using the American English spelling from the MPEG-7 standard.

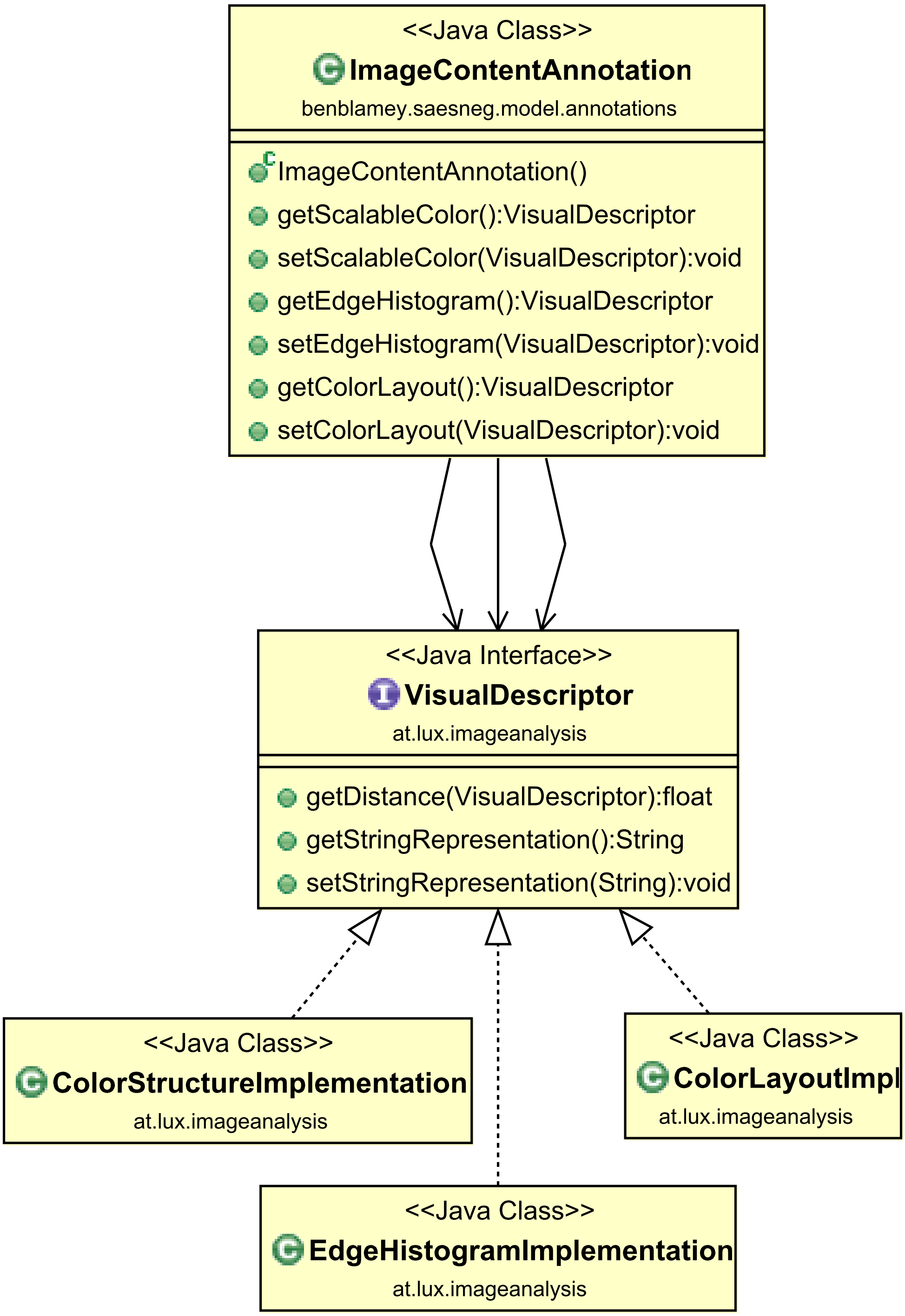

<<Java Class>>
ImageContentAnnotation
benblamey.saesneg.model.annotations
ImageContentAnnotation()
getScalableColor():VisualDescriptor
setScalableColor(VisualDescriptor):void
getEdgeHistogram():VisualDescriptor
setEdgeHistogram(VisualDescriptor):void
getColorLayout():VisualDescriptor
setColorLayout(VisualDescriptor):void
<<Java Interface>>
VisualDescriptor
at.lux.imageanalysis
getDistance(VisualDescriptor):float
getStringRepresentation():String
setStringRepresentation(String):void
<<Java Class>>
ColorStructureImplementation
at.lux.imageanalysis
<<Java Class>>
ColorLayoutImpl
at.lux.imageanalysis
<<Java Class>>
EdgeHistogramImplementation
at.lux.imageanalysis

Figure 5.8: Class diagram showing `ImageAnnotation` and associated classes.

5.10 User Structures

The rationale of the 'user structures' information extraction, is that where documents have already been grouped by the user, it makes sense to use this information as it may indicate (or increase the likelihood) of document event commonality). For this initial implementation of *SAESNEG* we focus on the example of photos organised into photo albums – these may or may not indicate event commonality. The parent album of an photo document can be obtained by executing the code shown in listing 5.10 in the Facebook Query Language (FQL)[33] – an alternative to the Facebook Graph API:

```
SELECT aid FROM photo WHERE object_id = 1234567
```

The returned metadata includes an ID for the album, which can then be used to identify documents in the same album, and annotate accordingly. As shown in table 5.16, it was possible to obtain album information for most of the photos (save those where privacy settings forbade it).

Table 5.16: Summary of User Structure Annotations.

Total Documents	583
Total User Structure Annotations	331
Documents with ≥ 1 annotation	322
Mean annotations / document	0.57

5.11 Conclusions

This chapter has described Phase A of *SAESNEG* and documented the development and experimentation related to its constituent components. A framework of techniques has been presented for the extraction of information for the event clustering task with empirical work presented in association with each of the components (table 5.17 gives an overview of the extraction techniques used in the initial system, and ideas for future work). Algorithms and techniques have been selected (drawing on the literature review in chapter 3) for their suitability for the *differentially heterogeneous* source OSN data. Existing techniques and software have been used where possible, and new techniques have also been developed.

This chapter has also described the annotation layer representing the information extracted by the components in Phase A. The annotations are independent of both the kind of source data (i.e. image/text/metadata) and the type of source document (photo/status message/event/etc.) – facilitating the decoupling of Phase A and Phase B, allowing the strategies in Phase B to focus on high-level semantics, and allows their implementation to be less dependent on the source data.The next few sections outline more specific contributions associated with particular components.

[33] https://developers.facebook.com/docs/technical-guides/fql/

	Locations	Temporal	People	Social Events	Low-Level Image Content
Image	(Landmark recog.) (EXIF/GPS)	(Day/Night)	(Facial, Clothing)	(Event Recog.)	MPEG7
Text	Stanford CoreNLP / Nominatim / OSM	Temporal Expressions	Mentions of Friends	Social Event Tagger	N/A
Metadata	Text Lat/-Long	Various Date/Time Fields	Photo Tags, Event Attendees		N/A

Table 5.17: Information Extraction Techniques Implemented in Phase A. Future work in brackets.

5.11.1 Distributed Approach to Temporal Expressions

Section 5.5 presents a novel approach for the representation of temporal expressions with distributed semantics, where the meaning of the temporal information is represented through the means of a probability distribution – rather than a discrete range or a single instant. This work is a contribution of the thesis in its own right, but any useful temporal information extracted from the social media documents is used to facilate the wider event-clustering experiment. It is argued that this alternative paradigm in representing temporal expressions has several clear advantages over the traditional approach of using fixed intervals and ranges:

- It is able to provide definitions for a wider class of temporal expressions, supporting expressions where there is no single official definition (hence widening the temporal expression extraction task).
- It captures greater cultural richness and ambiguity – arguably a more accurate definition.
- It facilitates further processing, such as the consideration of a prior probability (as demonstrated with an example in section 5.5.4).

As a proof of concept, section 5.5.3 also sets out modifications to the state-of-the-art StanfordNLP suite, as well as a method for building a distributed definition from Flickr photo timestamps (based on an annual distribution), and building a compact representation of the expression using a Gaussian mixture model (using the Accord.Net implementation of the MaxEnt algorithm).

Section 5.6 has presented a tagger|parser for well-known social events. Support for birthdays and weddings was implemented as a proof-of-concept. The component supports a strategy for reasoning about document event commonality well-known events – discussed in section 6.6 (note that there the reasoning strategy is not text-specific). Where event-related information is directly available in the source data, it made sense to use it in for the event clustering task. A TokensRegex[34] based grammar was implemented as a plugin for StanfordNLP.

[34] http://nlp.stanford.edu/software/tokensregex.shtml

5.11.2 Named Entities

As discussed in the literature review, two popular approaches to the detection of locations as named entities in text are to use a gazetteer, or to use a grammar. For the locations, this study sought to combine the approaches as a proof of concept, both an extensive gazetteer (based on the OpenStreetMap resource) and Stanford's state-of-the-art probabilistic grammar were used; a simple heuristic was chosen for filtering candidate expressions from both: relying exclusively on the gazetteer matches produced a deluge of false positives, whilst many unambiguous locations were missed by the Stanford grammar. A deeper investigation of integration between both approaches is deferred for future work. For mentions of people, a simple gazetteer – derived from the Facebook friends list was used.

5.11.3 Summary

The SAESNEG framework represents novelty both in terms of combining existing techniques in a new way in order to solve a new problem: event clustering of a differently heterogeneous set of documents (section 1.6), and in terms of new techniques and algorithms developed.

Chapter 6

Phase B: Clustering

> “Non-monotonic logic and probabilistic models will be needed to automate the process of creating the life-story, analogous to the way that one might try to assemble the life of a loved one from artefacts they leave behind: ‘Sometimes simple details can put items back in context. Like the notes scrawled on the back of my grandfathers photos, valuable little pieces of information–time, place, and person–can help create a patchwork of detail that gives some structure to allow the family to reassemble the jigsaw pieces of life.’ ” (Banks, 2011, p. 4)

6.1 Introduction

This chapter concerns Phase B of *SAESNEG* responsible for organising documents into event clusters, based on the annotations extracted in Phase A. The components of Phase B process the annotations generated in Phase A, outputting the set of document event clusters, supporting the event-centric user interface. This user interface displays the user's social media footprint, where documents of different types, and from different social networks, are presented together in groups according to real-world events. *SAESNEG* therefore, (it is argued) overcomes the challenges and deficiencies associated with the current state-of-the-art user experience for the social media footprint, outlined in chapter 3.

Whereas Phase A (described in the previous chapter) processed documents individually, Phase B operates on the set of all documents associated with a particular users life story – comparing annotations from different documents in order to compute the degree of *document event similarity*, and hence event clusters (see section 1.4). As discussed previously, phases A and B are separated by means of an annotation layer, seen as key to overcoming the *differently heterogenous* nature of the source data (section 1.6). The challenge in Phase B is that particular annotations rarely imply event commonality with complete certainty – there is always a possibility that positive evidence associated with one kind of annotation can contradict that of another kind of annotation. The implementation must, ultimately make a binary decision about the event commonality for each pair of documents, given continuous-valued features, which are often sparse and contradictory.

To make the problem of document event clustering tractable, document event commonality is considered pairwise – strategies consider pairs of documents, evidence is combined for each pair (or edge) creating a weighted graph of document event commonality, reducing the document event commonality task into that of correlation clustering problem (section 6.10).

The next section sets out the notion of the strategies, and the other detail of the architecture of Phase B. Each strategy is then discussed in turn, drawing on the relevant sections of technical review in chapter 3, and the extracted features discussed in chapter 5. Remaining sections discuss issues around the evaluation of clustering, present an evaluation of the overall clustering results, and the event-based user interface. Note that these comparison strategies are the reason the annotations are extracted in Phase A; as part of the wider event clustering experiment.

6.2 Architecture and Strategies

This section describes the architecture of Phase B, with a rationale based on achieving maximum utility of the available data and best incorporating existing techniques. The software architecture of Phase B can be divided into three layers: document comparison strategies, an SVM edge classifier, whereby the problem is transformed into an instance of the *correlation clustering problem*, for which a state-of-the-art technique is applied to create the final clustering. This architecture is visualised in figure 4.5, in chapter 4.

The rationale for separating the comparison strategies from the information extraction is to separate the high-level intuition and business logic regarding the comparison of documents from the low-extraction, performed separately in Phase A. This separation is achieved by implementing a layer of annotations between the two phases to represent the extracted information. This arrangement overcomes some of the challenges associated with the dataset – the issue of the *differently heterogeneous* source data (section 1.6). By decoupling extraction from the annotation, the comparison of documents for the purposes of computing document event similarity can be informed by information derived from a range of kinds of data (i.e. image, text, metadata) and types of source document. The strategies in Phase B are based on a mixture of novel intuitions and existing approaches to computing document event similarity, such as those outlined in the motivating example in section 1.2.

These strategies (and the decoupling achieved with the annotation layer) allow ideas and existing work related to the document event commonality task, perhaps developed on some particular source dataset with particular kinds of data, or types of document, and apply this to the generic document event commonality problem studied in this thesis. For example, many existing studies have recognised the key importance of temporal information in the computation of event similarity. In *SAESNEG* a range of extraction techniques can be incorporated into Phase A – temporal information can be derived from text and metadata across a range of types of source documents. Then in Phase B, once the means of extraction and originating context abstracted away, methods of comparing temporal information can be reviewed and implemented. Hence, a range of information extraction techniques can be combined, and the best comparison approaches can be selected applied to that data independently. Such *late fusion*, has been applied to a variety of problems (Bredin and Chollet, 2007; Reddy, 2007; Gunes and Piccardi, 2005).

There is further separation within Phase B: strategies for document event comparison generate features (whose values may indicate positive or negative evidence for document event commonality) which are used for SVM-based classification of the edges between documents. New strategies can be added, and it does not matter if the annotations they create are dependent on other annotations (either by virtue of the underlying interdependencies of the source data, or directly through use of Phase A annotations – reused between different strategies).

After the strategies have processed all pairs of documents, an SVM classifier computes a similarity strength for each respective pair. The SVM classifier is trained using edge labels inferred from the ground truth documents (collected from the specifically developed online user interface). The SVM classifier is operating on the edges of the document graph, so if it were to output a binary edge label, there would be no guarantees that the graph would contain multiple components – it would be necessary to somehow infer the components (i.e. event clusters) from the graph. Instead, the SVM is configured to output probabilities for each of the two output classes (i.e. intra- and inter- event), for each edge.

These probabilities can trivially be converted into (signed) edge weights. At this stage, the document event clustering task (as set out formally in section 1.4) – event clusters must be chosen to maximise intra-cluster edge weights whilst minimising inter-cluster edge weights. This problem is known as *correlation clustering.* A state-of-the-art edge algorithm implementation is chosen to perform this final clustering, which is then evaluated.

6.3 Friends Strategy

Key Idea: People attend events with other people, and documents often contain information indicating the attendance of more than one person. Hence, the set of attendees discovered in documents can indicate or repudiate event commonality between those documents.

6.3.1 Introduction

This strategy uses annotations representing the attendance of people at events associated with documents, to generate positive or negative evidence of document event commonality.

For example, if two documents each respectively indicate the attendence of the same set of people, this is evidence in favour of document event commonality i.e. that the documents relate to the same event. Conversely, if two documents are respectively associated with sets of people, and those sets differ significantly, this is negative evidence for document event similarity, or that the documents relate to different events.

6.3.2 Implementation

As with other strategies, the implementation needs to generate a real-valued feature vector for each pair of processed, annotated documents. For the friends strategy, the annotations of interest are the `PersonAtEvent` annotations, extracted in Phase A.

Hence, the input data for the strategy is two sets of unique identifiers. For the initial implementation of *SAESNEG* the calculation is based simply on the size of the intersection of the two sets of attendees extracted from the two documents, to produce a single feature value in the range $[0, 1]$. Note that the user themselves is always assumed to be in the set of event attendees, hence the lower bound of 0. Listing 6.1 shows part of the implementation. The efficacy of this strategy is discussed in the results section (4.4).

Listing 6.1: Friends Strategy Implementation

```
        List<PersonAnnotation> peopleAtLeft = left.
           getAnnotations().People;
        List<PersonAnnotation> peopleAtRight = left.
           getAnnotations().People;

        HashSet<String> facebookIDsLeft = new HashSet<>();
        HashSet<String> facebookIDsRight = new HashSet<>();

        for (PersonAnnotation pae : peopleAtLeft) {
            facebookIDsLeft.add(pae.getFacebookID());
        }

        for (PersonAnnotation pae : peopleAtRight) {
            facebookIDsRight.add(pae.getFacebookID());
        }

        List<String> intersection = ListUtils.intersection(
           facebookIDsLeft, facebookIDsRight);

        // Ensure that the owner is always in the intersection.
        if (!intersection.contains(_theUserID)) {
            intersection.add(_theUserID);
        }

        return new DatumSimilarityEvidence(
                FestibusFeatures.Friends_InCommon,
                1.0 - 1.0/intersection.size(),
                "");
```

6.4 Locations Strategy

Key Idea: An event occurs at a particular location, and documents often contain information related to those locations. Hence, location-related information discovered in documents can indicate or repudiate event commonality between those documents.

6.4.1 Introduction

Geographic information exists in source data in a variety of forms; in the most trivial cases, metadata fields may indicate precise latitude and longitude co-ordinates. In the case of text the mention of a location needs to be detected and then resolved.

The location aspect is key to characterising events, is part of the Jain and Westerman event model (2007, also figure 2.4). The decision made in this thesis to restrict to so-called *real-world* events means that events can reasonably be characterised as occurring at a single location at a single time[1].

[1] Although the precise temporal and spatial boundaries are left undefined

Level	Description	UK	Distance (miles)
1	Super-national borders		200
2	National border	National Boundary	200
3	Sub-national border		200
4	Sub-national border	England/Scotland/Wales	200
5	National regions	Regions of England: e.g. North West	50
6	Districts, provinces	Unitary Authorities	30
7	Regions, villages, towns, districts		5
8	Municipalities	Districts, London Boroughs	5
9	Suburbs, small settlements		3
10	nneighbourhoods, tiny settlements	Parishes	3
11≤	Neighbourhoods, tiny settlements		1

Table 6.1: Distance Constraints, extended from the OpenStreepMap Wiki.

6.4.2 Implementation

The spatial similarity strategy needs to compute the similarity of a pair of documents, based on the geographic information contained in the associated documents, and output a normalised feature value in $[-1, 1]$.

In Phase A, all geographic information extracted from the source documents are represented as Location Annotations, extracted in Phase A 5.4, and have all been resolved to latitude and longitude co-ordinates. However, a single document may have multiple location annotations, so the *SAESNEG* implementation returns a feature-value based on the pairwise closest locations (i.e. the pairwise shortest distance), using Haversine great-circle[2] distances.

Because the location annotation uses a single latitude and longitude co-ordinate to represent the location, a downside of this approach is that entities with a large geographic area are considered to be geographically proximate, or identical. Suppose Wales, Cardiff and Llandaff are represented as single geographic points:

$$\mathrm{D}_{\text{haversine}}(\text{Wales}, \text{Wales}) = \mathrm{D}_{\text{haversine}}(\text{Cardiff}, \text{Cardiff}) = \mathrm{D}_{\text{haversine}}(\text{Llandaff}, \text{Llandaff}) = 0$$

In the context of the document event commonality measure, this is inappropriate – two documents which happen mention the same country could relate to two entirely different events, whereas two documents mentioning a city region more likely to represent the same event.

To avoid this issue, a simple solution is proposed whereby the distance is constrained by a lower bound based on the size of the geographic entities being compared. For example, a country has a lower bound of 50 miles, towns 5 miles. Fortunately, the OpenStreetMap search results from which most of the location annotations originate, include an `administrative_level` field, indicating the relative size/importance of the associated geographical entity. These levels are shown in table 6.1.

The value of the features is simply the reciprocal of the distance in miles (with one as a lower bound on the distance). Using this approach, a single feature value is returned indicating the geographical similarity (i.e. proximity) of the pairs of documents.The contribution of this strategy to the document event commonality task is discussed in the results (section 6.11).

[2] The great-circle between two points on the surface of a sphere is measured along the shortest path on the surface of the sphere.

6.5 Temporal Strategy

Key Idea: An event occurs at a particular time, and documents often contain time-related information. Hence, time-related information discovered in documents can indicate or repudiate event commonality between those documents.

The temporal event aspect is key to detecting events (Reuter and Cimiano, 2012, discussed in section 3.4.5) and is part of the Jain and Westerman event model (2007). Temporal information is both abundant and provides a clear and simple means of calculating similarity. Social network documents always have a timestamp associated with them, so even the most information-sparse documents contain some temporal information to serve as a basis for an event commonality calculation. Crucially, this means there is always some evidence for every pair of documents – every edge in the document graph has some evidence associated with it. Such temporal metadata fields are readily parsed without the difficulties and ambiguities of resolving other classes of named entities (such as locations).

6.5.1 Introduction

The temporal event aspect is known to be crucial, arguably the single most important facet in characterising and detecting events (Reuter and Cimiano, 2012, discussed in section 3.4.5) and is part of the Jain and Westerman event model (2007). Temporal information is both abundant and provides a clear and simple means of calculating similarity:

- Social network documents always have a timestamp associated with them, so even the most information-sparse documents contain some temporal information to serve as a basis for an event commonality calculation. Crucially, this means there is always some evidence for every pair of documents – every edge in the document graph has some evidence associated with it.
- Such temporal metadata fields are readily parsed without the difficulties and ambiguities of resolving other classes of named entities (such as locations). Traditionally, document similarity can be calculated based on the interval between two time points – a simple calculation.

This section builds on previous sections discussion of existing work (in chapter 3) and the work on the extraction of temporal information in Phase A, and the issues involved (chapter 5). We return to the development of the novel distributed approach to representing temporal information, and explores how that paradigm can then be applied in order to compute the similarity of the documents. Both traditional and density-based approaches are implemented for evaluation purposes.

6.5.2 Distributed Approach

This thesis has argued that there are disadvantages to the traditional 'discrete interval' model for representing temporal expressions (section 5.5.1), and has proposed a novel alternative *density-based* paradigm, with various advantages (section 5.5.5), to recap in summary:

1. Allowing a wider range of text expressions to be considered temporal expressions, by removing the requirement that it must be possible to represent the temporal semantics with a traditional fixed interval.
2. Better capturing the 'real' meaning of the expression, with a statistical representation of its usage.
3. The mathematics of the underlying probability distributions provides a consistent means of combining temporal information, able to accommodate temporal information represented in both traditional discrete intervals and probability distributions derived from other sources
4. Similarly, probability theory provides a mathematically rigorous means of computing the similarity of temporal information associated with two documents.

Section 5.5 has demonstrated how (1) and (2) were applied to the implementation of Phase A. In this section we explore how (3) and (4) can be applied to the implementation of Phase B.

Regarding implementation, as with the other strategies we wish to compute the similarity of a pair of documents, and express that similarity as a normalised feature value, in this case, based on temporal information.

When documents are processed during Phase A, multiple annotations can be added – temporal annotations are no different, and each document has a set of temporal annotations associated with it. Helpfully, there will always be at least one temporal annotation for each document, based on the metadata indicating when the content was uploaded.

Hence, the first task is to convert a set of temporal annotations into a single representation of the probably time the real-world event occurred – demonstrating advantage (3).

Suppose we consider each item of temporal information associated with the document, to be an 'observation' of the underlying actual date and time when the event occurred. We have a random variable T, the time of the event, and for each annotation, we have different probability distributions for T_i, each assumed to be observed independently.

$T =$ random variable, with PDF (probability density function):

$$P_T(a \leq t \leq b) = \int F_T(t)\, dt$$

We have T_0, .., T_N representing each item of temporal information associated with the document, and for each of those we have an associated (known) probability density function:

$$P_{T_i}(a \leq t \leq b) = \int F_{T_i}\, dt$$

The usual approach is to combine the input PDFs by multiplying them, (and re-normalising so that the density function integrates to unity over the pre-defined date range):

$$F_T(t) = \prod_{i=0}^{i=N} F_{T_i}(t)$$

In this way, information from each of the temporal annotations is combined into a final probability distribution, capturing all of the temporal information available in the source document. Note that temporal annotations may have originated from metadata, text or (in future work) image content.

Multiplying the PDFs in this way assumes that all temporal information found in the document actually relates to the time the event occurred, i.e. that all temporal information is *definitive* (see discussion and examples in chapter 5). However, as discussed not all temporal information is definitive, and including the PDFs of such non-definitive information in the product would disrupt the final representation. Hence, non-definitive temporal information is included in the product after being mixed with a uniform distribution over the pre-defined date range. Essentially, assuming that there is a 50% probability of the non-definitive information actually relating to the underlying real-world event.

Revising the equations accordingly, F_T (the combined PDF for the document) can be expressed as:

$$F_T(t) = \prod_{i=0}^{i=N} F'_{T_i}(t)$$

where:

$$F'_{T_i}(t) = \begin{cases} F_{T_i}(t) & \text{(for \textit{definitive} information)} \\ 0.5 + (0.5)F_{T_i}(t) & \text{(for \textit{non-definitive} information)} \end{cases}$$

In this way, a final representation of the temporal information can be calculated. In the implementation, a vector-based representation is used for simplicity, to model the underlying PDFs.

Given a single temporal representation for each document, the second task is to compare the information between documents and compute a measure of similarity, pairwise, which can be used as feature for the SVM-based clustering process. As discussed earlier, an advantage of the PDF-based approach is that it motivates a mathematically sound approach to compare temporal information associated with documents.

Say if we assume that an event occurs on a single day. The PDFs representing the aggregatied temporal information from extracted in Phase A allows the probability of an event occuring on a particular day to be estimated from the temporal information associated with a particular document d, as follows:

$$\begin{aligned} F_T &= \text{combined PDF for document } d. \\ \tau &= [t_{\text{start}}, t_{\text{end}}] = \text{a 24-hour day.} \\ P(E \text{ occurred during } \tau) &= \int_{t_{start}}^{t_{end}} F_T \, \mathrm{d}t \end{aligned}$$

Given the two documents d_1 and d_2 under comparison, with associated events E_1 and E_2.

If we further that, as a general rule, an event occurs on a single day, the days associated with these events can be written as 24-hour intervals:

$$\begin{aligned} \tau_1 &= [t_{start_1}, t_{end_1}] \\ \tau_2 &= [t_{start_2}, t_{end_2}] \end{aligned}$$

For the sake of simplicity, if we assume further that a user is only involved in one event in any given 24-hour day, we can write:

$$\begin{aligned} P(E_1 = E_2) &= P(\tau_1 = \tau_2) \\ &= \sum_{\tau_i \in R} P(\tau_1 = \tau_i) P(\tau_2 = \tau_i) \\ &= \sum_{\tau_i \in R} F_{T_1}(t_{start_i}) F_{t_2}(t_{start_i}) \end{aligned}$$

where:

$$\begin{aligned} R &= \text{The overall date range under consideration.} \\ \tau_i &= \text{each 24-hour period in } R \end{aligned}$$

In summary, if we take the dot product of the vectors representing the PDFs associated with the documents d_1 and d_2, the result is equal to the probability that they relate to the same event, ignoring any other evidence. This result, in the interval $[0, 1]$ can be directly used as value for a machine-learning feature for the pairwise (i.e. edgewise) clustering algorithms which follow.

Temporal annotations were also extracted from the text (see section 5.5). When computing the similarity between a document pair, a simple and robust similarity metric is employed: the difference in days between the events. Where one or both documents contain multiple annotations, an annotation is selected from each document to minimize the number of days between documents. Annotations representing the interval of an entire year are ignored.

The function used to compute the value of the similarity metric is as follows (see section 1.4).

6.5.3 Traditional Approach

In additional to the novel approach, temporal annotations were also extracted from the text (see section 5.5), adopting the usual fixed-interval representation (section 5.5). When computing the similarity between a ument pair, a simple and robust similarity metric is employed: the difference in days between the events. Where one or both documents contain multiple annotations, an annotation is selected from each document to minimize the number of days between documents. Annotations representing the interval of an entire year are ignored.

The function used to compute the value of the similarity metric is as follows (see section 1.4 for definitions):

$$\begin{aligned} D_i, D_j &\in F(u) \\ \text{Similarity}(D_i, D_j) &= \frac{1}{1 + (\text{min days between } D_i \text{ and } D_j)} \end{aligned}$$

6.6 Social Events Strategy

Key Idea: Some social events have a specific purpose, or type, such as a wedding or birthday. There is not usually more than one such *event kind* associated with an event. Hence, any information discovered in documents which indicates the event kind, can, if compared, indicate or repudiate document event commonality. Furthermore, extra details, who was getting married – or the age of the birthday can strengthen this evidence either way.

6.6.1 Introduction

The events strategy implements a means of document event not commonality not seen in previous work, based on information relating to specific social events discovered in source documents during Phase A.

As with all the other strategies in Phase B, the task is clear – to compute a measure of document event commonality for a given pair of documents, the strategy is then used to compare all pairs of documents, and resulting feature values are combined with those of other strategies to inform the event clustering task.

If there is event-related information in contained in documents, it makes sense to use it for the purposes of computing document event commonality. Not all documents will contain such information, so the coverage is likely to be poor, but in the cases where a pair of documents both have a social event annotation, comparison could lead to a high degree of certainty about the document event commonality task.

The social event annotations extracted in Phase A are discussed in section 5.6, and detailed information about the extraction of the events (as well as details of the annotations representing the extracted information, future work, and so forth) can be found in that section. To recap, the initial implementation extracts information related to birthday and wedding events. Well-known events have individual semantics, important for parsing and when computing similarity: annotations may include details of the detected social events, such as the age and/or name of the person with the birthday, and the names of the people getting married. The rationale for extracting this detailed information is that the details can increase the degree of certainty of the document event commonality judgement, as discussed below.

6.6.2 Implementation

Given a pair of documents, the strategy needs to compute the value for a feature describing the similarity of the documents in consideration of their social event annotations.

Each input document is associated with zero or more input social event annotations. As is the case with the other strategies, the initial task is to merge the input documents, and make some initial checks:

- If either document has no such annotations, no calculation is performed, and a default feature value of zero is used.
- If input annotations associated with either of the documents are a mixture of different kinds of events (i.e. a single document has both birthday and wedding annotations), no calculation is performed, and a default feature value of zero is used.

- If the social event annotations for the documents are different (i.e. one document is a birthday and the other is a wedding), a feature value of -1 is returned to represent this strongly negative evidence for document event similarity.

Otherwise, details are merged for each respective document so that metadata is combined. Then, available details can be compared between the two 'flattened' annotations, and a feature value can be generated indicating positive or negative evidence for event similarity. For the initial implementation of *SAESNEG* this comparison is based on the following details from the annotations:

- Name(s) of people getting married.
- Age of the *birthday person.*
- Names of the *birthday person.*

For the sake of brevity, the implementation details regarding annotation merging, handling of null/missing information, the means of comparison, and effect on output feature value are omitted here. The resultant feature value in the range $[-1, +1]$ is then passed to the next step: machine-learning based document event clustering.

6.7 Scene (Image Content) Strategy

6.7.1 Introduction

Key Idea: Photographs of the same event scene might be visually similar, hence visual similarity between photographs may mean that those photographs are more likely to relate to the same event. In this way, image content can inform judgements regarding document event commonality.

Photos are a highly popular type of social media document, and feature heavily in the sample dataset used in this study (75% of documents, table 4.4). There are a range of existing techniques for extracting both high and low level semantic information from photographs, which can then be compared with image and non-image derived information from other documents. This topic has been discussed extensively in the earlier literature review (section 3.3), and section of chapter 5 (section 6.7), and such discussion is not repeated here.

For the initial implementation of *SAESNEG* only low-level image features were used. In this section, the implementation of the Phase B image-processing strategy is discussed, the counterpart to the techniques in Phase A. To recap, during Phase A, a number of low-level image features are extracted from each document, following the MPEG-7 Standard (The Moving Picture Experts Group, 2010). In Phase B, these annotations are compared to compute values representing the degree of visual similarity between the images, to be used as features for the machine-learning based clustering, the next step in Phase B. This section merely describes the implementation of image content processing functionality in Phase B.

6.7.2 Implementation

The implementation of the `PhotoSceneStrategy` is reasonably straightforward. As with all the other strategies, the document comparison is pairwise: given a pair of documents (and their associated annotations) the strategy must generate a bag of features representing the positive or negative evidence for document event commonality for that pair.

The annotations extracted in Phase A hold a selection of image descriptors, extracted using the Caliph-Emir implementation of the MPEG-7 standard. These descriptors include: *scalable color*, *color layout*, and *edge histogram*. Unlike other strategies, no merging of annotations is required, and the complexities of handling inconsistent and partial information are obviated. Helpfully, the MPEG-7 standard also describes how each of these descriptors can be compared to compute image similarity, algorithms which are again implemented in Caliph-Emir. Hence, the implementation of the scene strategy simply needs to invoke the appropriate comparison mechanisms from the implementation.

For each of the descriptors compared, a separate feature is included in the output, and used in the subsequent machine-learning based clustering algorithm. High image similarity is interpreted as positive evidence event similarity, and values are rescaled to $[0, 1]$, ready for machine-learning. Where one or both documents does not have an associated image (and hence no image content annotations) no features are returned, and default values of zero are used.

6.8 User Structures Strategy

Key Idea: Users often organise social media documents into structures (the key example being photo albums), often (but not always) grouping according to real-world events. Hence, when attempting to organise source data into real-world events, existing user structures are likely to provide useful information.

This section describes the Phase B *user structures* strategy for comparing documents according to their existing groupings in their source social networks, and follows from the discussion in chapter 5, and corresponding implementation in Phase A (section 5.10).

6.8.1 Existing Work

The non-conclusive inclusion of photo album information into a system which calculates document event commonality (for social networking data) is a novel approach to an existing problem. However, use of photo album information is not new – other studies have for example used photo albums as a source of ground truth (Rabbath et al., 2012), but employ assumptions about the relationship between events and photo albums. The approach in this study is to investigate the relationship more objectively, and consider photo album commonality as a feature, which, depending on the independently-gathered ground truth, might be a useful clue in creating document event clusters. A key distinction between the work of Rabbath et al. (2010) and *SAESNEG* is their limitation to photos; because

SAESNEG operates on all types of documents, the complete social media footprint, using photo albums (or the equivalent for any other document type) as the basis for ground truth is not feasible.

6.8.2 Implementation

In this implementation of *SAESNEG* Facebook photo album information is used to compute feature value indicating where photo documents are associated with the same Facebook album. There is no support for other document types in this implementation, in these cases a default feature value of 0 is returned.

Photo album information is extracted in Phase A, and stored as annotations. In the Phase B *user structures* strategy, the implementation is trivial. If photo album information is available for both documents, then a feature value of +1 or -1 is returned. An implementation detail is the handling of what (for the purposes of this thesis) are termed *structural albums* – albums generated by Facebook to serve particular purposes, for example:

- Profile Pictures: used to hold all the user's previous personal profile images.
- Mobile Uploads: holds images uploaded from the Facebook mobile app.

Because these albums are 'built in' and serve specific purposes, they are excluded from the strategy. Note that these albums facilitate special interpretation for other purposes – photos in the mobile uploads folder are interpreted as *live* documents.

6.9 Type Strategy

6.9.1 Introduction

Key Idea: users create a variety of different types of social networking documents as they interact with social networks in a different of ways. When comparing pairs of social network documents, the types of the documents themselves may provide useful information regarding document event commonality.

The type of document may indicate some degree of information about the likelihood of document event commonality between a pair of documents. Many photos may be taken at a single event, hence, given two photos from the same user, there is some chance they may relate to the same event; however, the likelihood of two event documents being created relating to the same event is vanishingly small. Facebook document events are created to represent events, whereas photos, status messages and other types of documents are created for different reasons as users interact with social networks in different ways.

Hence, this strategy generates features indicating the combination of types for each document pair under comparison, in the hope that this information is useful for the machine-learning based clustering, the next step in Phase B.

6.9.2 Implementation

Again, as with all the other strategies, documents are processed pairwise. Implementation of the type strategy is trivial, the type of each of the documents is detected, and a feature value is chosen according to the pair of types – there is one feature for each pairwise combination of types. The appropriate feature is set to a value of +1, whilst the others remain at their default value of 0. In this way, the type strategy generates what is effectively a prior probability for the document event commonality, before any of the other information is considered. This technique is the recommended handling of such a *so-called* categorical feature, as described in the LIBSVM manual (Hsu et al., 2010, p. 3). The effect on the overall clustering performance is discussed in section 6.11.

6.10 Clustering

6.10.1 Introduction

This section discusses the event clustering in *SAESNEG*; various approaches to clustering are reviewed from the literature, appropriate techniques are identified and discussed, and an algorithm is proposed whereby machine-learning can be combined with a selected clustering algorithm to produce the final document event clusters – this section's main contribution to the thesis. The implementation of this algorithm, and integration with external libraries are described in overview.

6.10.2 Clustering: A Literature Review

Clustering, or more formally, *cluster analysis*, is the broad term to describe the task of partitioning a set into sub-sets such that elements in together same subset are somehow more similar to each other, and elements in different subsets are somehow less similar or dissimilar to one another.

Clearly, this is an extremely broad area of computer science, with a large number of techniques, and a very large number of published algorithms in the literature. Furthermore, algorithms which are able to classify elements of a set according to some notion of a class could, technically, be considered clustering algorithms[3] This would include many of those employing machine-learning e.g. SVM, naive Bayes and Neural Networks.

The intention here is not to present a comprehensive review of all approaches, techniques, and algorithms used for clustering in general. Instead, there is a brief discussion of the main approaches which might seem relevant to our task, and decisions made on a suitable algorithm.

To recap, the overall goal of this thesis is to partition the set of source documents from the users social media footprint, into sets (or clusters) representing real-world events; see section 1.4 for the formal problem description. Where examples of each cluster are available beforehand, the clustering problem can be viewed as a classification problem – and machine-learning classifiers could be used to identify events. In this case, there are a large number of events, and there is no way of knowing beforehand any information about them, so such algorithms can be immediately excluded as a direct means of performing the clustering – as we might do in a task such as sentiment classification (Blamey et al., 2012). However, in scenarios where the events *were* known in advance (and training data might be obtained) such as classifying Tweets according to which football match they related to, this approach might be useful.

Many popular clustering techniques (which could broadly be called *geometric clustering techniques*) rely on the source data consisting of points in a metric space (most commonly, a Euclidean space). This includes the centroid and k-means clustering techniques, the *distribution* based clustering techniques, and *density based clustering* (e.g. DBSCAN, Ester et al., 1996, among others). Such approaches are simple, robust easy to visualise, readily tractable, and are very popular. Of the four key properties of a metric space[4][5], the triangle inequality is of key interest:

$$d(x, z) \leq d(x, y) + d(y, z)$$

The Phase B strategies to compute measures of similarity do not, generally, respect the triangle inequality. Suppose that documents x and y were rated as highly similar according to some strategy (perhaps they both mention a birthday), and that documents y and z were rated as highly similar according to some other strategy using some other information (perhaps nearby locations) – supposing there is incomplete information to suggest that x and y, and y and z, respectively, are dissimilar. However, suppose that when x and z are compared, yet another strategy finds strong evidence for dis-similarity (i.e. negative evidence for document event commonality). Hence, $d(x, z)$ would be large and positive (representing low similarity), contradicting the lower bound of the triangle inequality.

[3] If the class label indicates the cluster.

[4] Non-negativity, Identity of Indiscernibles, Symmetry, and the Triangle Inequality.

[5] Or rather its associated metric, or measure.

The above is not intended as a mathematical proof, but a general argument that the geometric approach is perhaps not the most naturally suited to this task: with the sparsity of available features, the independence of the various Phase B strategies, and their source annotation data intrinsically not resembling metric spaces (and the strategies not resembling metrics) this does not seem the most appropriate approach.

Indeed, such algorithms are not used for social network analysis[6]; the detection of clusters, or communities of users. A range of algorithms have been developed especially for this community detection problem. However, our task is not that of community detection – our network consists of documents, not users – the graph structures are highly likely to be different. There is nothing to suggest that a community of users (with social ties) resembles the set of documents in an event (with similarity calculated by the *SAESNEG* strategies). Furthermore, many of these algorithms do not support edge weights, and those that do may not support negative edge weights – Phase B strategies may decide that there is strong evidence for two documents not relating to the same event, there is no analogy for this in community detection. Community detection algorithms (Girvan and Newman, 2001; Blondel et al., 2008; Clauset et al., 2004; Raghavan et al., 2007; Newman, 2006) instead generally search for sub-graphs with a high density of edges, or sub-graphs which are cliques, on graphs which often have very large numbers of vertices.

Instead, more general clustering algorithms could be used. Supposing the evidence from the strategies were somehow combined into a single, signed, edge weight – the task is to find a set of clusters where intra-event edges generally had positive weight, and inter-event edges had generally negative weight. This is discrete optimisation task is known as *correlation clustering*. A state-of-the-art correlation clustering algorithm is selected for this task (discussed in section 6.10.4) – while the next section focuses on the calculation of edge weights.

6.10.3 Computation of Edge Weights

Before the subsequent clustering step, edges need to have a weight associated with them. Therefore, it is necessary to convert the bag of features generated by the Phase B strategies associated with each edge; into a single edge weight representing the overall evidence for document event commonality along the particular edge.

This is accomplished by training an SVM classifier to recognise intra- and inter- event edges, using the ground truth event clusters gathered from users with the web interface described in chapter 4 (section 4.10). SVMs are ideally suited to the task, because they are fundamentally 2-class classifiers, and there is no assumption of independence between input features (unlike naive Bayes). *SAESNEG* utilises the LIBSVM implementation (Chang and Lin, 2011). The library is invoked as a separate process, with the input files being generated by *SAESNEG* and the results parsed. 10 fold cross-validation is used to train, test and evaluate the model. The model is trained to generate probability estimates, which are in turn used to calculate the edge weights, as shown in equation 6.1. With the classifier trained to output class propbabilities, these are multiplied by the ***complementary*** edge class frequencies – so that overall, the expected total edge weight is zero, this balances intra- and inter- edge weights.

[6] Whether in the context of OSNs or not.

$$\begin{aligned} W_e = {} & P(e \text{ is intra-event}) \cdot f_{\text{inter}} \\ & - P(e \text{ is inter-event}) \cdot f_{\text{intra}} \end{aligned} \tag{6.1}$$

Hence, the output from the Phase B strategies (which make use of the information extracted in Phase A) is transformed into a discrete graph problem.

6.10.4 Computation of Clusters

Having computed the edge weights using the support vector machine in the previous step, the graph is ready to be partitioned into clusters which should represent real-world events, so that the documents can be presented in the user interface.

The task is that of *correlation clustering*, with the state-of-the-art C++ implementation by Elsner and Schudy (2009) chosen. The input matrix files were auto-generated (using the edge weights obtained in the previous step) and the output clusters were then fed back into *SAESNEG* for additional processing. The authors' recommended configuration of chained solvers was used: `log` → `vote` → `boem`[7]. The results are then parsed by *SAESNEG* and the final event clustering is stored for display in the user interface.

6.10.5 Summary

This section has presented a brief review of the main techniques and approaches for clustering, and then explained how a suitable approach was chosen for use in this study. SVMs have briefly been discussed, their advantages, suitability for this study, and their use for the calculation of edge weights – thus translating the document event clustering task into that of a correlation clustering problem. State-of-the-art implementations of support vector machine classifiers and correlation clustering were selected. The next section discusses the performance evaluation, and presents results.

6.11 Results

6.11.1 Introduction

This section presents results of Phase B, including the overall performance of *SAESNEG* The experiment consisted of running the pipeline on the data for each user, for all those users who participated by contributing ground truth data. Each time the experiment was run on the users, two forms of evaluation were performed:

1. The calculation of edge-wise classification accuracies, across all users, utilising 10-fold cross-validation across the aggregated edge-set, the approach of the *rand index.*
2. The calculation of mutual-information based metrics for evaluating 'goodness' of the output clusters, in comparison to the ground truth clusters. A mean was simply calculated over the users.

[7] *'Basic one element move'.*

Label	Description
None	No strategies at all – a baseline.
Friends	The ‘Friends’ strategy only (section 6.3).
Events	The ‘Events’ strategy only (section 6.6).
Spatial	The ‘Spatial’ strategy only (section 6.4).
User Help	The ‘User Help’ strategy only (section 6.8).
Scene	The ‘Scene’ strategy only (section 6.7).
Kind	The ‘Type’ strategy only (section 6.9).
Temporal (dist.)	The ‘Temporal’ strategy only (section 6.5), configured to use the distributed approach to modelling temporal information.
Temporal (trad.)	The ‘Temporal’ strategy only, configured to use the classic approach to modelling temporal information.
All	All strategies, with the temporal strategy configured to use the traditional approach.

Table 6.2: Strategy Selections used for *SAESNEG* experiments.

This experiment, including the evaluation outlined above, could then be repeated under various configurations, such as turning particular components or sections of the pipeline on or off, tweaking pre-processing settings, etc. to allow performance comparison. For this section, the comparison is based on Phase B strategy selection: the (multi-user) experiment was repeated for various selections of the strategies for Phase B. The Phase B strategies process the annotations extracted in Phase A, generating the machine-learning features used for clustering in Phase B. By selecting individual strategies (and hence the corresponding machine-learning features), the contribution of each to the overall performance can be evaluated independently. The set of strategy/feature selections were used as the basis of the results are shown in table 6.2, these labels are used for the presentation of subsequent results.

Hence, performance of the system is measured in two ways (edgewise classification, and overall clustering), for various selections and configurations of the strategies. Before presenting results, the mutual-information based metric for evaluating cluster performance needs to be explained.

6.11.2 Normalised Mutual Information

The evaluation of clustering performance is non-trivial. This section is broadly a synopsis of the ‘evaluation of clustering’ section of the excellent book by Manning et al. (2008), applied to the social network document clustering task.

The participants in the study have contributed ground truth event clusters – gold truth partitionings of their social media footprint (or at least some small sample of it). The clusters generated by *SAESNEG* can be evaluated against this ground truth data. In the case of a simple classification task, this evaluation is straightforward, especially in binary classification: the confusion matrix can be calculated without difficulty, from which metrics such as *precision*, *recall*, *f-measure*, can be calculated.

In a clustering context, a metric analogous to recall is *purity*:

> “To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N.” (Manning et al., 2008).

The purity suffers from a analogous issue to that of using recall by itself as a classification metric — placing each document in its own cluster yields a purity of 1. An alternative is to include a precision-like measure, and compute the edge classification accuracy, known as the *rand index* – results using this edge-wise approach are discussed below. The usual challenge of fairly evaluating performance in the case of skewed classes remains; the F-measure can be used.

An alternative is to adopt an information-theoretic approach, and measure the mutual information between the ground-truth partitioning and the generated partitioning. One such measure is the normalised mutual information (Manning et al., 2008):

$$\mathrm{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega, \mathbb{C})}{(H(\Omega) + H(\mathbb{C}))/2}$$

Where Ω is the set of output clusters, $\mathbb{C}$ is the set of ground-truth clusters, H is the entropy, and I is the mutual information.

The metric expresses the mutual information between the generated and reference clusterings, as a proportion of the mean of both the entropy in the reference clustering, and the entropy in the generated clustering. This yields a result in the range [0,1], hence normalised to be independent of the underlying entropy. A complete recovery of the information in the reference clustering yields a score of 1, whilst the generation of a 'random' clustering (with respect to the reference clustering) yields a score of 0. Note that the various NMI results show the mean NMI across users. The ONMI utility[8] (McDaid et al., 2011) was used for measuring NMI.

6.11.3 Discussion

This section presents and discusses the results. Results are presented in various forms:

- Figure 6.1 presents the inter-edge classification accuracies according to strategy selection.
- Figure 6.2 presents the (mean) NMI according to strategy selection.
- Table 6.3 shows the all results illustrated in the three charts in numerical form.

The experiment gave a positive result. *SAESNEG* 'works' in the sense when all the strategies were included (and the classic approach to temporal information was used) the system is able to cluster the source documents sufficiently well to achieve a normalised mutual information of 0.66. To the author's knowledge, this is the first demonstration of social media document clustering given a mixture of social media document types. The experiment has shown that it is possible to successfully re-create event clusters in a general sample of the social media footprint.

Nothing beats the temporal aspect: the use of the temporal strategy (using a classical approach) alone is sufficient to match the performance of all the strategies combined. This highlights the importance of the temporal event aspect, and demonstrates that the finding of earlier studies (on photos) is also applicable to a general sample of social media documents. The temporal strategy out-performed any other individual strategy.

[8] https://github.com/aaronmcdaid/Overlapping-NMI

Some of the other strategies, in isolation, were also able to out-perform the baseline: each of the 'friends', 'spatial', 'user help' and 'scene' strategies were able improve clustering performance by themselves. However, the results show that each of those strategies (and the associated information extraction techniques in Phase A), were able to boost the clustering performance above the baseline, showing that information useful for the document event commonality judgement was successfully extracted from the source documents, and that the associated strategies were able to improve the document event clustering performance in those cases. Clearly the temporal aspect is key, and without temporal information, performance is severely limited, but the finding suggests that with further work, non-temporal information may be able to offer an improvement in document event clustering. Although in this study, with the data available, and the experimental setup described, they did not improve the performance when combined with the information from the temporal aspect.

After the 'temporal' strategy, the next most informative strategy was the 'user help' strategy. Showing that photo album information can be improve judgements about document event commonality. Clearly, it will not always be useful, and is unsuitable for non-photo documents – as such, it is unsuitable for use in the generation of ground truth for this task, as it is in other studies which study photos exclusively (Rabbath et al., 2010; Rabbath et al., 2011). It demonstrates the importance of not overlooking metadata; using a simple readily-available machine-readable ID in the source data greatly outperformed an NLP pipeline and gazetteer for the extraction of spatial information for example.

The 'type' strategy, in isolation, offered no discernible performance improvement. This is not surprising, simply knowing the types of the documents is unsufficient insufficient to make any judgement about document event commonality. The (key) temporal aspect has a nuanced relationship with the document type – photos usually relate to the moment, whilst status messages could be forward looking statements, statements 'looking back', or statements 'in the moment'. Despite the negative result for this strategy, the author is keen to investigate how type information can be considered in decisions of document event commonality in the future. Using an alternative classifier for the dominant photo-photo edge cases would be a simple approach.

Disappointingly, the novel 'event' strategy (using information about events such as birthdays and weddings), in isolation, failed to outperform the baseline. Although judging commonality on the basis of social event type alone is unhelpful (i.e. (birthday,birthday) $\not\Longrightarrow$ the same event), when distinguishing information is available, one would expect the information to be helpful. Future work will investigate the issue in more detail.

Finally, adopting the novel distributed approach to representing temporal information (suggested in chapter 5), failed to improve performance of the system in comparison to the classic approach (although performance was reasonable at 0.62). A major (claimed) advantage of the distributed approach is the extraction of a culturally nuanced representation of the temporal information – however, in this study, with students recruited from the UK, there is perhaps less rationale for this approach – that all temporal expressions could be interpreted accoridng to UK/USA/'Western' cultural norms is actually a fair assumption for this dataset. This does not undermine the rationale for the distributed approach: the mined representations of the temporal expressions in OSN photos from around the world clearly demonstrate the rationale for a more nuanced representation. Cultural and geographic biases in ascribing 'Western'/Northern hemisphere, meaning to expressions such as 'Summer' and 'New Year' are clearly incorrect, something the data in chapter 5 (appendix A) shows. The issue has been overlooked, and the approach outlined in this thesis is reasonable.

Photos formed a large part of the source dataset, and a PDF function was mined (section 5.5.2) and used to estimate event times for photos. However, a key assumption was the independence of the event time for photos, even if they happen to have been uploaded at the same time, and/or uploaded to the same album. This is plainly not true, album structures do contain information about event commonality (and hence about commonality of event time) – somehow this information needs to be incorporated into the probabilistic models used for comparing temporal information associated with photos. This challenge is left for future work. By contrast, under the classic approach to handling temporal information, the upload time was used, so photos uploaded at a similar time would be interpreted as having a high degree of document event commonality.

In summary, *SAESNEG* is able to successfully generate document event clusters from source social networking data, containing a variety of document types – the first system to do so. The isolation of individual strategies has demonstrated the key importance of the temporal aspect, and how useful it can be for clustering a mixture of different types of social networking documents. Positive results from other strategies has also demonstrated the importance of other event aspects.

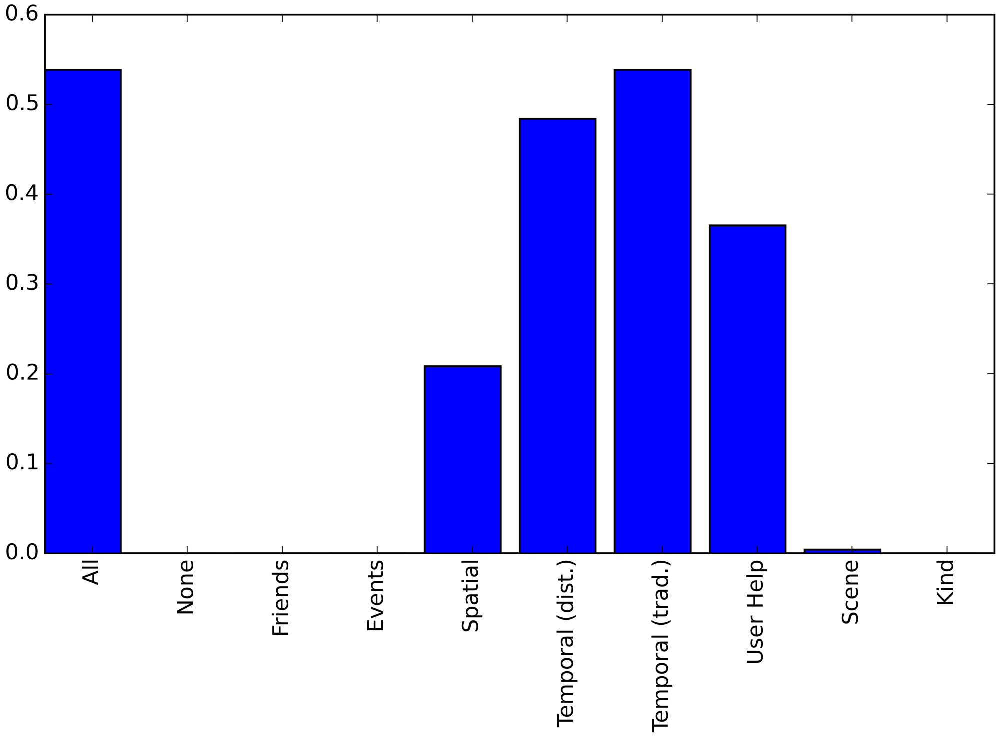


Figure 6.1: Intra-Event Edge Classification Accuracies by Strategy Selection

6.12 Summary

This chapter has described the implementation of Phase B of *SAESNEG* responsible for clustering the social network documents. Phase B processes the annotations representing information in Phase A using a set of strategies – each of which compare pairs of documents (or *edges*), for document event commonality – based on the information contained in the source annotations. These strategies generate a set of features for each edge, which are used as the basis for training an SVM classifier. The resulting classification of each edge is used to generate an edge weight, and the document event clustering task is hence transformed into that of *correlation clustering.*

This chapter has set out the architecture of Phase B, the rationale and implementation of each of the strategies, the ML edge classifier, use of external utilities for clustering (chained solver) and clustering evaluation (normalised mutual information).

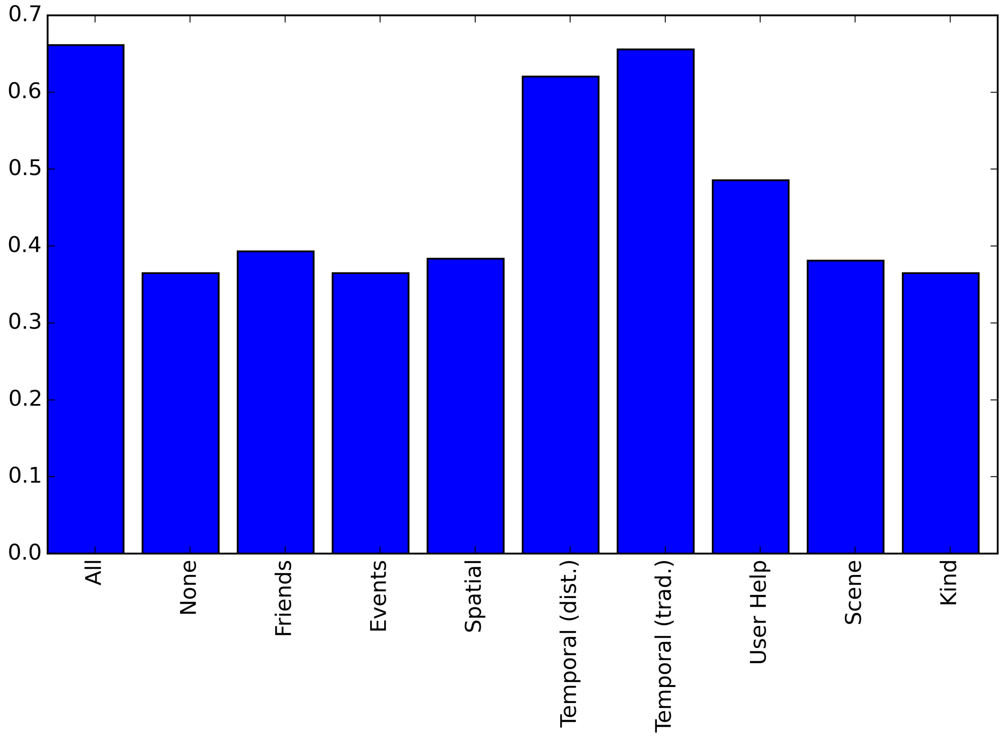


Figure 6.2: Mean Normalised Maximum Entropy by Strategy Selection

Table 6.3: Performance According to Strategy Selection

Strategy Set	Accuracy (Intra-Event Edges)	Accuracy (Inter-Event Edges)	Mean NMI
All	0.54	0.99	0.66
None	0.00	1.00	0.36
Friends	0.00	1.00	0.39
Events	0.00	1.00	0.36
Spatial	0.21	0.99	0.38
Temporal (dist.)	0.48	0.99	0.62
Temporal (trad.)	0.54	0.99	0.66
User Help	0.37	0.99	0.49
Scene	0.00	1.00	0.38
Kind	0.00	1.00	0.36

An evaluation of the performance of *SAESNEG* was presented in section 6.11. To summarise the results, it has been shown that *SAESNEG* is able to successfully generate document event clusters, on a differently heterogeneous sample of source social network data (section 1.6). The key importance of the temporal aspect has been confirmed on this novel dataset. Results for other strategies and event facets were encouraging, but failed to improve overall performance beyond that of the temporal facet/strategy.

Chapter 7

Conclusions

7.1 Introduction

This chapter identifies and clarifies the main contributions of the thesis. This introduction presents a brief overview of the study, highlights some core themes and concepts which re-occur throughout the thesis, and introduces some of the main areas of contribution.

This thesis has presented *SAESNEG*: a *System for the Automated Extraction of Social Network Event Groups*, combining a range of techniques to create a complete social-media processing pipeline, capable of performing event clustering on the entire personal social media footprint.

The intended application of the system is to download an archive of a user's personal social media footprint and organise the documents (or *documents*) contained within it according to real-life events, so that the archive can be presented in that manner. The idea is that a single real-world event, such as a holiday, may be associated with a variety of social media documents, perhaps distributed across different social networks, created before, during and subsequent to the event taking place. When this information is presented as an document event cluster, users should be able to navigate their personal timeline more easily, creating a user experience conducive to reminiscence, something that is lacking in today's online social networks.

SAESNEG achieves this by employing a wide range of information-extraction techniques from source social media documents and then applies a set of comparison strategies to compare the extracted information and ultimately make judgements about *document event commonality* i.e. whether documents are indeed associated with the same real-world event. Much of the thesis relates to the selection of these algorithms and techniques, the development of improvements and new, novel approaches – and how all the techniques and algorithms (and the data which they extracted) were combined together within the larger system.

To develop, train and evaluate *SAESNEG* social media data was collected from study participants, who were then asked to assemble ground truth (or 'gold') reference event clusters. The resultant data could be analysed (the results of which constitute findings in their own right), and then used for performance analysis of the system, presented towards the end of the previous chapter.

To recap, there are recurring topics or themes, which form the basis of the thesis:

1. The rationale: reminiscence, and the need for a 'reminiscence-friendly' user experience in the context of online social networks and the personal social media footprint, with associated issues such as the topic of lifelogging.
2. The event clustering task: to facilitate reminiscence, the collection of social media documents is organised into clusters representing real-world events (see section 1.4). Doing this entails making judgements on basis of comparing a range of extracted information from the source data.
3. This extraction and comparison is complicated by the nature of the source data – most crucially, its *differently heterogeneous* nature (section 1.6), and the sparsity of the events: few documents, private in nature with little external reference (section 1.7).
4. Overcoming the challenge of the *differently heterogenous* nature of the source data is motivation for *SAESNEG* and its architecture, selection and development of algorithms, and is what distinguishes *SAESNEG* from existing work – the first system (to the author's knowledge) capable of performing event clustering on a mixture of different types of social media document.

Event clustering is not a new research task, *SAESNEG* builds on a range of existing work to achieve event clustering–a variety of systems have sought to organise personal documents in this manner, supported by a range of techniques. The areas of contribution include:

Socio-Technical contributions: Identifying the need for a system to create an archive of the personal social media footprint, and facilitate a reminiscence-friendly experience of that dataset; *SAESNEG* is presented as an effort towards such a solution. Some of the findings relate to topics such as lifelogging, privacy, ownership, identity, digital death and legacy. Within data-mining, cultural issues are highly pertinent.

Proposed Novel Algorithms and Techniques: Various new algorithms, approaches and techniques have been proposed. Some, such as the combined clustering technique (section 6.10), and the overall *SAESNEG* processing pipeline, its unique architecture, supporting components; are more relevant to this specific research task. However, others, such as the distributed approach to representing temporal expressions and the social event detection grammar have implications for wider research tasks, and should find application outside this thesis.

Empirical Results: Including the analysis (and characterisation) of ground truth event clusters, the creation of distributed definitions of temporal expressions. The thesis culminates in a performance evaluation – testing the end-to-end performance of the system from raw data through to computed clusters, measured against ground truth clusters created by the participants of the study, using their own data.

Future Work: it is hoped that the study highlights and motivates a range of issues for investigation in future work, both in terms of individual algorithms and techniques, and regarding *SAESNEG* itself. Ideas for future work are presented in section 7.6.

7.2 Contributions

The key contributions of the thesis are listed below, organised into four groups, forming the structure for the discussion. The groups of contributions are organised generally in the order in which they appear in the thesis, but do not map precisely to chapters.

Rationale, Context, Existing Techniques:

1. The motivation to provide a user interface (and indeed user experience) of the personal social media footprint – the rationale for *SAESNEG* (and the experiment) as identified from the literature.

2. A literature review of a range of data-mining and event clustering techniques, some overview of historical trends, recent developments, and identifying state of the art approaches, techniques, algorithms, abd libraries – allowing a novel combination of range of techniques for text and image analysis to be combined with document event clutering strategies according to the *SAESNEG* archictecture. Approaches to handling temporal information in information extraction are critiqued.

***SAESNEG*, Ground Truth Documents and the Characterisation of the Social Media Footprint**:

3. The Collection and Analysis of Ground Truth Clusters.

4. The characterisation of the data found in the personal social media footprint, and the ground truth events – differently heterogenous data with small, private events unlike large 'public' events – with various implications for the selection of techniques and design of experiments (sections 1.6 and 1.7).

5. The arcitecture of *SAESNEG* – with the separation of the concerns of *extraction* (Phase A); with seperate pipelines for image, text and metadata respectively – from *clustering* (Phase B) by means of a layer of abstract annotations. The document event clustering task is also formalised.

Novel Techniques:

6. The Distributed Paradigm for temporal expressions – entirely novel, mathematical rationale, overcoming cultural bias in data-mining; along with other advantages, methodology for the creation of distributed temporal definitions, and application to *SAESNEG*.

7. The technique for social event mining, using the StandfordNLP `TokensRegex` language to describe rules for the discovery and extraction of social events (such as birthdays and weddings) from text, with application to the document event clustering task.

8. The approach to the event clustering itself: pairwise application of SVMs for edge classification, to induce a signed edge weighting, coupled with the existing chained solvers implementation (Section 6.10.4) – adapting the document event clustering task into that of correlation clustering.

Results and Performance Evaluation:

9. The empirical evaluation of strategies and their comtribution to the overall performance of *SAESNEG* – an experiment in supervised machine learning.

10. The performance of *SAESNEG* on the document event clustering task: **It works!**

The remaining sections of this chapter expand on the contributions claimed above, and justify them by reference to evidence contained in the thesis, finishing with discussion of future work, and concluding remarks.

7.3 Context, Rationale, and Existing Techniques

In this section, the purpose and theoretical contributions of the two literature review chapters (2 and 3) are summarised. Broadly, chapter 2 reviews literature relating to the wider context, and establishes a rationale for the study – showing that the problem is worth investigating. Subsequently, chapter 3 reviews a range of existing work relevent to implementing the system – showing that it would be possible to create *SAESNEG* by combining a range of existing techniques, algortihms and libraries in a novel way. The main contributions of each chapter are summarised below.

Chapter 2 discusses the wider context and rationale of the study: the value we place on personal documents, the practice of reminiscence, and essentially seeks to problematise the issue of reminiscence of the context of social media. A number of central topics are introduced (section 2.2): reminiscence, and lifelogging – the thesis seeks to explore the various meanings of the term, placing lifelogging at the core of the thesis, whilst touching on issues digital death and digital legacy.

Building on these foundational definitions, the growing challenges of acheiving reminiscence and preservation in today's ever-changing social media lanndscape are discussed as a series of interrelated trends, with increasing:

- volume of data being captured (section 2.3.1);
- number of social networks and devices we use, (section 2.3.2);
- variety of documents created (section 2.3.3).

The thesis argues that these trends have created a number of distinct challenges:

- Handling the volume (section 2.3.5): how can such quantities of data be displayed to the user effectively?
- Fragmentation: the technical difficulties of collecting and archiving data from diverse networks in diverse formats (section 2.3.6).
- Finally, the risk of data loss: abandonement as users transit among services, service closure, or mere accident (section 2.3.7).

This is the rationale for the experiment: is the novel combination of various data mining techniques, coupled with machine learning and correlation clustering able to compute the event clusters automatically, and hence facilitate a event-centric, reminiscence-friendly user experience, with the necessary structure mined automatically from the users social media footprint?

A number of systems are discussed which are partial solutions to these challenges, or seek to make such a claim. These are grouped, and critically reviewed from a socio-technical viewport. The event-based system (in a broad sense) is introduced, and a range of commercial and academic systems of this kind are discussed, from both commercial and academic realms. The history of organising personal data (including the personal social media footprint) according to an event structure, in both academic studies and commercial systems is presented as a rationale for the choice to pursue an event-based system for the representation of the social media footprint for this study – the event is the de facto structure for organising and presenting personal data, and it was deemed sensible to pursue this model for *SAESNEG*

In light of this decision, given the volume of social media data under inspection, the research task is clear: to create a system that is able to automatically organise the personal social media footprint into event groups, with the primary intended application of a user interface conducive to reminiscence-centred experience of the social media footprint.

Whilst the non-technical content of chapter 2 may appear to be incongruous with the remainder of the thesis, the establishment of a clear rationale is a key socio-technical contribution, and is justified with an extensive literature review.

The second literature review (chapter 3) reviews a range of existing data-mining algrotihms and techniques, and some similar existing systems (mostly event-clustering systems supporting a single type of document, and using only a single kind of data). The review covers data-mining and event clustering techniques, some overview of historical trends, recent developments, and identifying state of the art approaches, techniques, algorithms, and libraries.

The function of *SAESNEG* is to perform clustering of *documents* (social media documents) from the mixture of document types found in the social media footprint. Hence *SAESNEG* builds on a history of event clustering systems, but is differentiated by its ability to process and compare a variety of documents: different types of document, and different kinds of source data (the so-called *differently heterogenous* social media footprint), unlike many existing systems which handle a single type of document, and a single kind of data within it.

Hence, *SAESNEG* needs to employ a variety of information extraction techniques, for text; image and metadata, extracting information likely to be useful for the purposes of computing document event commonality, according to the strategies described in chapter 6. Within the field of NLP, the literature review focuses on techniques and approaches to named entity extraction and recognition for names of places, people, and temporal expressions. This creates scope to use *SAESNEG* as a basis for developing new and improved information extaction techniques, as discussed chapter 5, for example: the *distributed approach to temporal expressions.*

Low-level NLP techniques are also explored, such as tokenisation and part-of-speech recognition, with discussion of the difficulties presented by the unnatural language of social media, and recent research efforts to overcome these issues.

Whilst NLP was a key focus, *SAESNEG* extracts and compares infomation from image content (where it is present in documents); for this initial implementation of *SAESNEG* only low-level extraction techniques are incorporated, accordingly, such techniques are reviewed, and appropriate algorithms selected. See future work for a discussion of high-level image feature extraction.

Finally, a range of existing event-centric systems are critically reviewed, in terms of the technical details of how they achieve document event clustering (in contrast to the socio-technical/user experience perspective in chapter 2). There are a number of approaches of “event detection”, and the algorithms used to, for example, detect global media events (such as football World Cup matches) on Twitter (section 3.4.6), or partitioning GPS time-series data from a military patrol, are quite different to computing which documents from an individuals social media footprint relate to the same birthday party. Where suitable, strategies for computing event commonality are identified from existing studies, the techniques of Sandhaus, Boll, Rabbath et al. (2010; 2012; 2013) being a key influence.

After this extensive review, there are a range of techniques and algorithms for both *information extraction* and *event clustering*, to be incorporated into *SAESNEG*

7.4 *SAESNEG* Ground Truth Document Clusters and the Characterisation of the Social Media Footprint

The contributions listed in this area relate to social media data itself, its colection, analysis, defining characteristics and the associated implications for the design of *SAESNEG*. The document event clustering task is formalised in section 1.4.

A ground truth collection tool (and supporting components) has been proposed (contribution 3) – a novel drag-and-drop user interface, presented as a web page. The interface allows users to organise their social media data into events, forming the ground dataset. By requiring users to create ground truth documents 'from scratch' – rather than being pre-populated by heuristic as in some other studies, it is argued that the resulting ground truth document clusters are a more realistic sample of real-life events; rather than the cases which happened to fit a particular heuristic. Clearly, users have their own memories of the associated events: in some cases there may simply be insufficient information available in the source documents for accurate partitioning into event clusters – a theoretical upper bound on the performance of a system such as *SAESNEG* this is one avenue to explore in future work, although analysis by an expert user requires access to the private and highly personal information. The implementation and details of the ground truth web interface are discussed in section 4.10, with similar discussions for other supporting components of the system: the user database (section 4.7),the fetcher daemon (section 4.8.3) and subsequent serialisation (section 4.9) of the collected social networking data.

Analysis of the collected ground truth data (contribution 4) has shown that the number of participants (who actually created events) were in line with similar studies (table 4.1), and that event clusters are small – typically only a few documents (table 4.3). Furthermore, a range of types of documents were selected by users for inclusion in event clusters, meaning that document event clusters often contained a mixture of different event types, and that various types of documents should be considered to be part of the social media footprint, and processed as part of the event clustering.

Hence, this thesis argues that the social media footprint is *differently heterogenous* (see section 1.6 for a definition). Furthermore that event groups are shown to be small, often relating to private events, in contrast to larger public events, for which the related social media is often used as the basis for studies in event clustering. Focusing on the personal, private social media footprint, these two characteristics have influenced the design of the architecture and algorithms comprising *SAESNEG*.

These characteristics have greatly influenced the design of the *SAESNEG* architecture (contribution 5) – the separation into an extraction Phase A and clustering Phase B, by means of annotations independent of the source data. Clearly, successful clustering of the documents has required comparison of information from these various sources. The clear separation of information extraction and information comparison creates an opportunity to develop and evaluate algorithms for each, independently.

By representing extracted information independently of the source data kind and document type, such concerns are abstracted away from the document comparison strategies in Phase B, with the intention of allowing grestest use of the extracted data – a key priority given the small size of the event clusters, and the sparsity of available source data. An important difference between *SAESNEG* and exsiting systems is the handling of different types of documents found in the social media footprint.

In summary, this thesis has proposed a means of collecting ground truth document clusters from users. Analysis of the resulting data has identified two key properties of the data (a) that document event clusters are small – often relating to private events and (b) that social networking data is differently heterogeneous – with events containing various types of documents (photos, status messages, etc.) and kinds of source data – text, image, metadata.

7.5 Results and Performance Evaluation

The performance of *SAESNEG* was evaluated in section 6.11. Results were presented to evaluate the system's performance in two distinct ways: both for per-edge classification accuracy, and a mutual-information based metric for overall clustering performance. This evluation is the evaluation of the overarching experiment: can the SAESNEG (with its novel combination of data-mining techniques, event commonality strategies, machine learning and correlation clustering) organize the documents in a user's social media footprint into events (this is the task described formally in section 1.4).

The related findings were as follows:

- *SAESNEG* was able to successfully perform the document event clustering task, on the sample dataset collected for this study. To the author's knowledge, this is the first time that the event clustering task has been automatically performed on a sample of the social media footprint containing multiple types of documents.
- The key importance of the temporal aspect, a finding in other studies, is confirmed for this dataset.
- Some of the other strategies, in isolation, were able to out-perform the baseline. This indicates that the corresponding information extraction techniques (in Phase A) contained information likely to be useful for the document event clustering task. With further development, the other strategies should be able to boost the performance above that achieved with the temporal strategy alone even although this was not the case under this study.
- The proposed distributed approach to representing temporal information was not able to achieve a performance boost; this outcome is discussed in section 6.11.3.

7.6 Future Work

There are many obvious avenues for future work on *SAESNEG* perhaps incorporating new techniques to take advantage of the wealth of information available in the source data – to enhance the accuracy of the event clustering task. It is hoped that the novel techniques developed in this study can find wider application, and be further developed in the future. More widely, the study has brought existing state-of-the-art techniques (and entirely new ones) from NLP into the domain of event clustering, which is commonly associated with multimedia and more image-centric research. A range of ideas for future work are outlined below:

7.6.1 SAESNEG

1. Socio-Technical Issues: explore in greater depth, working with entirely different repositories of personal information: emails, personal finance information, and so forth – and reconcile *SAESNEG* with emerging paradigms of ownership and privacy.

2. New Applications: explore applications outside of reminiscence, utilising *SAESNEG*'s event clustering functionality as a means of detecting fraudulent financial transactions, applications in cyber security, or developing new event-based privacy models which straddle multiple social networks.

3. Event Summarisation: can names for events be auto-generated based on a combination of heuristics and annotations (who/what/where/when)? Clustering means that not all documents in an event need to be shown in the timeline – how can documents be chosen to represent the event?

4. Performance: Improve *SAESNEG* by investigating alternative machine-learning approaches, and investigating why some of the strategies failed to improve overall clustering performance.

7.6.2 Temporal Information

Temporal information is key to the event clustering task. Future work could explore the distributed approach to representing temporal expressions in more depth, pursuing the associated techniques in their own right, within the field of NLP and temporal expression parsing, widening the scope of temporal expressions.

It is hoped that the work could be extended by modelling semantics at alternative scales, considering a wider range of expressions, including durations – which means expanding support for SUTime's temporal operators, alternative asymmetric distributions could be included in the mixture model, with an appropriate algorithm to determine initial parameters, to achieve a better fit to some of the distributions that were found. Furthermore, the intention is to develop a framework for evaluating the distributional approach against the existing approaches, and explore the philosophical issues, such as the dividing instant problem, in the context of the distributed approach to temporal expressions (section 5.5.5).

7.6.3 Friends

Some of the assumptions made about event attendance are an oversimplification: a future study could investigate whether status messages did indeed indicate physical attendance of the owner (and tagged users) at events, based on a variety of factors, perhaps using machine-learning. Such associations could be represented probabilistically, instead of as a concrete assertion (something similar was achieved with the temporal expressions in section 5.5).

There are wider issues here: "virtual" event participation – the football World Cup for example: are event participants...playing on the pitch? watching from the stalls? from a bar? or the highlights on TV? Work from other disciplines (outside the scope of this thesis) is potentially relevant: can future work reconcile the event model (Westermann and Jain, 2007) with the geographies of *supermodernity* (Augé, 2008).

7.6.4 Social Event Parser

Future work could greatly expand the social events supported by *SAESNEG* the details extracted, and improve their means of comparison – perhaps supporting 'known nicknames' of individuals. A wider scope of events could be considered, and the specific semantics modelled. If a system was able to reliably detect a handful of key life events, the extracted data could serve as a skeleton around which to structure the a life story, an approach shown to be effective in other studies (Alahmari, 2012), and one currently used by the Facebook Timeline (although the latter is based on explicit metadata).

7.6.5 User Structures

Photo album metadata provides a rich vein for future research into document event commonality, especially given the high popularity of photos in social media – and the way they are used to showcase event-related social networking data. Future work will examine the relationship between photo albums and real-world events in greater depth; for example: what proportion of albums do indeed represent real world events? Can this be inferred from album titles and descriptions? How can textual information from albums be incorporated into document similarity calculations, and should all photos in an album 'inherit' information associated with their parent album?

7.6.6 Image Content

SAESNEG system architecture is intended to be extensible, and has been designed to allow future support for other image extraction techniques (including those developed in existing work without this specific application in mind), with the decoupling of phases A and B through the use of the layer of annotations.

For example, an additional image-content based processor could be added to the collection of Phase A annotators, and generate location annotations based on landmarks recognised in photos (as in the work of Abbasi et al., 2009). The location annotations generated by this component would be aggregated with the other locations annotations extracted from text and metadata sources, for consideration by the location document event commonality strategy in Phase B – geographic information in the form of annotations, originating from text, metadata and image sources can then be compared regardless of their source context (thus overcoming the challenge of handling *differently heterogeneous* source data, a key challenge of this study).

Existing work on facial recognition (Suh and Bederson, 2007) could be incorporated in a similar way in contexts where such information was not already available as metadata – generating 'people' annotations for consideration by the appropriate component in Phase B. Whilst the recognition of location-related objects in photos (or indeed other objects) would doubtless be useful in some instances, such techniques work well where large datasets are available, but whether any useful information could be obtained in the context of sparse, private events for the document event commonality task would be an open research question and could be investigated in future work.

Authors have noted the limits of what can be extracted from digital photographs, alone, noting that "identifying an event, such as a birthday party or its location would be particularly hard." (Suh and Bederson, 2007, p. 5) – whether such high-level features would even be useful in the event-commonality task, in the context of the (perhaps sparse) personal social media footprint is an open research question.

7.7 Closing Discussion

This thesis has presented *SAESNEG* a *System for the Automated Extraction of Social Networking Event Groups* – explained why it was built, how it was built, and demonstrating that it was able to cluster social media documents into event clusters; as set out in the formal problem statement at the start of this thesis (section 1.4).

SAESNEG is an experimental pipeline comprising a novel framework of data mining techniques to extract annotations, strategies to reason about event commonality based on those annotations, using machine learning (with labels derived from user-created ground truth), and correlation clustering to convert the output of the machine learning into the final set of computed event clusters. It was used to conduct the overarching experiment in this thesis: can such a system be trained to automatically partition a user's social media footprint into datum event clusters?

SAESNEG is able to aggregate and cluster the variety of documents found in a social media footprint into events. The rationale for doing so was the facilitation of a integrated, reminiscence-friendly user experience of social networking data, something thought to be lacking in the user interfaces of today's online social networks.

SAESNEG's two-phase Architecture was developed to accommodate a variety of information extraction techniques to process the variety of image, text and metadata, with corresponding strategies to compute document event commonality, found in Phase B. *SAESNEG* incorporates a wide range of existing techniques, as well as more novel approaches such as the social event parser (section 5.6).

SAESNEG is a response to the research challenges of information extraction and clustering in the context of heterogeneous data, as identified by Bontcheva and Rout (2014) (as per quote on page 80), and the wider socio-technical challenges of preserving a lifetimes' worth of social media content, as explored by Banks (2011).

Whilst the previous section has set out a variety of avenues for future work, an overarching research goal for the future is to migrate *SAESNEG* to the personal private cloud (Wu et al., 2009; Ardissono et al., 2009; Tian et al., 2011), giving users an archive of their social media footprint; fully under their control, organised into a lifelogging style event-timeline for ease of navigation. As the OSN ecosystem begins to mature (n.b. Facebook is now ten years old), archival is a sensible logical step; ongoing privacy concerns will only increase the motivation. The realisation of this vision rests on a number of compelling research questions:

- What are the challenges facing the personal cloud?
- How do we architect the personal cloud, to overcome these challenges? How do we interact with other cloud services and paradigms?
- How do we store data in the personal cloud? How can novel architectures facilitate new business models?
- What are the socio-technical barriers to personal cloud migration – both from user and tech-business standpoints?
- What are the other novel uses of the personal cloud stack?

- Having brought together this great wealth of personal information – how can it be organised and displayed?
- How can we build solutions from existing open-source software to realise the personal cloud vision?
- Greening the personal cloud: how can hardware and software reduce the environmental impact of the personal cloud?

It is hoped that *SAESNEG* constitutes progress towards the goal of empowering individuals to explore and experience their personal social media footprint, and preserve it for posterity. Hence empowering users to achieve the ultimate lifelogging experience in a world of ever expanding online social networks, mixtures of devices and ever-growing variety of documents and information preserved online.

Appendix A

Examples of ***Distributed*** Definitions of Temporal Expressions

This appendix shows example mined *distributed* definitions for temporal expressions (according to the techiques discussed in section 5.5.2.

A range of phrases were chosen for the creation of a distributed definition, some are typical temporal expressions defined in SUTime (e.g. 'Summer'), and others more novel examples seen in text from the social media documents collected in the wider study.

For Bonfire Night (figure A.1) (5th November, United Kingdom) the primary concentration is near the primary date, but with more variance than is with the case with April Fools' Day. A number of other distributions in the fit have between 1-2% mixing coefficient, with means at 8th January (possibly relating to the solemnity of John the Baptist on 16th January), 26th June (Midsummer's Eve, 23rd June is popular for bonfires in Ireland), and 2nd May (Bonfires are popular in Slavic Europe on 1st May).

Christmas (figure A.2) (commonly 25th December) starts early, with 10% of the probability density contributed by a distribution with a mean of 20th November. In cultures using the Julian calendar, Christmas is celebrated on 7th January and 19th January, perhaps explaining some of the probability density seen in January. In the case of New Years' Eve, more than 92% of the probability mass is centred around 31st of December. A normal with mean of January 15th was found, possibly relating to the Chinese New Year on (23rd January, 2012).

The data for Halloween (figure A.3) clearly has a skewed distribution about its official data, 31st October – with much more activity seen in the preceding weeks, with activity rapidly dropping off afterwards. A similar distribution is exhibited in the case of Valentines' Day on (14th February). A limitation of the work is that asymmetric distributions were not included in the mixture model; such distributions are fitted to a cluster of normal distributions with appropriately decaying mixture coefficients.

Freshers' Week is a term used predominantly in the UK to describe undergraduate initiation at university, usually in September or October. With the obvious differences between educational calenders between regions and institutions, a complex pattern is unsurprising. In the case of Last Day of School (figure A.5), assuming a precise date range in the general case is clearly impossible. Many universities have multiple Graduation (figure A.4) ceremonies a year, with loose conventions on dates, reflected in the clustering of the data.

Definitions of seasons show a significant bias toward northern hemisphere definitions, to be

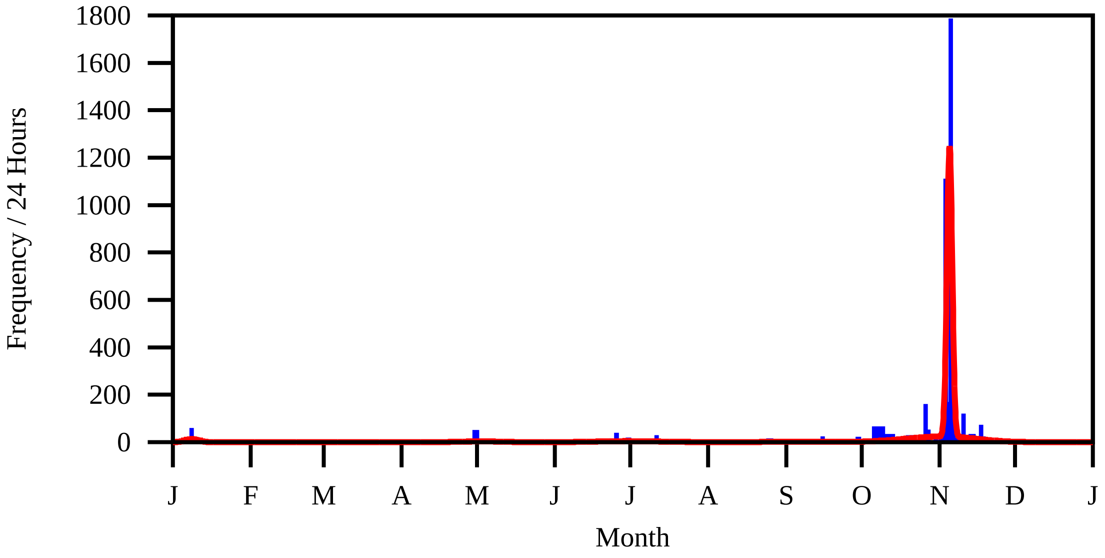


Figure A.1: Distribution of “Bonfire Night”

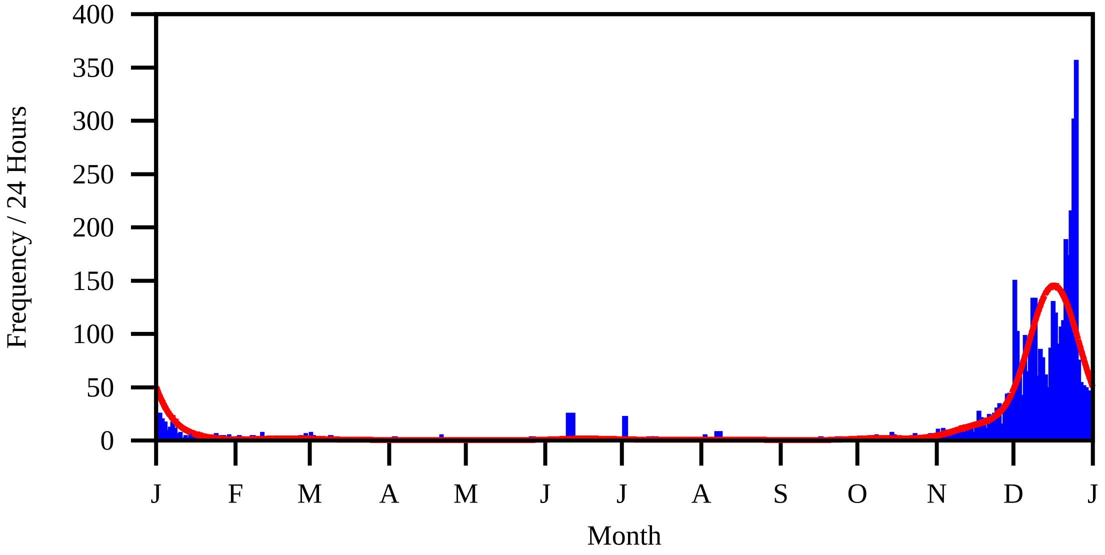


Figure A.2: Distribution of “Christmas”

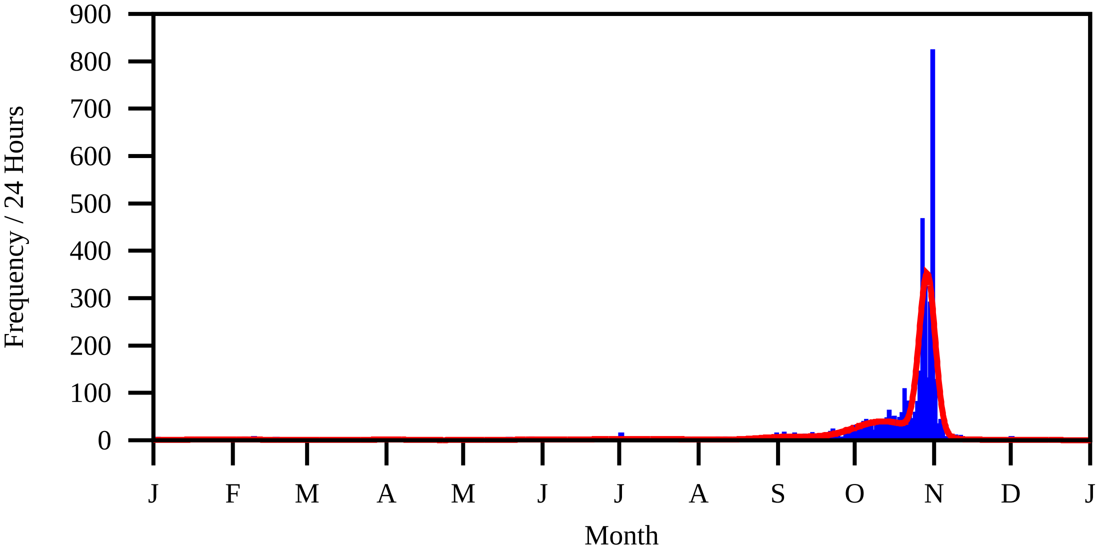


Figure A.3: Distribution of "Halloween"

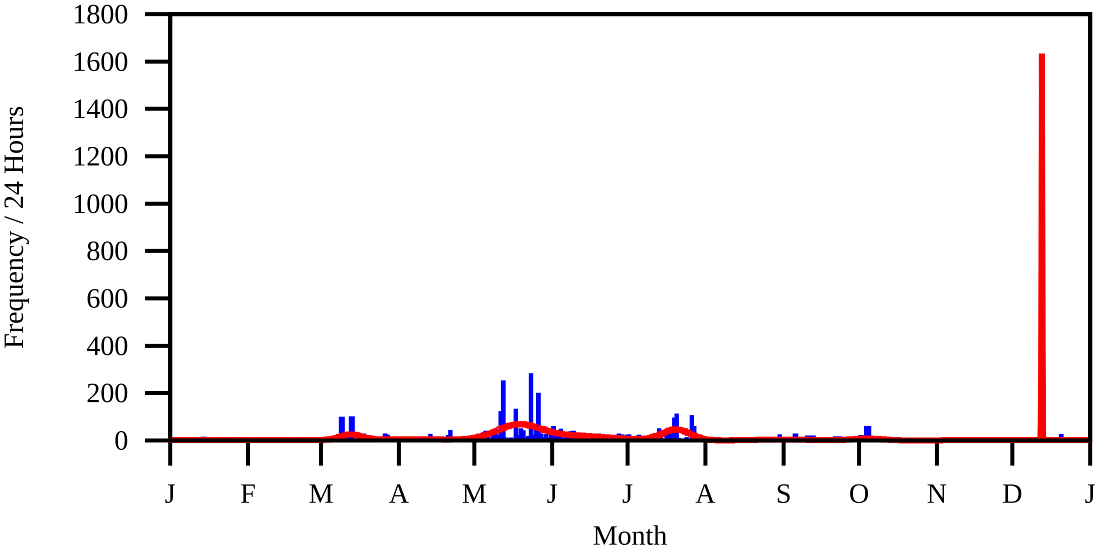


Figure A.4: Distribution of "Graduation"

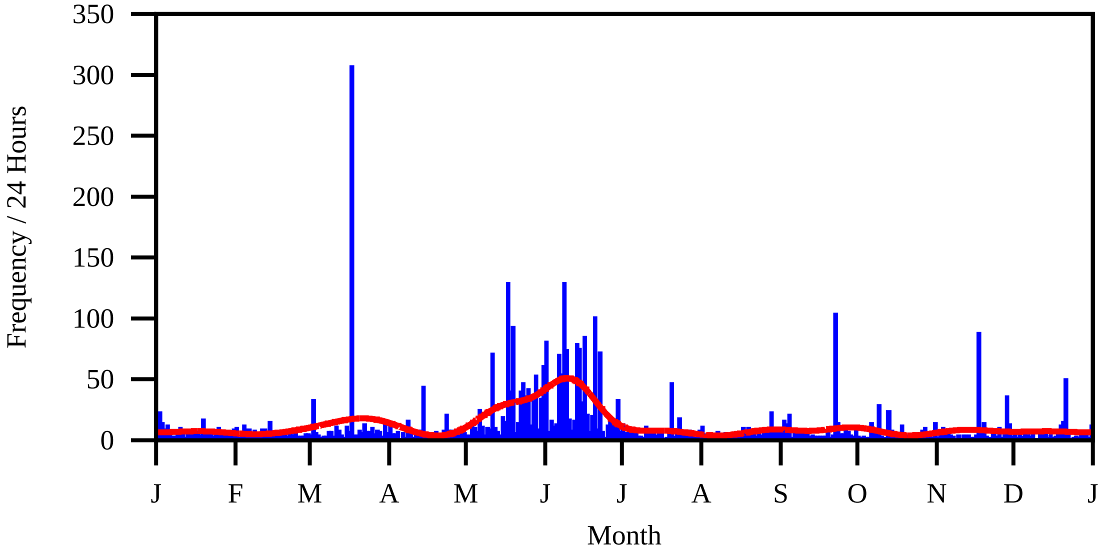


Figure A.5: Distribution of "Last Day of School"

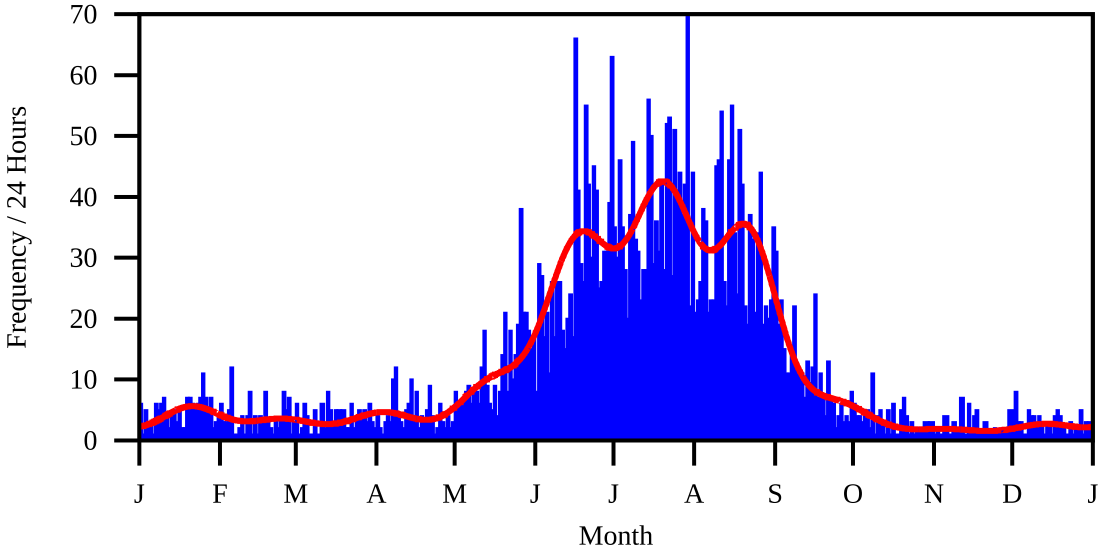


Figure A.6: Distribution of "Summer"

expected with the bias towards English language. However, density is found at the antipodal dates in each case, modelled by normal distributions of appropriate means. For example, Winter has a distribution with a mean of 13th July, with a mix coefficient of 1%, with a similar phenomenon in the other samples. It is clear from the distribution of the data that all season terms in the study are used year-round. The data for Summer is shown in figure A.6, exhibiting a similar antipodal peak.

Some degree of background “noise” was present in many of the examples: the mixing coefficient was typically in the region of 2%.

Appendix B

Anonymized Event Data from Study Participant

In this study, the term *event* is defined as a set of social media documents (see: section 1.4). Due to the large number of event documents (583) a sample from one of the participants is shown here in full. Events do not have any names or other identifiying features – they are simply sets of documents; any clues regarding the event commonality of the documents needs to be determined by the system as part of the experiment.

The proceeding pages present ground truth event data from one of the participants; note the following:

1. The data presented is a summary of the raw data as downloaded from Facebook, the inclusion of the full original JSON documents would have been prohibitively verbose (as indeed would the inclusion of data from all participants).

2. Subject to the provisions of the original research ethics proposal, explicit approval has been granted from the individual concerned for the publication of this data.

3. Various details (such as names) have been anonymised – although the pseudonyms are consistent to allow comparison between documents.

4. A small amount of this individuals data has been excluded because they did not wish the information to be published.

5. Many of the photo documents do not include the associated photo, it not always being available from the Facebook API, often due to the privacy settings of the user(s) involved. Similarly, album information is only sometimes available. This sparsity of information is noted as a challenge of the research (section 1.7).

6. Note the recurring distinction between a *Facebook Event* (a document explicitly representing an event, for handling RSVP, etc.) and an *event cluster* – defined as a set of documents (section 1.4) – documents which themselves may be Facebook events.

Event 0

Event

Date: 2013/04/21
Desc.: The UKs largest roaming vintage fair is headed back to Cardiff's City Hall this April!...
Location: Cardiff
Organizer: UF

Event 1

Event

Date: 2013/02/16
Desc.: Hope you can join me in the Sofa Bar of The Artesian Well for birthday drinks from 8pm, followed by dancing! +...
Location: The Artesian Well, 693 Wandsworth Road, London, SW8 3JF
Organizer: DE

Event

Date: 2013/02/16
Desc.: It would be lovely to see you all before my birthday party for drinks at the flat and a meal at Kaosarn (Thai ...
Location: Kaosarn, 110 St John's Hill, London, SW11 1SJ
Organizer: DE

Event 2

Event

Date: 2013/03/14
Desc.: Cardiff University's Graduate College are pleased to announce that the 6th Annual Spotlight on Social Sciences...
Location: The Graduate Centre - Cardiff University
Organizer: MG

Event 3

Event

Date: 2012/12/16
Desc.: Badges, investitures and a party as well! Help will be needed by parents of all sections.
Location: 49th Scout Hall
Organizer: NC

Event 4

Photo

Album Name: Tampere y más
Date: 2013/05/31
People: SP, TM, KS, FT, UP, NO, HP,

Photo

Album Name: Tampere y más
Date: 2013/05/31
People: FT, TM, NO, SP, HP, KS,

Photo

Album Name: Tampere y más
Date: 2013/05/31
People: TM, FT, KS, NO, HP, SP,

Photo

Album Loc.: Tampere, Finland
Album Name: Climbing down the Observation Tower
Date: 2013/05/31
People: SP, FT,

Photo

Album Loc.: Tampere, Finland
Album Name: Climbing down the Observation Tower
Date: 2013/05/31
People: FT, UP, SP,

Photo

Album Loc.: Tampere, Finland
Album Name: Climbing down the Observation Tower
Date: 2013/05/31
People: FT, EP, SP,

Photo

Album Loc.: Tampere, Finland
Album Name: Climbing down the Observation Tower
Date: 2013/05/31
People: FT, SP, EP,

Photo

Album Loc.: Tampere, Finland
Album Name: Climbing down the Observation Tower
Date: 2013/05/31
People: FT, UP, SP,

Status Message

Date: 2013/04/04
Status: Southern Finland in June- anybody got any tips?

Status Message

Date: 2013/06/09
Status: Almost hit an elk in a Smart car....My money was on the elk.

Status Message

Date: 2013/06/13
Status: Was in Tampere when Bon Jovi were playing, they were in Cardiff last night... Am I unwittingly turning into a ...

Photo

Album Name: Profile Pictures
Date: 2013/05/27

Photo

Album Name: Profile Pictures
Date: 2013/06/19

Event 5

Photo

Album Name: Jen and Tim's Wedding
Date: 2013/04/15
People: UM, SM, SP,

Photo

Album Name: Jen and Tim's Wedding
Date: 2013/04/15
People: CB, UM, SM, GM, BG, TH, SP, SA,

Photo

Album Desc.: Birthday and Wedding
Album Name: London '13
Date: 2013/04/07
People: UM, GM, TH, SP,
Photo Name: Wedding Reception, Windsor

Photo

Album Desc.: Birthday and Wedding
Album Name: London '13
Date: 2013/04/07
People: UM, SM, SP, SA,
Photo Name: Wedding ceremony of Jen and Tim

Photo

Album Desc.: Birthday and Wedding
Album Name: London '13
Date: 2013/04/07
People: UM, SP,
Photo Name: Wedding ceremony of Jen and Tim

Photo

Album Name: Jen and Tim's Wedding
Date: 2013/04/08
People: TH, UM, GM, SP,

Photo

Album Desc.: Birthday and Wedding
Album Name: London '13
Date: 2013/04/07
People: UM, GM, TH, SP,
Photo Name: Wedding Reception, Windsor

Photo

Album Name: Jen and Tim's Wedding
Date: 2013/04/08
People: UM, GM, TH, SP,

Event 6

Photo

Album Name: Heatwave!!! And other fun:-)
Date: 2013/09/08
People: ML, SP, TY,

Event 7

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, UM, SM, SP,
Photo Name: The end of a fab weekend in Leeds!

Photo

Album Name: Winter '11
Date: 2011/12/04
People: SM, SP,
Photo Name: Night in Leeds!

Photo

Album Name: Winter '11
Date: 2011/12/04
People: UM, SP,
Photo Name: We made it!!

Photo

Album Name: Winter '11
Date: 2011/12/04
People: UM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, SM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, UM, SM, SP,
Photo Name: Half way up!

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, UM, SM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, UM, SM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: SM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, UM, SM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, UM, SM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: IE, UM, SM, SP,

Photo

Album Name: Winter '11
Date: 2011/12/04
People: SM, SP,
Photo Name: Leeds Dec '11

Event 8

Photo

Album Name: Spring- Summer '12
Date: 2012/06/23
People: DS, SP, SA,

Photo

Album Name: Spring- Summer '12
Date: 2012/06/23
People: DS, SP, SA,

Event 9

Photo

Album Desc.: People, Alcohol, Bowling, Raving Rabbits, Dr Shoe and Alan Sugar
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Tom and Hannah's - 2011
Date: 2011/10/23
People: CB, UM, SP,

Photo

Album Desc.: People, Alcohol, Bowling, Raving Rabbits, Dr Shoe and Alan Sugar
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Tom and Hannah's - 2011
Date: 2011/10/23
People: SP,

Photo

Album Desc.: People, Alcohol, Bowling, Raving Rabbits, Dr Shoe and Alan Sugar
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Tom and Hannah's - 2011
Date: 2011/10/23
People: UM, SP,

Photo

Album Name: Autumnal Times of 2011
Date: 2011/10/23
People: SP,

Photo

Album Name: Autumnal Times of 2011
Date: 2011/10/23
People: UM, SP,

Event 10

Photo

Album Name: Mobile Uploads
Date: 2013/01/18

Photo

Album Name: Mobile Uploads
Date: 2013/01/18

Photo

Album Name: Mobile Uploads
Date: 2013/01/18

Photo

Album Name: Mobile Uploads
Date: 2013/01/18

Event 11

Photo

Date: 2013/10/05

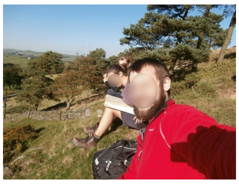
Photo

Date: 2013/10/05

Photo

Date: 2013/10/05

Photo



Date: 2013/10/05

Photo

Date: 2013/10/05

Photo

Date: 2013/10/05

Event 12

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: SP,

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: SP,

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: CB, SP,

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: CB, SP,

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: SP, SB,

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: DB, CB, SB, SP,

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: CB, SP,

Photo

Album Name: Cardiff- Ben's Birthday
Date: 2011/02/21
People: CB, SP,

Event 13

Photo

Album Desc.: These are some pictures of one of our training sesions in Llanedeyrn High school that we have every Friday eve...
Album Loc.: Llanedeyrn
Album Name: CRG Training
Date: 2013/05/15
People: SP, TB, DH,

Photo

Album Desc.: These are some pictures of one of our training sesions in Llanedeyrn High school that we have every Friday eve...
Album Loc.: Llanedeyrn
Album Name: CRG Training
Date: 2013/05/15
People: KB, DH, JH, SP,

Photo

Album Desc.: These are some pictures of one of our training sesions in Llanedeyrn High school that we have every Friday eve...
Album Loc.: Llanedeyrn
Album Name: CRG Training
Date: 2013/05/14
People: SP, SL, BM, TB, NZ, KB,

Event 14

Photo

Album Name: Dad and Evies bday bash
Date: 2011/03/29
People: SP, CB,

Photo

Album Name: Dad and Evies bday bash
Date: 2011/03/29
People: DR, MY, SP,

Photo

Album Name: Dad and Evies bday bash
Date: 2011/03/29
People: SP,
Photo Name: sleepy

Event 15

Photo

Album Desc.: Random parties in Andover. Now with extreme frisbee. And a new house!
Album Name: Tom & Hannah's - 2012 & 2013
Date: 2012/05/27
People: CB, SP,

Photo

Album Desc.: Random parties in Andover. Now with extreme frisbee. And a new house!
Album Name: Tom & Hannah's - 2012 & 2013
Date: 2012/05/27
People: SP, IE, SM,

Photo

Album Desc.: Random parties in Andover. Now with extreme frisbee. And a new house!
Album Name: Tom & Hannah's - 2012 & 2013
Date: 2012/05/27
People: CB, SP, BP,

Photo

Album Desc.: Awesome party with lots of good food, good drinks and good friends.
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Eurovision Party 2012 - Baku (near Andover)
Date: 2012/05/28
People: IE, UM, EH, SP, EP, SE,

Photo

Album Desc.: Awesome party with lots of good food, good drinks and good friends.
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Eurovision Party 2012 - Baku (near Andover)
Date: 2012/05/28
People: SE, UM, CB, EH, SP, IE,

Photo

Album Desc.: Awesome party with lots of good food, good drinks and good friends.
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Eurovision Party 2012 - Baku (near Andover)
Date: 2012/05/28
People: CB, SP,
Photo Name: Ben hiding behind Rhi's hair...

Photo

Album Desc.: Awesome party with lots of good food, good drinks and good friends.
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Eurovision Party 2012 - Baku (near Andover)
Date: 2012/05/28
People: IE, CB, SP,

Photo

Album Desc.: Awesome party with lots of good food, good drinks and good friends.
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Eurovision Party 2012 - Baku (near Andover)
Date: 2012/05/28
People: CB, IE, UM, SP, EH,

Photo

Album Desc.: Awesome party with lots of good food, good drinks and good friends.
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Eurovision Party 2012 - Baku (near Andover)
Date: 2012/05/28
People: SP,

Photo

Album Desc.: Awesome party with lots of good food, good drinks and good friends.
Album Loc.: Andover, Hampshire, United Kingdom
Album Name: Eurovision Party 2012 - Baku (near Andover)
Date: 2012/05/28
People: SP, EH, SE,
Photo Name: That is an awesome expression Dan has there!

Event 16

Status Message

Date: 2012/09/14
Status: Come to the Senedd in Cardiff Bay tomorrow at 12pm for Coast Along- walk for water so others don't have to!

Event 17

Status Message

Date: 2012/06/14
Status: 10 kellog's variety packs, 60 apples & bananas, a few hundred biscuits, 5 packs of spaghetti, a tonne of mince...

Event 18

Status Message

Date: 2012/08/01
Status: Vietnam was amazing... I want to go back already.

Status Message

Date: 2012/07/15
Status: Hello Mr. Giant Jelly fish....no I don't want a hug!

Status Message

Date: 2012/07/09
Status: no PhD for three whole weeks......not sure how I feel about that!

Status Message

Date: 2012/06/28
Status: In two weeks will be on a boat in Halong Bay...that's if we've avoided all the flight scams, visa scams, taxi ...

Status Message

Date: 2012/04/04
Status: Can't believe we just booked flights to Vietnam! Has anybody got any tips of where to stay/ what to do?

Event 19

Status Message

Date: 2012/08/13
Status: Leanne Wood is at Green Man Festival (legend!)....is it acceptable to accost her for a PhD interview?!

Event 20

Status Message

Date: 2012/02/20
Status: Worst place in Wales to break your ankle? Summit of Mt Snowdon. Fact.

Status Message

Date: 2012/02/27
Status: Wooohoo bright pink cast!!!!

Status Message

Date: 2012/04/02
Status: Hello skinny, hairy, purpley leg. So nice to see you again!

Status Message

Date: 2012/04/10
Status: can walk!

Event 21

Photo

Album Name: Profile Pictures
Date: 2012/01/31

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP,

Photo

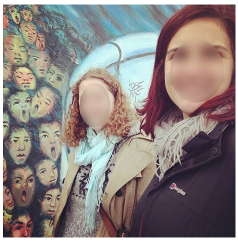
Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, NW,

Event 22

Photo

Album Name: Profile Pictures
Date: 2012/11/19

Photo



Album Name: Profile Pictures
Date: 2012/11/19

Event 23

Photo

Album Desc.: February onwards :)
Album Name: Spring '12
Date: 2012/04/08
People: CB, EB, TH, SP,

Photo

Album Desc.: February onwards :)
Album Name: Spring '12
Date: 2012/04/08
People: CB, EH, EB, TH, SP,

Photo

Album Desc.: February onwards :)
Album Name: Spring '12
Date: 2012/04/08
People: EH, EB, TH, SP,

Photo

Album Desc.: February onwards :)
Album Name: Spring '12
Date: 2012/04/08
People: CB, TH, SP,

Photo

Album Desc.: February onwards :)
Album Name: Spring '12
Date: 2012/04/08
People: CB, TH, SP,

Photo
Album Desc.: Wet Easter trip to see friends and family.
Album Name: Easter Trip to Wales
Date: 2012/04/09
People: UM, SP, TH, CB,

Photo
Album Desc.: Wet Easter trip to see friends and family.
Album Name: Easter Trip to Wales
Date: 2012/04/09
People: UM, LW, SP, TH, CB,

Event 24

Photo
Album Name: Autumn: such a cosy season <3
Date: 2012/09/29
People: UM, EH, SP,

Photo
Album Name: Autumn: such a cosy season <3
Date: 2012/09/29
People: UM, SP,

Photo
Album Name: Autumn: such a cosy season <3
Date: 2012/09/29
People: UM, EH, SP,
Photo Name: Tah-dah‼

Photo
Album Name: Autumn: such a cosy season <3
Date: 2012/09/29
People: SP,
Photo Name: Concentration

Photo
Album Name: Autumn: such a cosy season <3
Date: 2012/09/29
People: SP,

Photo
Album Name: Autumn: such a cosy season <3
Date: 2012/09/29
People: SP,
Photo Name: Cake- stand making!

Photo
Album Name: Mobile Uploads
Date: 2012/09/22
People: UM, SP,
Photo Name: So proud of their homemade cake stand

Event 25

Photo

Album Name: Spring '13
Date: 2013/03/13
People: SP, SA,

Photo

Album Name: Spring '13
Date: 2013/03/13
People: GM, SP, SA,

Photo

Album Name: Spring '13
Date: 2013/03/13
People: SM, SP,

Event 26

Photo

Album Name: Timeline Photos
Date: 2012/09/23
People: SP, EH,
Photo Name: Nan's Birthday

Photo

Album Name: Mobile Uploads
Date: 2012/01/21
People: SP, CB,
Photo Name: Posh birthday no. 2

Event 27

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, NW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, NW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, NW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

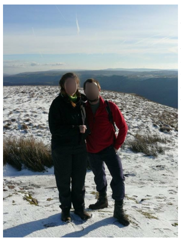
Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Photo



Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Photo

Album Loc.: Brecon Beacons National Park
Album Name: Winter 2012
Date: 2012/01/29
People: SP, BW,

Event 28

Status Message

Date: 2013/05/25
Status: Day spent chilling out in Tampere with 30 lovely people from all over the world. Okay, so maybe academia is th...

Event 29

Event

Date: 2013/07/13
Location: Rumney Scout Hall
Organizer: NC

Event 30

Event

Date: 2013/07/25
Desc.: As most of you must know by now, Ed and I are sadly (and excitedly) leaving the city, the country and indeed t...
Location: 60 Africa Gardens
Organizer: MW

Event 31

Event

Date: 2013/06/27
Desc.: Tanwen and I will be spending a very long weekend in Cardiff at the end of June and basically this is just a w...
Location: Cardiff
Organizer: SD

Event 32

Photo

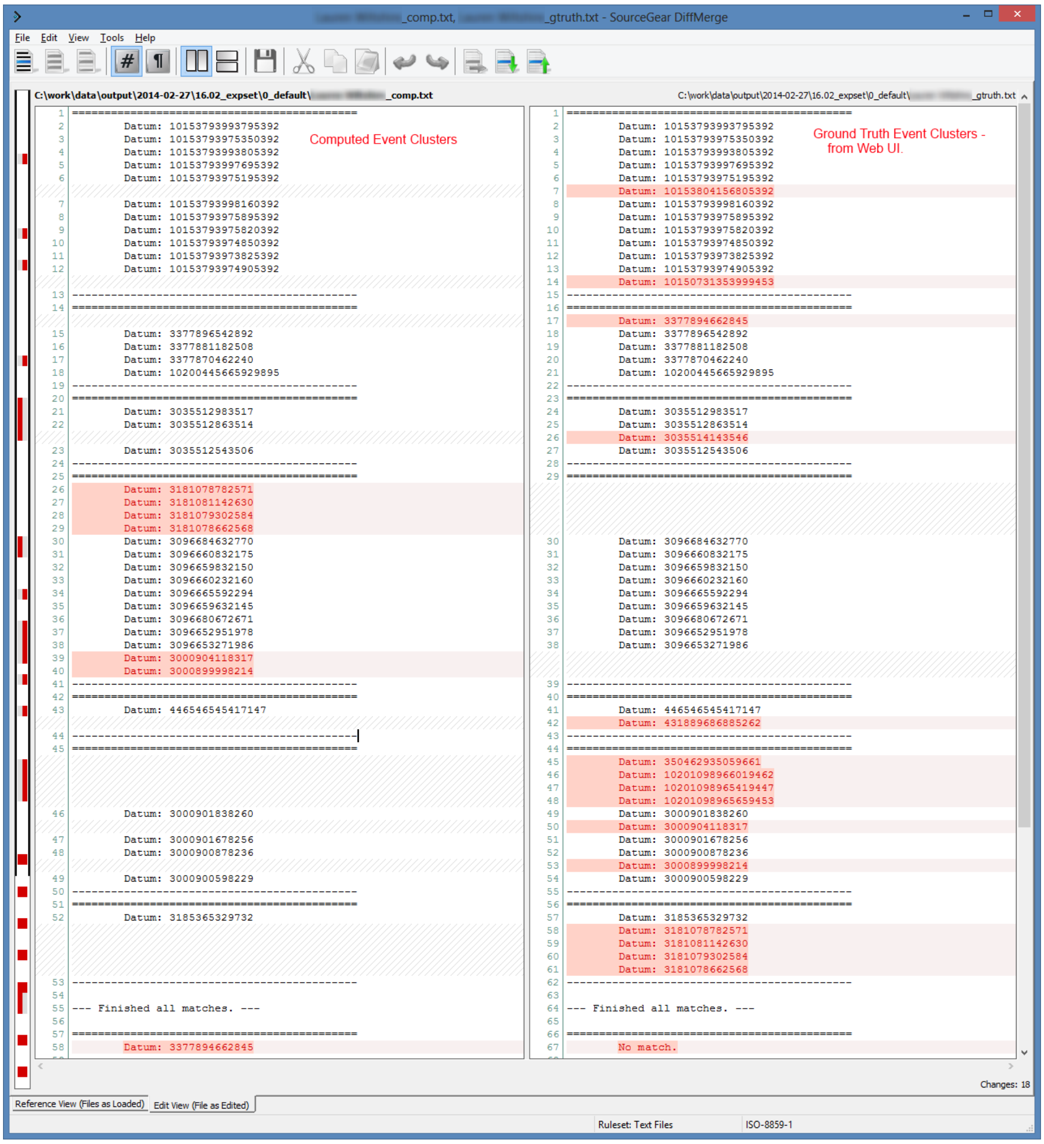
Album Loc.: Bridgend
Album Name: 80s party
Date: 2013/10/20
People: NM, SP, ED,
Photo Name: Awesome party last night, thank you all for coming :)

Photo

Album Loc.: Bridgend
Album Name: 80s party
Date: 2013/10/20
People: SP,

Appendix C

Using a *diff* Tool for Evaluation of Clustering Performance



[redacted]_comp.txt, [redacted]_gtruth.txt - SourceGear DiffMerge

C:\work\data\output\2014-02-27\16.02_expset\0_default\[redacted]_comp.txt

Computed Event Clusters

```
============================================
        Datum: 10153793993795392
        Datum: 10153793975350392
        Datum: 10153793993805392
        Datum: 10153793997695392
        Datum: 10153793975195392
        Datum: 10153793998160392
        Datum: 10153793975895392
        Datum: 10153793975820392
        Datum: 10153793974850392
        Datum: 10153793973825392
        Datum: 10153793974905392
--------------------------------------------
============================================
        Datum: 3377896542892
        Datum: 3377881182508
        Datum: 3377870462240
        Datum: 10200445665929895
--------------------------------------------
============================================
        Datum: 3035512983517
        Datum: 3035512863514
        Datum: 3035512543506
--------------------------------------------
============================================
        Datum: 3181078782571
        Datum: 3181081142630
        Datum: 3181079302584
        Datum: 3181078662568
        Datum: 3096684632770
        Datum: 3096660832175
        Datum: 3096659832150
        Datum: 3096660232160
        Datum: 3096665592294
        Datum: 3096659632145
        Datum: 3096680672671
        Datum: 3096652951978
        Datum: 3096653271986
        Datum: 3000904118317
        Datum: 3000899998214
--------------------------------------------
============================================
        Datum: 446546545417147
--------------------------------------------
============================================
        Datum: 3000901838260
        Datum: 3000901678256
        Datum: 3000900878236
        Datum: 3000900598229
--------------------------------------------
============================================
        Datum: 3185365329732
--------------------------------------------

--- Finished all matches. ---

============================================
        Datum: 3377894662845
```

C:\work\data\output\2014-02-27\16.02_expset\0_default\[redacted]_gtruth.txt

Ground Truth Event Clusters - from Web UI.

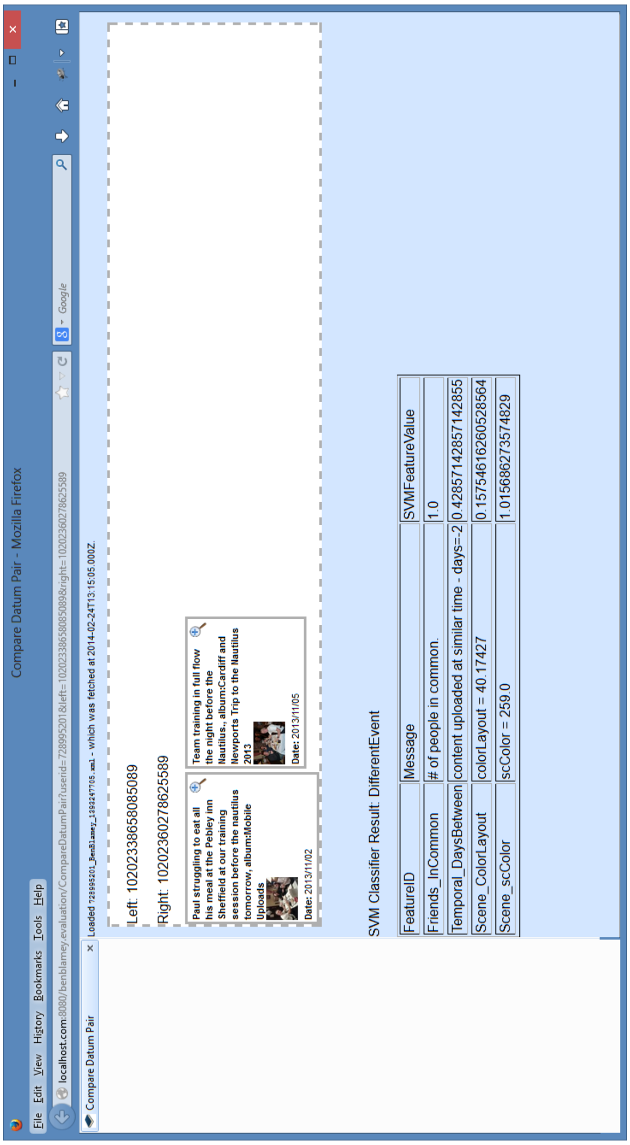
```
============================================
        Datum: 10153793993795392
        Datum: 10153793975350392
        Datum: 10153793993805392
        Datum: 10153793997695392
        Datum: 10153793975195392
        Datum: 10153804156805392
        Datum: 10153793998160392
        Datum: 10153793975895392
        Datum: 10153793975820392
        Datum: 10153793974850392
        Datum: 10153793973825392
        Datum: 10153793974905392
        Datum: 10150731353999453
--------------------------------------------
============================================
        Datum: 3377894662845
        Datum: 3377896542892
        Datum: 3377881182508
        Datum: 3377870462240
        Datum: 10200445665929895
--------------------------------------------
============================================
        Datum: 3035512983517
        Datum: 3035512863514
        Datum: 3035514143546
        Datum: 3035512543506
--------------------------------------------
============================================
        Datum: 3096684632770
        Datum: 3096660832175
        Datum: 3096659832150
        Datum: 3096660232160
        Datum: 3096665592294
        Datum: 3096659632145
        Datum: 3096680672671
        Datum: 3096652951978
        Datum: 3096653271986
--------------------------------------------
============================================
        Datum: 446546545417147
        Datum: 431889686885262
--------------------------------------------
============================================
        Datum: 350462935059661
        Datum: 10201098966019462
        Datum: 10201098965419447
        Datum: 10201098965659453
        Datum: 3000901838260
        Datum: 3000904118317
        Datum: 3000901678256
        Datum: 3000900878236
        Datum: 3000899998214
        Datum: 3000900598229
--------------------------------------------
============================================
        Datum: 3185365329732
        Datum: 3181078782571
        Datum: 3181081142630
        Datum: 3181079302584
        Datum: 3181078662568
--------------------------------------------

--- Finished all matches. ---

============================================
        No match.
```

Changes: 18

Figure C.1: Using a *diff* Tool for Evaluation of Clustering Performance

Appendix D

Comparing a Document Pair



Compare Datum Pair - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost.com:8080/benblamey.evaluation/CompareDatumPair?userid=728995201&left=10202338658085089&right=10202360278625589
Google
Compare Datum Pair

Loaded 728995201_BenBlamey_1393247705.xml - which was fetched at 2014-02-24T13:15:05.000Z.

Left: 10202338658085089

Right: 10202360278625589

Paul struggling to eat all his meal at the Pebley inn Sheffield at our training session before the nautilus tomorrow, album:Mobile Uploads

Date: 2013/11/02

Team training in full flow the night before the Nautilus., album:Cardiff and Newports Trip to the Nautilus 2013

Date: 2013/11/05

SVM Classifier Result: DifferentEvent

FeatureID	Message	SVMFeatureValue
Friends_InCommon	# of people in common.	1.0
Temporal_DaysBetween	content uploaded at similar time - days=-2	0.42857142857142855
Scene_ColorLayout	colorLayout = 40.17427	0.15754616260528564
Scene_scColor	scColor = 259.0	1.0156862735574829

Figure D.1: Comparing a Document Pair in the Web Debug View

Bibliography

Abbasi, R., S. Chernov, W. Nejdl, R. Paiu, and S. Staab (2009). "Exploiting Flickr Tags and Groups for Finding Landmark Photos". In: *Advances in Information Retrieval*, pp. 654–661.

Aggarwal, C.C. (2011). *Social Network Data Analytics*. Springer. ISBN: 9781441984616.

Aiken, Justin (2014). "Building an Open Source Social Media Aggregation Timeline". MA thesis. Southern Utah University.

Alahmari, Muteeb Saad (2012). "Personal Semantic Timeframe". MA thesis.

Alias-i (2008). *LingPipe 4.1.0*. URL: `http://alias-i.com/lingpipe`.

Allen, James F. (1981). "An Interval-Based Representation of Temporal Knowledge". In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI'81)*. Morgan Kaufmann, pp. 221–226.

Amatriain, Xavier (2012). "More data or better models?" URL: `http://technocalifornia.blogspot.co.uk/2012/07/more-data-or-better-models.html`.

Angeli, Gabor, Christopher Manning, and Daniel Jurafsky (2012). "Parsing Time: Learning to Interpret Time Expressions". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 446–455.

Apted, Trent, Judy Kay, and Aaron Quigley (2006). "Tabletop Sharing of Digital Photographs for the Elderly". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, pp. 781–790.

Ardissono, Liliana, Anna Goy, Giovanna Petrone, and Marino Segnan (2009). "From Service Clouds to User-Centric Personal Clouds". In: *IEEE International Conference on Cloud Computing (CLOUD'09)*, pp. 1–8.

Augé, Marc (2008). *Non-Places. An Introduction to Supermodernity*. Trans. by John Howe. 2nd ed. Verso.

Banks, Richard (2011). *The Future of Looking Back*. Microsoft Press. ISBN: 0735658064 9780735658066.

Banos, Vangelis, Nikos Baltas, and Yannis Manolopoulos (2012). "Trends in Blog Preservation". In: *Proceedings of the 14th International Conference on Enterprise Information Systems (ICEIS2012)*.

Barzilay, Regina (2010). *Natural Language Processing (Lectures)*. MIT. URL: `http://people.csail.mit.edu/regina/6864/`.

Baum, Leonard E, Ted Petrie, George Soules, and Norman Weiss (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains". In: *The Annals of Mathematical Statistics* 41 (1), pp. 164–171.

Becker, Hila, Dan Iter, Mor Naaman, and Luis Gravano (2012). "Identifying content for planned events across social media sites". In: *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. WSDM '12. Seattle, Washington, USA: ACM, pp. 533–542.

Becker, Hila, Mor Naaman, and Luis Gravano (2010). “Learning Similarity Metrics for Event Identification in Social Media”. In: *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. ACM, pp. 291–300.

Bederson, Benjamin B (2001). “PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps”. In: *Proceedings of the 14th annual ACM Symposium on User Interface Software and Technology*. ACM, pp. 71–80.

Bell, C.G. and J. Gemmell (2009). *Total Recall: How the E-Memory Revolution Will Change Everything*. Dutton. ISBN: 9780525951346.

Belli, R.F. (1998). “The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys”. In: *Memory* 6 (4), pp. 383–406.

Blamey, Ben, Tom Crick, and Giles Oatley (2012). “R U :-) or :-(? Character- vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora”. In: *AI-2012: The Thirty-second SGAI International Conference*. Springer.

Blamey, Ben, Tom Crick, and Giles Oatley (2013). “‘The First Day of Summer’: Parsing Temporal Expressions with Distributed Semantics”. English. In: *Research and Development in Intelligent Systems XXX*. Ed. by Max Bramer and Miltos Petridis. Springer International Publishing, pp. 389–402.

Blei, David M and Jon D McAuliffe (2007). “Supervised Topic Models”. In: *Advances in Neural Information Processing Systems*, pp. 121–128.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008). “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* (10).

Boll, Susanne, Philip Sandhaus, Ansgar Scherp, and Sabine Thieme (2007). “MetaXa – Context– and content–driven metadata enhancement for personal photo albums”. In: *Prodeedings of the International Multimedia Modeling Conference (MMM 2007)*.

Bontcheva, Kalina, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani (2013). “TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

Bontcheva, Kalina and Dominic Rout (2014). “Making sense of social media streams through semantics: a survey”. In: *Semantic Web Journal* (5), pp. 373–403.

Brants, Thorsten (2000). “TnT: a statistical part-of-speech tagger”. In: *Proceedings of the 6th Conference on Applied Natural Language Processing*. Association for Computational Linguistics, pp. 224–231.

Bredin, Hervé and Gérard Chollet (2007). “Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2.

Brill, Eric (1992). “A simple rule-based part of speech tagger”. In: *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, pp. 112–116.

Brody, Samuel and Nicholas Diakopoulos (2011). “Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 562–570.

Brown, Peter F., Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai (1992). “Class-Based n-gram Models of Natural Language”. In: *Computational Linguistics* 18 (4), pp. 467–479.

Brucato, Matteo, Leon Derczynski, Hector Llorens, Kalina Bontcheva, and Christian S Jensen (2013). “Recognising and Interpreting Named Temporal Expressions”. In: *Recent Advances in Natural Language Processing (RANLP 2013)*.

Buckler, Ott Tucker (2006). “An Introduction to the History of Scrapbooks”. In: *Scrapbooks in American Life*. Temple University Press, pp. 1–25.

Bunescu, Razvan C and Marius Paca (2006). “Using Encyclopedic Knowledge for Named entity Disambiguation.” In: *Proceedings of the European Chapter of the ACL Conference*. Vol. 6, pp. 9–16.

Byrne, Daragh, Aiden Doherty, Cees Snoek, Gareth Jones, and Alan Smeaton (2010). “Everyday Concept Detection in Visual Lifelogs: Validation, Relationships and Trends”. In: *Multimedia Tools and Applications* 49 (1), pp. 119–144.

Cardie, Claire (2011). *Claire Cardie on Applications of Natural Language Processing*. URL: `http://www.cornell.edu/video/?videoID=1766` (Accessed on: 02/03/2012).

Carpenter, Bob (2010). *Yahoo Group Message Discussion*. URL: `https://groups.yahoo.com/neo/groups/LingPipe/conversations/messages/917`.

Carpenter, Bob (2011). *LingPipe: Sentiment Analysis Tutorial*. URL: `http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html`.

Cashmore, Pete (2009). *Women Rule the Social Web*. URL: `http://mashable.com/2009/10/03/women-rule-the-social-web/`.

Cavnar, William B. and John M. Trenkle (1994). “N-Gram-Based Text Categorization”. In: *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)*, pp. 161–175.

Chang, Angel X. and Christopher Manning (2012). “SUTime: A library for recognizing and normalizing time expressions”. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC’12)*.

Chang, Angel X. and Christopher D. Manning (2014). *TokensRegex: Defining cascaded regular expressions over tokens*. Tech. rep. CSTR 2014-02. Department of Computer Science, Stanford University.

Chang, Chih-Chung and Chih-Jen Lin (2011). “LIBSVM: A Library for Support Vector Machines”. In: *ACM Transactions on Intelligent Systems and Technology* 2 (3). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 27:1–27:27.

Charniak, Eugene, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz (1993). “Equations for part-of-speech tagging”. In: *Conference of the Association for the Advancement of Artificial Intelligence*, pp. 784–789.

Cheng, Jacqui (2012). *Three years later, deleting your photos on Facebook now actually works*. URL: `http://arstechnica.com/business/2012/08/facebook-finally-changes-photo-deletion-policy-after-3-years-of-reporting/`.

Chinchor, Nancy (1997). *MUC-7 Named Entity Task Definition*. URL: `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html` (Accessed on: 04/11/2014).

Cho, Sung-Bae, K-J Kim, Keum Sung Hwang, and In-Ji Song (2007). “AniDiary: Daily cartoon-style diary exploits Bayesian networks”. In: *Pervasive Computing* 6 (3), pp. 66–75.

Chomsky, N. (1956). “Three models for the description of language”. In: *Information Theory, IRE Transactions on* 2 (3), pp. 113–124.

Choney, Suzanne (2012). *Facebook Timeline poll: ‘Overwhelming negative’ reaction*. NBC News. URL: `http://www.nbcnews.com/technology/technolog/facebook-timeline-poll-overwhelming-negative-reaction-84717`.

Clauset, Aaron, Mark EJ Newman, and Cristopher Moore (2004). "Finding community structure in very large networks". In: *Physical Review E* 70 (6), p. 066111.

Cluley, Graham (2012). *Poll reveals widespread concern over Facebook Timeline.* URL: http://nakedsecurity.sophos.com/2012/01/27/poll-reveals-widespread-concern-over-facebook-timeline/.

Collins, Michael (2002). "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms". In: *Proceedings of the ACL-02 conference on Empirical methods in Natural Language Processing.* Vol. 10. Association for Computational Linguistics, pp. 1–8.

Conway, M. A. (1999). "Autobiographical knowledge and autobiographical memories". In: *Studies in Autobiographical Memory.* Ed. by David C. Rubin.

Copestake, Ann (2007). *Natural Language Processing (Lectures).* University of Cambridge.

Corinna Cortes, Vladimir Vapnik (1995). "Support-Vector Networks". In: *Machine Learning* (20). Ed. by Lorenza Saitta, pp. 273–297.

Crockford, Douglas (2006). "JSON: The Fat-Free Alternative to XML". In: *XML 2006.* URL: http://www.json.org/fatfree.html.

Croll, Alistair (2015). *Year Zero: Our life timelines begin.* URL: http://radar.oreilly.com/2015/03/year-zero-our-life-timelines-begin.html.

Cunningham et al. (2001–2014). *Developing Language Processing Components with GATE Version 8 (a User Guide).* University of Sheffield Department of Computer Science. URL: http://gate.ac.uk/sale/tao/split.html.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan (2002). "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications". In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).*

Daumé, Hal III (2006). *Natural Language Processing blog: Beating an N-Gram.* URL: http://nlpers.blogspot.co.uk/2006/06/beating-n-gram.html.

Dietze, Stefan, Diana Maynard, Elena Demidova, Thomas Risse, Wim Peters, Katerina Doka, and Yannis Stavrakas (2012). "Entity Extraction and Consolidation for Social Web Content Preservation". In: *International Workshop on Semantic Digital Archives.* Vol. 912, pp. 18–29.

DiMicco, Joan Morris and David R. Millen (2007). "Identity Management: Multiple Presentations of self in Facebook". In: *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP'07).* ACM, pp. 383–386.

Edelstein, Orit, Michael Factor, Ross King, Thomas Risse, Eliot Salant, and Philip Taylor (2011). "Evolving Domains, Problems and Solutions for Long Term Digital Preservation". In: *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011).*

Elsner, Micha and Warren Schudy (2009). "Bounding and comparing methods for correlation clustering beyond ILP". In: *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing.* Association for Computational Linguistics, pp. 19–27.

Endomondo (2012). *About Endomondo.* URL: http://blog.endomondo.com/about/.

Ester, Martin, Hans-peter Kriegel, Jörg S, and Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of The Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231.

Facebook (2014a). *Shared Albums.* URL: https://www.facebook.com/help/151724461692727/ (Accessed on: 03/26/2014).

Facebook (2014b). *Statement of Rights and Responsibilities.* URL: https://www.facebook.com/legal/terms (Accessed on: 03/26/2014).

Facebook (2014c). *What happens when a deceased person's account is memorialized?* URL: https://www.facebook.com/help/103897939701143 (Accessed on: 03/26/2014).

Facebook (2012a). *Introducing Timeline.* URL: https://www.facebook.com/about/timeline.

Facebook (2012b). *SEC Registration Statement (IPO).* URL: http://www.sec.gov/Archives/edgar/data/1326801/000119312512175673/d287954ds1a.htm.

Garfinkel, Simson and David Cox (2009). *Finding and Archiving the Internet Footprint.* URL: http://simson.net/clips/academic/2009.BL.InternetFootprint.pdf.

Gemmell, Jim, Gordon Bell, and Roger Lueder (2006). "MyLifeBits: a personal database for everything". In: *Communictions of the ACM* 49 (1), pp. 88–95. ISSN: 0001-0782.

Gerlitz, Carolin (2012). "Acting on Data. Temporality and Self-Evaluation in Social Media". URL: http://eprints.gold.ac.uk/7076.

Gimpel, Kevin et al. (2011). "Part-of-speech tagging for twitter: Annotation, features, and experiments". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2.* Association for Computational Linguistics, pp. 42–47.

Girvan, Michelle and Mark EJ Newman (2001). "Community structure in social and biological networks".

Go, Alec, Richa Bhayani, and Lei Huang (2009). "Twitter Sentiment Classification using Distant Supervision". In: *Processing* 150 (12). Ed. by Roger GEditor Walker, pp. 1–6.

Gold, Kevin (2012). *Norvig vs. Chomsky and the Fight for the Future of AI.* URL: http://www.tor.com/blogs/2011/06/norvig-vs-chomsky-and-the-fight-for-the-future-of-ai?goback=.gde_130377_member_176125409.

Gong, Bo, Utz Westermann, Srikanth Agaram, and Ramesh Jain (2006). "Event Discovery in Multimedia Reconnaissance Data Using Spatio-Temporal Clustering". In: *Proceedings of the AAAI Workshop on Event Extraction and Synthesis (EES'06).*

Gouveia, R. and E. Karapanos (2013). "Footprint Tracker: Supporting Diary Studies with Lifelogging". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, pp. 2921–2930.

Greene, Barbara B and Gerald M Rubin (1971). *Automatic Grammatical Tagging of English.* Department of Linguistics, Brown University.

Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Discovery.* Springer.

Gunes, Hatice and Massimo Piccardi (2005). "Affect recognition from face and body: early fusion vs. late fusion". In: *Proceedings of IEEE International Conference on Systems, Man and Cybernetics.* Vol. 4. IEEE, pp. 3437–3443.

Gurrin, Cathal, Zhengwei Qiu, Mark Hughes, Niamh Caprani, Aiden R Doherty, Steve E Hodges, and Alan F Smeaton (2013). "The smartphone as a platform for wearable cameras in health research". In: *American Journal of Preventive Medicine* 44 (3), pp. 308–313.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks (2006). "A Closer Look at Skip-gram Modelling". In: *Proceedings of the 5th International Conference on Language Resources And Evaluation.* Genoa, Italy, pp. 1222–1225.

Habernal, Ivan, Tomá Ptáek, and Josef Steinberger (2014). "Supervised Sentiment Analysis in Czech Social Media". In: *Information Processing & Management* 50 (5), pp. 693–707.

Hagiwara, Masato (2010). *Unnatural Language Processing Contest 2nd will be held at NLP2011.* URL: http://blog.lilyx.net/2010/11/28/unnatural-language-processing-contest-2nd-will-be-held-at-nlp2011/.

Halevy, A., P. Norvig, and F. Pereira (2009). “The unreasonable effectiveness of data”. In: *Intelligent Systems, IEEE* 24 (2), pp. 8–12.

Harvilla, Rob (2010). *Here Is What Christopher R. Weingarten's "Tweetbox," a/k/a The Custom-Built Card Catalog Containing All 1,000 Of His 2009 Twitter-Based Record Reviews, Looks Like.* URL: http://blogs.villagevoice.com/music/2010/03/here_is_what_ch.php.

Hodges, Steve, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood (2006). “SenseCam: A Retrospective Memory Aid”. In: *Proceedings of the 8th International Conference on Ubicomp*, pp. 177–193.

Hsu, Chih-Wei, Chang, Chih-Chung, and Chih-Jen Lin (2010). *A Practical Guide to Support Vector Classification.* URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Huang, Qian and B. Dom (1995). “Quantitative methods of evaluating image segmentation”. In: *Proceedings of the International Conference on Image Processing.* Vol. 3, pp. 53–56.

Internet Engineering Task Force (2012). *The OAuth 2.0 Authorization Framework.* Ed. by D. Hardt. URL: https://tools.ietf.org/html/rfc6749.

James, Jeff (2011). “How Facebook Handles Image EXIF Data”. In: URL: http://windowsitpro.com/blog/how-facebook-handles-image-exif-data.

Jordan, A and A Ng (2002). “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes”. In: *Advances in Neural Information Processing Systems* 14, p. 841.

Jurafsky, Daniel and James Martin (2009). *Speech and Language Processing.* 2nd ed. Prentice Hall Series in Artificial Intelligence. Pearson.

Kirk, David S., Shahram Izadi, Abigail Sellen, Stuart Taylor, Richard Banks, and Otmar Hilliges (2010). “Opening up the family archive”. In: *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work (CSCW '10).* ACM, pp. 261–270.

Kleene, S.C. (1951). *Representation of Events in Nerve Nets and Finite Automata.* Tech. rep.

Klein, Dan, Joseph Smarr, Huy Nguyen, and Christopher Manning (2003). “Named entity recognition with character-level models”. In: *Proceedings of the 7th Conference on Natural language Learning of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (2003).* Vol. 4. CONLL '03. Association for Computational Linguistics, pp. 180–183.

Klein, Sheldon and Robert F. Simmons (1963). “A Computational Approach to Grammatical Coding of English Words”. In: *Journal of the ACM* 10 (3), pp. 334–347.

Krupka, George R and Kevin Hausman (1998). “IsoQuest Inc.: Description of the NetOwl (TM) Extractor System as Used for MUC-7”. In: *Proceedings of the 7th Message Understanding Conference (MUC-7).* Vol. 7.

Kuchinsky, A., C. Pering, M.L. Creech, D. Freeze, B. Serra, and J. Gwizdka (1999). “FotoFile: a Consumer Multimedia Organization and Retrieval System”. In: *Proceedings of the SIGCHI Conference on Human factors in Computing Systems: The CHI is the Limit.* ACM, pp. 496–503.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence

Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*, pp. 282–289.

Landry, Brian M. (2008). "Storytelling with digital photographs: supporting the practice, understanding the benefit". In: *Human Factors in Computing Systems (CHI'08)*. Florence, Italy: ACM, pp. 2657–2660.

Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce (2006). "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE, pp. 2169–2178.

Lee, Pei, Laks V. S. Lakshmanan, and Evangelos E. Milios (2013). "Event Evolution Tracking from Streaming Social Posts". URL: `http://arxiv.org/abs/1311.5978`.

Li, Xing, Pin Wang, and Hong Li (2015). "Design and Algorithm Optimization of P2P Mobile Monitoring Network-Based Facial Recognition System". In: 721, pp. 771–774.

Liang, Percy (2005). "Semi-Supervised Learning for Natural Language". MA thesis. Massachusetts Institute of Technology.

Lim, Joo-Hwee, Qi Tian, and Philippe Mulhem (2003). "Home Photo Content Modeling for Personalized Event-Based Retrieval". In: *IEEE Multimedia* 10, pp. 28–37.

Lin, Chin-Yew (2004). "Rouge: A package for automatic evaluation of summaries". In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson (2005). *TIDES 2005 Standard for the Annotation of Temporal Expressions*.

Liu, Bing, Chee Wee Chin, and Hwee Tou Ng (2003). "Mining topic-specific concepts and definitions on the web". In: *Proceedings of the 12th International Conference on World Wide Web*. ACM, pp. 251–260.

Liu, Yan, Alexandru Niculescu-Mizil, and Wojciech Gryc (2009). "Topic-link LDA: joint models of topic and author community". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, pp. 665–672.

Lu, Ye, Chunhui Hu, Xingquan Zhu, HongJiang Zhang, and Qiang Yang (2000). "A unified framework for semantics and feature based relevance feedback in image retrieval systems". In: *Proceedings of the 8th ACM International Conference on Multimedia*. ACM, pp. 31–37.

Lundeen, Rich (2013). *Common OAuth Issue You Can Use To Take Over Accounts*. URL: `http://webstersprodigy.net/2013/05/09/common-oauth-issue-you-can-use-to-take-over-accounts/`.

Lux, Mathias (2009). "Caliph & Emir: MPEG-7 Photo Annotation and Retrieval". In: *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*. Beijing, China: ACM, pp. 925–926.

Ma, Jixin and Brian Knight (2003). "Representing The Dividing Instant". In: *The Computer Journal* 46 (2), pp. 213–222.

Mani, Inderjeet and George Wilson (2000). "Robust temporal processing of news". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 69–76.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze (2008). "Evaluation of Clustering". In: *Introduction to Information Retrieval*. Cambridge University Press.

Manning, Christopher and Dan Klein (2003). "Optimization, Maxent Models, and Conditional Estimation without Magic". In: *HLT-NAACL 2003 and ACL 2003*.

Marcus, Adam, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller (2011). “Twitinfo: aggregating and visualizing microblogs for event exploration”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 227–236.

Massimi, M. and A. Charise (2009). “Dying, death, and mortality: towards thanatosensitivity in HCI”. In: *Proceedings of the 27th International Conference on Human Factors in Computing Systems*. ACM, pp. 2459–2468.

Massimi, Michael and Ronald M. Baecker (2010). “A death in the family: opportunities for designing technologies for the bereaved”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. ACM, pp. 1821–1830.

McCallum, Andrew and Wei Li (2003). “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons”. In: *Proceedings of the 7th Conference on Natural Language Learning of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (2003)*. Vol. 4, pp. 188–191.

McCown, Frank and Michael L. Nelson (2009). “What Happens when Facebook is Gone?” In: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL’09)*. ACM, pp. 251–254.

McDaid, Aaron F, Derek Greene, and Neil Hurley (2011). “Normalized mutual information to evaluate overlapping community finding algorithms”.

Mei, T., B. Wang, X.S. Hua, H.Q. Zhou, and S. Li (2006). “Probabilistic multimodality fusion for event based home photo clustering”. In: *IEEE International Conference and Expo on Multimedia*, pp. 1757–1760.

Melcombe, Melissa (2011). “Women’s Perceptions of Identity Construction on Facebook”. MA thesis. Gonzaga University.

Mihalcea, Rada and Andras Csomai (2007). “Wikify!: linking documents to encyclopedic knowledge”. In: *Proceedings of the 16th ACM Conference on Information and Knowledge Management*. ACM, pp. 233–242.

Mikheev, Andrei, Marc Moens, and Claire Grover (1999). “Named entity recognition without gazetteers”. In: *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1–8.

Milne, D. and I.H. Witten (2008). “Learning to link with wikipedia”. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM, pp. 509–518.

Morgan, Richard, Roberto Garigliano, Paul Callaghan, Sanjay Poria, Mark Smith, and Chris Cooper (1995). “University of Durham: description of the LOLITA system as used in MUC-6”. In: *Proceedings of the 6th Conference on Message Understanding*. Association for Computational Linguistics, pp. 71–85.

Mulvenna, M., A. Astell, H. Zheng, and T. Wright (2009). “Reminiscence Systems”. In: *Proceedings of the First International Workshop on Reminiscence Systems (RSW-2009)*. CEUR.

Nagarajan, Meenakshi, Karthik Gomadam, Amit P Sheth, Ajith Ranabahu, Raghava Mutharaju, and Ashutosh Jadhav (2009). “Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences”. In: *Web Information Systems Engineering (WISE 2009)*. Springer, pp. 539–553.

Narr, Sascha, Michael Hülfenhaus, and Sahin Albayrak (2012). “Language-Independent Twitter Sentiment Analysis”. In: *KDML, LWA 2012*.

Newman, Mark EJ (2006). “Modularity and community structure in networks”. In: *Proceedings of the National Academy of Sciences* 103 (23), pp. 8577–8582.

Nichols, Jeffrey, Jalal Mahmud, and Clemens Drews (2012). "Summarizing sporting events using twitter". In: *Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI'12).* ACM, pp. 189–198.

Oatley, Giles, Tom Crick, and Mohamed Mostafa (2015). "Digital Footprints: Envisaging and Analysing Online Behaviour". In: *Proceedings of the Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour.*

O'Connor, Brendan, Michel Krieger, and David Ahn (2010). "TweetMotif: Exploratory Search and Topic Summarization for Twitter". In: *International AAAI Conference on Web and Social Media.* Ed. by William W. Cohen, Samuel Gosling, William W. Cohen, and Samuel Gosling. The AAAI Press.

Odom, William, Richard Banks, David Kirk, Richard Harper, Siân Lindley, and Abigail Sellen (2012). "Technology heirlooms?: considerations for passing down and inheriting digital materials". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, pp. 337–346.

Orlowski, Andrew (2014). *Google wearables: A solution looking for a rich nerd. Some revolutions never happen. This might be one of them.* URL: `http://www.theregister.co.uk/2014/03/20/google_the_world_and_wearables_why_its_still_a_solution_looking_for_a_rich_nerd/?page=2`.

Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, and N. Schneider (2012). *Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances.* Tech. rep. CMU-ML-12-107. Carnegie Mellon University.

Owoputi, O., Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith (2013). "Improved part-of-speech tagging for online conversational text with word clusters". In: *Proceedings of North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pp. 380–390.

Pang, Bo and Lillian Lee (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". In: *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 271–278.

Peng, Fuchun, Dale Schuurmans, and Shaojun Wang (2003a). "Language and Task Independent Text Categorization with Simple Language Models". In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pp. 110–117.

Peng, Fuchun, Dale Schuurmans, Shaojun Wang, and Vlado Keselj (2003b). "Language independent authorship attribution using character level language models". In: *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics (EACL'03).* Vol. 1. Association for Computational Linguistics, pp. 267–274.

Ptácek, Tomá, Ivan Habernal, and Jun Hong (2014). "Sarcasm Detection on Czech and English Twitter". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 213–223.

Ptaszynski, Michal, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi (2011). "Research on Emoticons: Review of the Field and Proposal of Research Framework". In: *Proceedings of the 17th Annual Meeting of The Association for Natural Language Processing (NLP-2011), Organized Session on Un-Natural Language Processing.* Toyohashi, Japan, pp. 1159–1162.

Rabbath, M., P. Sandhaus, and S. Boll (2011). "Large scale flexible event-based clustering from photos in social media". In: *Proceedings of the Third International Conference on Internet Multimedia Computing and Service.* ACM, pp. 26–29.

Rabbath, Mohamad and Susanne Boll (2013). "Detecting Multimedia Contents of Social Events in Social Networks". In: *Social Media Retrieval.* Ed. by Naeem Ramzan, Roelof Zwol, Jong-Seok Lee, Kai Clüver, and Xian-Sheng Hua. Computer Communications and Networks. Springer, pp. 87–111.

Rabbath, Mohamad, Philipp Sandhaus, and Susanne Boll (2010). "Automatic Creation of Photo Books from Stories in Social Media". In: *Proceedings of the 2nd ACM Special Interest Group on Multimedia Workshop on Social Media*. ACM, pp. 15–20.

Rabbath, Mohamad, Philipp Sandhaus, and Susanne Boll (2012). "Analysing Facebook features to support event detection for photo-based Facebook applications". In: *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR'12)*. ACM, 11:1–11:8.

Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara (2007). "Near linear time algorithm to detect community structures in large-scale networks". In: *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* 76 (3), p. 036106.

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D Manning (2009). "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Vol. 1. Association for Computational Linguistics, pp. 248–256.

Rao, Delip, Paul McNamee, and Mark Dredze (2013). "Entity Linking: Finding Extracted Entities in a Knowledge Base". In: *Multi-source, Multi-lingual Information Extraction and Summarization*. Springer, pp. 93–115.

Ratnaparkhi, Adwait et al. (1996). "A maximum entropy model for part-of-speech tagging". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Vol. 1, pp. 133–142.

Read, Jonathon (2005). "Using emoticons to reduce dependency in machine learning techniques for sentiment classification". In: *Proceedings of the ACL Student Research Workshop at The Association of Computationl Linguistics*. Vol. 43. June. Association for Computational Linguistics.

Reddy, Bakkama Srinath (2007). "Evidential Reasoning for Multimodal Fusion in Human Computer Interaction". MA thesis. University of Waterloo.

Redi, Miriam and Bernard Merialdo (2011). "Saliency moments for image categorization". In: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR'11)*. ACM.

Reiner, Fageth, Susanne Boll, and Philippp Sandhaus (2008). "Image selection: no longer a dilemma". In: *Proceedings of the IS&T/SPIE 20th Annual Symposium Electronic Imaging Science and Technology.*

Reuter, Timo and Philipp Cimiano (2012). "Event-based Classification of Social Media Streams". In: *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR'12)*. ACM, 22:1–22:8.

Rhee, Youngho, Juyoun Lee, and IlKu Chang (2010). "Designing Mobile Social Networking Service Through UCD Process: LifeDiary". In: *International Journal of Human-Computer Interaction* 26 (11), pp. 1052–1076.

Ritter, Alan, Sam Clark, Mausam, and Oren Etzioni (2011). "Named Entity Recognition in Tweets: An Experimental Study". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1524–1534.

Ritter, Alan, Mausam, Oren Etzioni, and Sam Clark (2012). "Open domain event extraction from twitter". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. ACM, pp. 1104–1112.

Robertson, Andrew (2014). *Download Your Tagged Facebook Photos Before Your Friends Remove Them.*

Robertson, Scott P., Ravi K. Vatrapu, and Richard Medina (2009). “The social life of social networks: Facebook linkage patterns in the 2008 U.S. presidential election”. In: *Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government*, pp. 6–15.

Rodden, Kerry and Kenneth R. Wood (2003). “How do people manage their digital photographs?” In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’03).* ACM, pp. 409–416.

Rother, Carsten, Lucas Bordeaux, Youssef Hamadi, and Andrew Blake (2006). “AutoCollage”. In: *Proceedings of the Association Computing Machine Special Interest Group on Computer Graphics and Interactive Techniques.* New York, NY, USA: ACM, pp. 847–852.

Rowe, Stacy (1989). “Snapshot Versions of Life”. In: *Visual Anthropology Review* 5 (1), pp. 46–47.

Rybina, Kateryna (2012). “Sentiment Analysis of Contexts Around Query Terms in Documents”. MA thesis. Technische Universitat Dresden.

Salvetti, Franco and Nicolas Nicolov (2006). “Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach”. In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* Association for Computational Linguistics, pp. 137–140.

Sandhaus, Philipp and Susanne Boll (2009). “From usage to annotation”. In: *Proceedings of First ACM SIGMM workshop on Social media.* ACM.

Sandhaus, Philipp and Susanne Boll (2011). “Semantic analysis and retrieval in personal and social photo collections”. In: *Multimedia Tools and Applications* 51 (1), pp. 5–33.

Sandhaus, Philipp, Mohamad Rabbath, and Susanne Boll (2010). “Blog2Book: transforming blogs into photo books employing aesthetic principles”. In: *Proceedings of International Conference on Multimedia 2010 (MM ’10).* ACM.

Sandhaus, Philipp, Sabine Thieme, and Susanne Boll (2008). “Processes of Photo Book Production”. In: *Multimedia Systems* 14 (6), pp. 351–357.

Sas, Corina and Steve Whittaker (2013). “Design for Forgetting: Disposing of Digital Possessions After a Breakup”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems‘.* ACM, pp. 1823–1832.

Schenkel, Ralf and Marc Spaniol (2011). *Lecture “Web Dynamics”.* Figures from Facebook Website, quoted in a lecture. Max-Planck-Institut für Informatik. URL: `http://www.mpi-inf.mpg.de/departments/d5/teaching/ss10/dyn/slides/dyn-kap1.pdf` (Accessed on: 11/08/2012).

Scoble, Robert (2012). *The coming automatic, freaky, contextual world and why we’re writing a book about it.* URL: `http://scobleizer.com/2012/07/17/the-coming-automatic-freaky-contextual-world-and-why-were-writing-a-book-about-it/`.

Sellen, A.J. and S. Whittaker (2010). “Beyond total capture: a constructive critique of lifelogging”. In: *Communications of the ACM* 53 (5), pp. 70–77.

Shamma, David A, Lyndon Kennedy, and Elizabeth F Churchill (2009). “Tweet the Debates: Understanding Community Annotation of Uncollected Sources”. In: *Proceedings of the First ACM Special Interest Group on Multimedia Workshop on Social Media*, pp. 3–10.

Shannon, Claude Elwood (1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 3 (27), pp. 379–423.

Sharifi, Beaux, Mark-Anthony Hutton, and Jugal Kalita (2010). "Summarizing microblogs automatically". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 685–688.

Shimojo, Akira, Saori Kamada, Shinsuke Matsumoto, and Masahide Nakamura (2010). "On integrating heterogeneous lifelog services". In: *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services*. ACM, pp. 263–272.

Sinn, Donghee and SueYeon Syn (2014). "Personal Documentation on a Social Network Site: Facebook, a Collection of Moments From Your Life?" In: *Archival Science* 14 (2), pp. 95–124.

Smith, Andrew, Trevor Cohn, and Miles Osborne (2005). "Logarithmic opinion pools for conditional random fields". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 18–25.

Smith, Andrew and Miles Osborne (2006). "Using Gazetteers in Discriminative Information Extraction". In: *In CoNLL-X, Tenth Conference on Computational Natural Language Learning*, pp. 10–8.

Souza, C. R. (2012). *The Accord.NET Framework*. URL: `http://accord.googlecode.com`.

Spitkovsky, Valentin I. and Angel X. Chang (2012). "A Cross-Lingual Dictionary for English Wikipedia Concepts". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Spoustová, Drahomra, Jan Haji, Jan Raab, and Miroslav Spousta (2009). "Semi-supervised Training for the Averaged Perceptron POS Tagger". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*. Association for Computational Linguistics, pp. 763–771.

Steiner, Thomas, Ruben Verborgh, Raphaël Troncy, Giuseppe Rizzo, José Luis Redondo Garcia, Joaquim Gabarró Vallés, and Rik Van de Walle (2012). "Modeling and Reconciling Nightlife Events from Public Event Databases for the Automatic Generation of Magazines". In: *Proceedings of the Second Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*.

Stelmaszewska, Hanna, Bob Fields, and Ann Blandford (2008). "The roles of time, place, value and relationships in collocated photo sharing with camera phones". In: *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction (BCS-HCI '08)*. Vol. 1. British Computer Society, pp. 141–150.

Stevens, Molly M, Gregory D Abowd, Khai N Truong, and Florian Vollmer (2003). "Getting into the Living Memory Box: Family Archives & Holistic Design". In: *Personal and Ubiquitous Computing* 7 (3-4), pp. 210–216.

Strötgen, Jannik and Michael Gertz (2010). "HeidelTime: High quality rule-based extraction and normalization of temporal expressions". In: *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, pp. 321–324.

Suh, Bongwon and Benjamin B. Bederson (2007). "Semi-Automatic Photo Annotation Strategies using Event Based Clustering and Clothing Based Person Recognition". In: *Interacting with Computers* 19 (4), pp. 524–544.

Sumi, Yasuyuki, Ryuuki Sakamoto, Keiko Nakao, and Kenji Mase (2002). "ComicDiary: Representing Individual Experiences in a Comics Style". In: *Ubiquitous Computing (UbiComp 2002)*. Springer, pp. 16–32.

Swan, Laurel and Alex S. Taylor (2008). "Photo Displays in the Home". In: *Proceedings of the Seventh ACM Conference on Designing Interactive Systems (DIS2008)*, pp. 261–270.

Takahashi, Kohei, Shinsuke Matsumoto, Sachio Saiki, and Masahide Nakamura (2014). "Exploiting No-SQL DB for Implementing Lifelog Mashup Platform". In: *Soft Computing in Big Data Processing*. Vol. 271. Springer, pp. 39–49.

Tang, Jiliang, Xufei Wang, Huiji Gao, Xia Hu, and Huan Liu (2012). "Enriching short text representation in microblog for clustering". In: *Frontiers of Computer Science in China* 6 (1), pp. 88–101.

The Moving Picture Experts Group (2010). *MPEG-7*. URL: http://mpeg.chiariglione.org/standards/mpeg-7.

The Stanford Natural Language Processing Group (2014). *The Stanford Natural Language Processing Group - Stanford NER CRF FAQ*. URL: http://nlp.stanford.edu/software/crf-faq.shtml#pos (Accessed on: 04/10/2014).

Tian, Yuan, Biao Song, and Eui-Nam Huh (2011). "Towards the Development of Personal Cloud Computing for Mobile Thin-Clients". In: *International Conference on Information Science and Applications (ICISA)*. IEEE, pp. 1–5.

Tossell, Ivor (2009). *On the Web, forever has a due date*. URL: http://www.theglobeandmail.com/technology/on-the-web-forever-brhas-a-due-date/article4287810/.

trendwatching.com (2004). *LIFE CACHING – An emerging consumer trend and related new business ideas*. URL: http://www.trendwatching.com/trends/life_caching.htm.

Twitter (2014). *Developer Display Requirements*. (Accessed on: 03/26/2014).

Twitter (2015). *GET statuses/user_timeline*. URL: https://dev.twitter.com/rest/reference/get/statuses/user_timeline (Accessed on: 03/14/2015).

UzZaman, Naushad, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky (2013). "TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations". In: *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval'13)*, pp. 1–9.

Van House, Nancy A. (2007). "Flickr and public image-sharing: distant closeness and photo exhibition". In: *Human Factors in Computing Systems (CHI'07)*. ACM, pp. 2717–2722.

Verhagen, Marc and James Pustejovsky (2008). "Temporal processing with the TARSQI toolkit". In: *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 189–192.

Wang, Peng (2012). "Semantic Interpretation of Events in Lifelogging". PhD Thesis. Dublin City University. URL: http://doras.dcu.ie/16766/.

Ward, Grady (1996). *Moby Lexicon - Part*. URL: http://icon.shef.ac.uk/Moby/mpos.html.

Weikum, G., J. Hoffart, N. Nakashole, M. Spaniol, F. M. Suchanek, and M. A. Yosef (2012). "Big Data Methods for Computational Linguistics". In: *IEEE Data Engineering Bulletin* 35 (3), pp. 46–55.

Weingarten, Christopher R. (2010). *Would you buy a box set of rock review tweets?* The Guardian. URL: http://www.theguardian.com/music/2010/feb/18/twitter-rock-reviews-box-set.

Wenyin, Liu, Yanfeng Sun, and Hongjiang Zhang (2000). "MiAlbum - a system for home photo management using the semi-automatic image annotation approach". In: *Proceedings of the eighth ACM international conference on Multimedia*. ACM, pp. 479–480.

Westermann, U. and R. Jain (2007). "Toward a Common Event Model for Multimedia Applications". In: *MultiMedia* 14 (1), pp. 19–29.

Williams, Thomas and Colin Kelley (2013). *gnuplot 4.6.* URL: `http://www.gnuplot.info/documentation.html`.

Wilson, C., B. Boe, A. Sala, K.P.N. Puttaswamy, and B.Y. Zhao (2009). "User interactions in social networks and their implications". In: *Proceedings of the 4th ACM European Conference on Computer Systems*, pp. 205–218.

Wolff, Eberhard (2013). "NoSQL: An Architects Perspective". In: *Berlin Expert Days.*

Wu, Xiaoxin, Wei Wang, Ben Lin, and Kai Miao (2009). "Composable IO: A Novel Resource Sharing Platform in Personal Clouds". In: *Cloud Computing.* Springer, pp. 232–242.

Xue, Nianwen (2003). "Chinese Word Segmentation as Character Tagging". In: *Computational Linguistics and Chinese Language Processing* 8 (1), pp. 29–47.

Zhao, Xuan, Niloufar Salehi, Sasha Naranjit, Sara Alwaalan, Stephen Voida, and Dan Cosley (2013). "The many faces of Facebook: Experiencing social media as performance, exhibition, and personal archive". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, pp. 1–10.