HUMAN CENTRED EVENT LINKING

A thesis submitted to the University of Manchester
for the degree of Master of Philosophy
in the Faculty of Engineering and Physical Sciences

2012

By
Paul Waring
School of Computer Science

Contents

Total word count (including footnotes but excluding bibliography): 32786.

List of Tables

List of Figures

Abstract

The World Wide Web contains a vast corpus of information describing a variety of events, both current and historical. However, this information is not as well connected as it could be, with many stories standing on their own without links to related events. Previous research suggests that generating these links could increase the scope for users to serendipitously discover events related to the story which they are currently reading. Existing approaches to this problem have focused on connecting events within small corpora containing only a few hundred documents, and have been more concerned with treating the problem as a challenge in information retrieval, with precision and recall being the measures of success. Our approach differs by tackling this problem in the human factors domain, with user perception as an indicator of success, as opposed to strictly technical measures.

The Human Centred Event Linking (HuCEL) project provides a solution to this problem which is lightweight, extendable and builds upon existing work. By extracting keywords from the main content of a news Web page, we are able to automatically generate search queries to scan the Web for related events, and display this additional information to users next to the original story. A technical evaluation indicates that users find our queries to be related to the story, demonstrating that our algorithm is producing quality keywords. A qualitative and quantitative user study of the links generated by the HuCEL platform also demonstrates that users find these associations to be related to the story under discussion.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I gratefully acknowledge the help, support and guidance of my supervisors, Simon Harper and Andrew Brown, and my advisor, Carole Goble. I would also like to thank Yeliz Yesilada and Darren Lunn for their assistance with integrating the COHSE and SADIe projects respectively.

This work was funded by a Doctoral Training Award from the Engineering and Physical Sciences Research Council (EPSRC).

Chapter 1

Introduction

The World Wide Web contains a vast corpus of information describing a variety of events, both current and historical. However, this information is not as well connected as it could be, with many stories standing alone without links to related events. Existing approaches to the problem of connecting events have focused on forming associations within small corpora, often containing only a few hundred documents. In addition to this, previous approaches have treated the problem of associations between events as a challenge within the information retrieval domain, where precision and recall are considered the measures of success. We use the following definitions for precision and recall throughout the thesis (Baeza-Yates & Ribeiro-Neto 1999, p.75):

Precision: Out of the set of all retrieved documents, the percentage which are relevant. A precision of 100% (or 1.00) indicates that all retrieved documents are relevant.

Recall: Out of the set of all relevant the documents, the percentage which have been retrieved. A recall of 100% (or 1.00) indicates that all relevant documents have been retrieved.

Whilst this approach is a valid one to take, we suggest that by focusing on improving the precision and recall of retrieval algorithms, the existing research ignores what is arguably the most important factor: the user's perception of the results obtained.

Human Centred Event Linking (HuCEL) is a project which aims to aid users in their navigation of the vast information space on the Web, and stimulate

opportunities for serendipitous discovery of related events. This chapter presents an introduction to our work, describing our motivations, contributions and an outline of the remainder of the thesis.

1.1 Problem Statement

Whilst there is a significant amount of information describing events on the Web, much of this is not as well connected as it could be, and therefore opportunities for users to discover information about related events are lost. By creating links between these pages, we can assist users in navigating the complex information space of the Web. In order to perform this linking, we have developed an algorithm for extracting keywords from the main content of news pages. This process is acknowledged as a difficult task that can be fraught with errors, particularly in domains where ungrammatical language is used (Halpin & Moore 2006). The unstructured and inconsistent nature of the Web also presents difficulties for extracting and analysing this information (Dill et al. 2003).

1.2 Motivation

Our principal motivation for undertaking this research is to address the issue of links between related events from a human factors perspective. We suggest that by approaching the problem from this angle, as opposed to the measures of precision and recall, we will increase our understanding of how events can be connected using links. Furthermore, the ultimate test of any associations is not whether they meet specific information retrieval measures such as precision and recall, but whether users find them to be useful when browsing, as 'only a human can make the final determination of whether or not retrieved items are indeed relevant' (Hayes & Dekhtyar 2005). This motivation is perhaps best summed up by Saracevic (1997):

> The success or failure of any interactive system and technology is contingent on the extent to which user issues, the human factors, are addressed right from the beginning to the very end, right from theory, conceptualization, and design process on to development, evaluation, and to provision of services.

1.3 Solution

Despite the unstructured and inconsistent nature of the Web, we demonstrate that it is possible to apply an algorithm which generates a set of links to related events from a given Web page from a news site. In order to evaluate our solution to this problem, we have performed two experiments. The first was of a technical nature, with the aim of assessing whether our algorithm produces keywords which users perceive as being related to the event described on the page. The second experiment built on this, by providing participants with links generated by user keywords and our algorithm, which they were asked to rate in terms of their relatedness to the story.

1.4 Contributions

The principal contribution of the HuCEL project is an algorithm which extracts keywords from the main content of a Web page on a news site. These attributes can then be used to automatically generate a search query, which returns links to pages describing related events. In addition, we have developed a method for retrieving and displaying links to these related stories to users alongside the original story.

Through the analysis and evaluation of data produced from our technical experiment, we have demonstrated that users find our keywords to be related to the story in question. This suggests that our algorithm is producing high quality keywords from a human-centred perspective, even if such results might be considered to be of low quality if traditional information retrieval measures were applied to them.

To back up our technical evaluation, we also performed a user study of the links generated by the HuCEL platform. This process enabled us to obtain qualitative and quantitative feedback relating to the results of the queries produced by our algorithm. The results show that users find these associations to be related to the story under discussion in the majority of cases. Furthermore, the links generated by our algorithm generally equalled or outperformed the results obtained using the keywords supplied by participants in the technical experiment. These results suggest a mismatch in user preconceptions of keywords and results, with user keywords generating links to the same event and our keywords producing

links to related events.

In the process of our research, we have also created a lightweight and extendable tool for automatically generating links between events. This utilises our algorithm for identifying key words and phrases which can be used to locate related events, as well as building upon existing research such as the COHSE (Carr et al. 2001) and SADIe[1] projects. In order to facilitate further research in this area, this tool is freely available from an online repository, along with a technical report detailing its functionality and use (Waring 2009*a*).

1.5 Thesis Organisation

The remainder of this thesis is organised into the following chapters.

Chapter 2: Background and Related Work discusses the existing literature in the areas of event detection, user browsing strategies and dynamic linking. In particular, we can see that previous research has largely focused on improving algorithms for identifying events in the information retrieval domain, with evaluations based upon measurements such as precision and recall. Furthermore, a significant amount of this related work has been applied to fixed corpora consisting of only a few hundred documents. We conclude that limited work has been undertaken to address the problem of associating events in the human factors domain by evaluating user perceptions of the results obtained from such research – this being the area in which we intend to situate our work.

Chapter 3: Understanding Events discusses the development of our algorithm for extracting keywords from a news Web page and producing links to related events. The algorithm isolates proper nouns within the text of a story, using a combination of techniques including part-of-speech tagging, lexicons and word expansion based on previous occurrences within the text. This provides us with a set of keywords which we can automatically feed into a search engine to obtain links to related events. Initial tests on sample pages show that selecting proper nouns, particularly entity names and locations, produces search queries which return links to related events. Based

[1] `http://hcw.cs.manchester.ac.uk/research/sadie/`

on these results, we conclude that these keywords and phrases are important when locating related events.

Chapter 4: Methodology and Results discusses our methodical approach to obtaining data relating to the utility of our work. We demonstrate how, even with the undue influence of memory bias, users still rate our keywords as relevant to the story in question, and in some cases display a marked preference for our keywords. We can also clearly see how users find the links produced by our queries to be related to the original story. Finally, even though users often state that they prefer their keywords in our first and second experiments, the results of our third experiment clearly demonstrate that our algorithm actually produces more related links.

Chapter 5: Analysis and Discussion presents a detailed analysis of the results from our experiments. The analysis of keyword preferences shows that users display a weaker preference for their keywords than we would expect, suggesting that our keywords are making users think again about what they entered. Furthermore, results which indicate that users prefer the results produced by our algorithm suggest that user keywords are returning links to the same event, whilst our results are focused on links to related events.

Chapter 6: Conclusions and Future Work discusses our overall conclusions and presents avenues for future work which build upon the research presented in this thesis. Our primary conclusion is that there is a perception mismatch between keywords which users feel will generate links to related events and the results which they prefer, and that an algorithm can produce better keywords than the majority of users. Finally, we propose several possible directions in which future work in this area could take, including the incorporation of temporal information in event detection and analysing the links to ensure that the most related ones are returned.

1.6 Technical Reports

In addition to the thesis, we have produced a number of technical reports containing the data from our experiments and the code for our algorithm described in Chapter 3.

HuCEL: Keywords Experiment Manual describes how to re-run our keywords experiment, as well as providing the data obtained from the experiment (Waring 2009*c*).

HuCEL: Keywords Experiment II Manual describes how to re-run our second keywords experiment, and contains all the data obtained from the experiment (Waring 2009*b*).

HuCEL: Links Experiment Manual describes how to re-run our links experiment, and contains all the data obtained from the experiment (Waring 2009*d*).

Deploying the HuCEL Java Servlet describes how to deploy the Java servlet which contains the complete HuCEL platform, including our event parser (Waring 2009*a*).

Chapter 2

Background and Related Work

In this chapter we discuss the existing literature which is closely related to our work, including the areas of event detection, user browsing strategies and dynamic linking. In particular, we demonstrate that previous research has largely focused on improving algorithms for identifying events, through the information retrieval domain, with evaluations based upon technical measurements such as precision and recall. Furthermore, a significant amount of this related work has been applied to fixed corpora consisting of only a few hundred documents. We conclude that limited work has been undertaken to address the problem of associating events in the human factors domain by evaluating user perceptions of the results obtained from such research – this being the area in which we intend to situate our work.

2.1 What is an Event?

Although there is a growing corpus of research related to the identification and tracking of events, the word 'event' in itself is often used without a formal definition of what is meant by this term. For example, Brants & Chen (2003) describe 'new event detection' as 'the task of detecting stories about previously unseen events in a stream of news stories', but fail to provide any definition of what an event is, bar a few unconnected examples.[1] Petras et al. (2006) also do not define the term, either in their own words or by reference to an existing definition, despite discussing 'placing events in temporal and geographic context'. As Makkonen et al. (2003) point out, whilst thinking about what an event is appears

[1] 'e.g. an airplane crash, and earthquake, governmental elections, etc.' (Brants & Chen 2003)

to be intuitive, 'it is difficult to establish a solid definition.' Nevertheless, defining what is meant by 'an event' is an important task if work is to be undertaken on identifying and linking events, and some researchers and projects have attempted to tackle this problem.

One of the earliest works in the area of event tracking suggests that 'a possible definition of event is something that happens at a particular time and place' – in other words, an event has both a spatial and temporal attribute, and both of these attributes are clearly defined (Allan et al. 1998). This definition appears to have been accepted by a number of other researchers (Makkonen & Ahonen-Myka 2003, Li et al. 2005, Zhang et al. 2007), who specifically refer to it as the definition of 'an event' rather than suggesting their own alternative. Some examples of cases which might not be considered events under this definition have been highlighted, such as those which continue over a long period of time (Makkonen et al. 2002). However, such cases could possibly be broken up into smaller parts, each of which would constitute an individual event on its own, so this is not necessarily a problem.

Building on this definition, it has been stated that 'the specific location and time of an event differentiate it from broader classes of events', suggesting that these attributes are the means by which any given event can be uniquely identified (Allan et al. 1998). In other words, an event is different to another event if at least one of these attributes differs, and conversely two events are identical if they involve the same 'something that happens' at the same place and time. A simple example of this can be seen in the eruptions of Mount Vesuvius – the same thing happens in both cases (a volcanic eruption) and the spatial location (the Bay of Naples) is the same, but each of these events can be distinguished by their temporal attribute, AD 79 and AD 1631. Other researchers agree with this suggestion, stating that for two different events involving the same occurence 'it would seem that the location and the time ... are the terms that make up the difference' (Makkonen et al. 2002).

In their study of retrospective and on-line event detection, Yang et al. (1998) present a similar definition to Allan et al. (1998) when stating that 'the only guideline explicitly given [to researchers] for event definition was that an event should identify *something (non-trivial) happening in a certain place at a certain time*' (emphasis original). This definition is reiterated in Yang et al. (1999) and accepted by several other scholars (Smith 2002, Nallapati et al. 2004, Feng &

Allan 2007, Zhang et al. 2007).

From a broader and less technical point of view of what people think of as representing an event, Scholes (1980) offers us the suggestion that 'a real event is something that happens: a happening, an occurence, an event.' Whilst perhaps not the most useful of definitions in itself, Scholes does go on to suggest that 'a narrated event is the symbolization of a real event: a temporal icon', indicating that time is an important aspect of an event – though in this case a specific type. Furthermore, Scholes suggests that 'a narration is the symbolic presentation of a sequence of events connected by subject matter and related by time. Without temporal relation we have only a list.' This would suggest that the temporal aspect is not only part of each individual event, but it is a way by which different events can be linked – in this case a sequence of events which occur after one another.

Again from a non-technical perspective, Fogelson (1989) offers the definition of an event as 'in simplest terms, an event can be defined as that which occurs at a given time and place'. As with previous definitions, the temporal aspect is mentioned, along with a geographical attribute – an alternative way of wording the definition given by Allan et al. (1998). Fogelson also states that 'events are also considered to have properties and relationships', though unfortunately he provides no indication as to what these properties and relationships might be. However, if an event is considered to have properties and relationships these could be used to link events, in that events which have the same value for a given property could be considered to have a relationship based on that particular attribute.

Continuing on the theme of emphasising the importance of time when defining an event, Makkonen & Ahonen-Myka (2003) state that 'clearly, an event as well as the news-stream itself are intrinsically sensitive to time.' Furthermore, whilst the authors make it clear that detection and tracking of events should not rely solely on time-based information, the various expressions of time within a given document appear to be useful when organising documents mentioning events into particular topics.

Although agreeing on the importance of the temporal aspect of events, Vendler (1967, p.141) attempts to separate the concept of an event from the concept of an object, arguing that:

> Fires and blizzards, unlike tables, crystals, or cows, can occur, begin,

> and end, can be sudden or prolonged, can be watched and observed – they are, in a word, events and not objects.

The justification Vendler offers for making this distinction is that objects exist in space but not in time, as they cannot be said to occur, begin or end (Vendler 1967, p.143). Events, on the other hand, are 'primarily temporal entities' and thus do not exist in space and cannot be said to occur in a particular location (Vendler 1967, p.144). Rattenbury et al. (2007) appear to agree with this definition, as they define 'event' and 'place' as two different things – specifically event tags exhibit 'significant temporal patterns' and place tags exhibit 'significant spatial patterns'. Whilst this is perhaps a fair distinction to make, all events involve objects of some kind – an occurence cannot just 'happen', it must happen *to* something or someone, and likewise an object cannot really be said to exist outside of time. Whilst events might arguably lack a spatial attribute if we consider them separately from objects, if we take into account the binding relationship between these two concepts then an event can be said to be more than just a temporal entity. Zacks & Tversky (2001) provide an excellent evaluation of this point from the psychological perspective of how people perceive events and objects, suggesting that 'one can reasonably argue for treating events as one treats objects', given that 'objects have boundaries in space' and 'events have boundaries in time'.

In a slightly different vein, Smith (2002) suggests that 'for narrative documents, questions of "what happened?", "where?" and "when?" are natural points of entry'. Again, this clearly defines an event as something which happens at a particular place and time, and Smith suggests that this enables users 'to browse document collections by the common and well-understood dimensions of time and space.' Smith stands out in that this particular piece of work concentrates more on identifying the geographical locations of events and is less interested in the temporal aspect, which is a marked difference to all the literature surveyed so far. Indeed, Smith specifically draws attention to this, stating that 'despite the definition of an event, however, as occuring in a certain place, most TDT[2] systems do not directly take geographical location into account.'

[2]Topic Detection and Tracking (TDT) was a research project pursued under the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) programme. The project has now ended, and results can be viewed online at: `http://www.nist.gov/speech/tests/tdt/`

When examining the work to date, one point which all of the literature examined agrees on is that time is an important aspect of defining an event. Indeed, several authors see time as the most important aspect of an event (Vendler 1967, Scholes 1980). In addition, many scholars also mention the location of an event as being part of its definition. Therefore, drawing together all of the definitions given so far, the following definition of an event would appear to broadly represent all of these views: something which happens, at a given place and time. In other words, *what*, *where* and *when*. However, one obvious element which is missing from this definition is the question of *who* was involved in the event. Although events can occur without people present,[3] for the purposes of human factors research the primary area of interest lies in the experiences of people, including what events mean to them and how the same events are reported by different people. Whilst people do write, often at length, about events which did not involve any human participants, they cannot be said to have experienced those events, or even to be drawing upon the experiences of people who were present.

However, there are a number of papers that allude to or incorporate the concept of people being involved with an event. Allan (2002, p.2) states that a particular event occurs 'not only at some particular time, but in a specific location, and usually with an identifiable set of particupants.' This definition is refined later to 'an event is something that has a specific time, location and people associated with it.' (Allan 2002, p.13). Wei & Lee (2004) agree, declaring that a news story generally reports event properties including 'when the event occured, who was involved, where it took place'.

Nakahira et al. (2007) also mention the importance of the people who are involved in an event, defining a historical event 'by five elements: person, cause, object, location and time.' Whilst cause is not of immediate interest to our work, the other four elements are a useful indicator of the attributes which we may consider an event to have. Lavrenko et al. (2002) are in agreement, stating that an event 'occurs in a specific place and time, with specific people involved.'

Makkonen et al. (2002) also include the concept of people within their definition of an event, stating that a report of an event should include at least '*what* happened, *where* it happened, *when* it happened, and *who* was involved' (emphasis original). Furthermore, they suggest that each of these attributes can be

[3]An example of this would be the Big Bang, which most people would probably consider to be a major scientific and historical 'event'.

represented as a semantic class, specifically *names* (of people), *temporals* (expressions of time), *locations* and *terms* (nouns and adjectives which do not fit into any of the other classes). These semantic classes may well represent the properties alluded to by Fogelson (1989).

Bringing together all the literature surveyed, we can suggest that an event might be best described as:

1. Something which happened.
2. The place where it happened.
3. The time when it happened.
4. The set of individuals involved.

This is our proposed definition of an event which we shall be using for the rest of this work.

2.1.1 Types of Event Descriptions

Broadly speaking, descriptions of events on the Web can be divided into two types, *structured* and *unstructured*. Structured descriptions are those which, perhaps unsurprisingly, have a well-defined and labelled structure. For example, all events on Upcoming[4] have a time period, location and description – a standard structure which is applied across the site. Once we are aware of this structure, extracting information about events from this particular source is a trivial task. On the other hand, unstructured descriptions consist of natural language text, with no consistency as to where event attributes (e.g. locations) appear. Unstructured descriptions of events can be ambiguous, especially when taken out of context (Feng & Allan 2007).

In particular, temporal expressions can be difficult to parse – for example 'last Tuesday was the State Opening of Parliament' relies on the context of knowing when this statement was made in order to change the relative date ('last Tuesday') into a absolute date (e.g. 'Tuesday, 4 December 2007') (Makkonen et al. 2002, Makkonen et al. 2003). Relative dates within a single time frame, whether this be an entire document or a section thereof, are not necessarily an issue, but as soon as we start to compare dates across different documents we need to resolve

[4] `http://www.upcoming.org/`

relative dates to their absolute equivalents (Mani & Wilson 2000). Unfortunately, few temporal expressions within text are fully specified and therefore need to be computed from the context of the surrounding text (Dale & Mazur 2007). In addition, a pilot experiment found that only 25% of clauses examined contained explicit time expressions and as a result, only using explicit times would be not be sufficient for anchoring events (Mani et al. 2003). Hobbs & Pan (2004) also demonstrate how difficult temporal arithmetic can be, including working with temporal descriptions which rely on context in order to be fully understood.

2.1.2 Topics, Stories and Events

Although the majority of the literature refers to events consistently, there are several papers which refer to topics as well as, or instead of, events. These papers are usually involved with the Topic Detection and Tracking (TDT) project, so this use of the word is perhaps understandable, but there is a difference between how researchers view the two terms, with Kumaran & Allan (2004) and Yang et al. (2000) being the most obvious contrasting opinions. Kumaran & Allan (2004) use 'topic' and 'event' as synomyms, suggesting that 'an example of a topic could be the sinking of an oil tanker.' Furthermore, the authors suggest that 'every time a new topic was found and tracked by a topic tracking system, it was equivalent to finding a new event.' On the other hand, Yang et al. (2000) are careful to draw a line between the definition of an event and that of a topic, with the distinction being that an event is 'localized in space and time' and 'typically short in duration'. Using this definition, the sinking of a specific oil tanker would be classed as an event, whereas the general category of 'accidents at sea' would be classed as a topic. In general, the distrinction drawn by Yang et al. (2000) appears to be supported more by other scholars than the method of treating topics and events as synonyms.

Throughout the thesis we use 'event' when referring to the work on our algorithm and parser, and 'story' when referring to a news item describing an event – i.e. 'event' and 'story' are interchangeable in this context. We do not discuss the notion of a 'topic' in its use as a 'category of events' beyond this section.

2.2 Connecting Events

Once a number of events have been identified, we can begin to connect them together based on the four attributes which define an event – i.e. the event itself, the location, the time and the people involved. There are two types of connection which we will consider – *clustering* and *linking* – and they are outlined in the following pages.

2.2.1 Clustering

Liu (2005, p.118) defines clustering as 'the process of organizing data instances into groups [i.e. clusters] whose members are similar in some way.' How similarity is defined and to what degree it is applied varies from application to application. In some instances, clustering may only place identical items into the same cluster,[5] whereas in other instances a clustering algorithm may require only one out of many possible attributes in common in order to class two data items as being 'similar'. Following on from this, the fact that items in the same cluster have a certain degree of similarity implies that items in different clusters have a degree of dissimilarity, a feature which can be useful in certain situations.[6]

In the TDT programme, clustering is viewed as an extension of new event detection. Each story in a news stream is processed to determine whether or not it discusses a topic which has not been seen previously. If a story discusses a topic which has already been encountered previously, it is placed in an existing 'bin' (i.e. a cluster) with all other stories discussing the same topic, and if the story relates to a topic which has not been seen before a new bin is created for that topic (Allan et al. 2005). Lam et al. (2001) follow a similar approach, noting that 'a story is inserted to the most similar cluster or formed as a seed of a new cluster.'

Clustering in general is a problem which has been the source of much attention in the past, and the field can be said to be well studied (Chakrabarti et al. 2006). We will not be aiming to make significant contributions to this area, rather we shall be using the existing techniques to further the unique aspects of our work.

[5] One practical use of this would be to remove duplicate Web pages from search results, by only returning one result from a cluster of identical documents.

[6] For example, if we wish to separate documents which refer to different events (Smith 2002).

2.2.2 Linking

Some confusion can arise with the use of the word 'linking', as it has several meanings. The TDT project has an evaluation task, 'link detection', which 'requires determining whether or not two randomly selected stories discuss the same topic' (Lavrenko et al. 2002). This is also different to our work on linking, as we will be looking to connect events which we consider to be related, but which may not be part of the same topic. Whilst it may be the case that some of the events we link together based on commonality in attributes also happen to be part of the same topic, we are not specifically aiming to link events based on topic membership. Instead, our aim is to create connections (expressed as hyperlinks) between events based on common values for the four event attributes which we have mentioned previously – i.e. *what* happens, *where* it happens, *when* it happens and *who* it happens to.

The issue of linking events has been discussed previously in Feng & Allan (2007), though under the title of 'event threading'. Whilst the underlying concept is similar, their approach is from an information retrieval viewpoint, with the aim of finding the most efficient and precise method for extracting and linking event information from news stories. Our approach differs from this in two ways. First, we shall be approaching the problem from a human-centred perspective, with the aim of presenting event-related information on the Web in a structured way which is easier for users to understand than the current unstructured mass of text which exists. Second, we are not confining the problem to a small fixed corpus, but will instead be taking into account the large and constantly changing set of pages of which the Web is composed.

The advantage linking provides over clustering is that it enables the serendipitious discovery of related events, whereas clustering only allows us the discovery of similar descriptions of the same event. For example, the run on Northern Rock in September 2007, where thousands of savers withdrew their deposits from the bank, was widely reported in the news. Around the same time, the Nationwide building society saw a surge in its deposits, caused largely by people who had previously held Northern Rock accounts looking for a 'safer' place to deposit their money. By using clustering techniques on that day's news, we would be able to see several news outlets reporting the run on Northern Rock, but the stories reporting the surge in deposits at Nationwide would be overlooked as they do not

discuss the same event.[7] However, linking *would* pick up this relation – i.e. that both events involve the same people, 'Northern Rock savers'. We would suggest that anyone interested in the mass withdrawals from Northern Rock would also be interested in where those deposits were going, and so by displaying the stories relating to Nationwide, we can provide further related information for the user.

2.2.3 Gaps in Previous Work

Whilst some research does exist on the subject of connecting events, the literature which has been examined so far focuses on identifying the same event in a number of different news stories (i.e. clustering), whilst there appears to be no current work being conducted into how to link different events to one another or discover new events based on knowledge of events which we are already aware of. Most of this clustering work comes from the TDT project, where the general aim appears to be the identification of events within news stories, followed by the clustering together of news stories which mention the same event under the umbrella of a 'topic'.

2.3 Browsing Hypertext

A fundamental part of hypermedia and the Web, which is regularly engaged in by users, is the concept of browsing through documents to obtain information (Carmel et al. 1992, Yesilada et al. 2007). Browsing is often differentiated from searching on the basis that searching assumes the user knows what she is looking for, or at least is aware of a number of keywords which are likely to be contained in documents of interest and can therefore be combined into a query to be performed on a corpus of data. For example, searching, 'looking for a known target', can be contrasted with browsing, 'looking to see what is available in the world' (Jul & Furnas 1997). The difference between these two concepts can also be defined as *finding* (i.e. searching), 'using the Web to find something specific', and *browsing*, which involves 'having no specific goal in mind' (Sellen et al. 2002). Alternatively, browsing can be described as 'the art of not knowing what one wants until one finds it' (Cove & Walsh 1988), as opposed to searching, where the goal is known

[7] A broader application of clustering on 'bank events' might pick up this relation, but would also result in a larger cluster of documents. Such an application would also miss the relation if the withdrawn deposits were used in some other way – e.g. invested in the stock market.

beforehand. However, whilst there are differences between the two techniques, searching and browsing are not mutually exclusive – the two methods may be considered complementary (Jul & Furnas 1997), and both are often employed in the user's quest for the information she seeks (Catledge & Pitkow 1995). Even when users are aware of their information needs, keyword searching is not necessarily the preferred method of obtaining information. For example, a study conducted by Teevan et al. (2004) suggested that only 39% of user queries involved keyword searches.

Browsing can also be divided into smaller sub-categories, such as the ones suggested by Cove & Walsh (1988)[8] and accepted by various other scholars (Carmel et al. 1992, Catledge & Pitkow 1995, Yu & Roh 2002, Thakor et al. 2004), which are:

1. *Search browsing:* Where the goal is already known before browsing begins – this is similar to the broad topic of 'searching'.

2. *General purpose browsing:* The regular consultation of several sources based on the assumption and likelihood that these sources contain information which the user is seeking.

3. *Serendipity browsing:* A 'purely random, unstructured, and undirected activity.'

For our purposes, the final category holds the most interest, as we intend to present the user with dynamically generated links which she can then explore to serendipitously discover new pages of interest.

In addition to how users browse, there is also the question of what the user is looking for. Gibson (2004) suggests three possibilities for the user's task when browsing:

1. Navigating to previously visited pages.

2. Discovering new pages.

3. Assessing information for possible later use.

[8]Similar sub-categories, described as *patterns*, can be found in Salomon (1990).

Although knowledge workers appear to split their time equally amongst all three tasks,[9] for our work we will be concentrating on the second of these three tasks, as we are looking to help users serendipitously discover new pages which they would otherwise not have encountered.

2.3.1 Serendipitous Browsing

Of the three main types of browsing suggested in Cove & Walsh (1988), we are concentrating primarily on serendipitous browsing. The benefits of serendipitous browsing have already been demonstrated in other domains. In particular, the way books are organised on shelves using the Dewey Decimal or Library of Congress systems has been shown to stimulate serendipitous browsing, by placing books on similar subjects next to one another (Liestman 1992). For example, research undertaken by Stelmaszewska & Blandford (2004) concluded that 'most of the subjects experienced serendipity during their library session and for them it was a natural way to find material.' Serendipity can be considered to be an essential aid to the process of discovery across disciplines, both in the humanities (Delgadillo & Lynch 1999) and the sciences (Foster & Ford 2003).

In our work, we will be focusing on generating links to pages which describe events related to the story that the user is currently reading. By presenting these links next to the current story, we aim to provide opportunities for serendipitious browsing in a similar way to how libraries organise books on shelves.

2.4 Dynamic Link Generation

Generally speaking, the nature by which links are created can be split into three distinct types (Ashman et al. 1997):

1. Links individually created by a user (*hand-made links*).
2. Links automatically created ahead of time as the result of a computation (*pre-computed links*).
3. Links created when needed as the results of a computation (*dynamically computed links* or *dynamic links*).

[9] A recent study found that partipants split the majority of their activity on the Web between information gathering (35%), finding (24%) and browsing (27%) (Sellen et al. 2002).

In order to connect pages which contain related events, we will need to generate links from the page which the user is currently viewing to other pages which contain information about related events. This linking will be performed dynamically, based on the event attributes which have been extracted from the page. The following reasons have been suggested as to why dynamic link generation may be desirable (Yan et al. 1996):

1. Links can be customised for an individual user, based on the content in which she has expressed an interest so far.

2. Due to the continuous changes to the content of a Web site, dynamic linking can provide more up to date information than a static set of links.

3. As the number of categories and amount of content increases, it becomes more and more difficult for a designer to offer static links.

The first benefit is of less interest to our work than the other two, as we will be dynamically generating links based on the content of the page – more specifically, the events mentioned within the content – rather than previous interests shown by the user. In other words, we are assuming that the user will be interested in events related to those under discussion on the current page. However, both the second and third benefits are relevant to our work, although we will be examining the Web as a whole as opposed to focusing on individual sites. We can therefore represent these two benefits in the following modified ways.

Firstly, because the content of the Web in general changes continuously,[10] dynamic links are the only feasible way to connect pages which mention related events. Attempting to manually maintain large collections of links is both expensive and inefficent (El-Beltagy et al. 2001), as it requires a significant amount of human intervention (Dalal et al. 2004). Some scholars have gone as far as to suggest that the maintenance issue of links means that webmasters should 'link sparingly, if at all' (Lynch & Horton 1997), though we consider this to be an overreaction to the problem. Furthermore, the likelihood of any given URL being available decays over time, and the lifetime of any given URL is limited, with URLs having a *half-life*[11] between nine months and one year (Fetterly et al. 2003, Bar-Yossef et al. 2004, Dalal et al. 2004, Ntoulas et al. 2004). Even

[10] A study by Cho & Garcia-Molina (2000) found that 40% of pages on over 200 popular sites changed on a weekly basis.

[11] The average time it takes for 50% of pages to become unavailable.

in the area of scientific research and publications, where we might expect additional effort to be put into ensuring Web references are persistent, surveys have shown that 10-20% of URL citations are unavailable one year after their creation (Markwell & Brooks 2002, Dellavalle et al. 2003, Spinellis 2003, Wren 2004).

Secondly, as the number of pages on the Web grows, connecting related events becomes a task which is increasingly difficult to perform manually (Wilkinson & Smeaton 1999), which is a possible reason for why so few sites do so at present.[12]

In addition to these benefits, dynamic link generation has also been shown to significantly reduce the amount of time required by users to perform a specific task. In a study conducted by El-Beltagy et al. (2001), users were asked to answer a given set of questions on a particular topic, first by using only a search engine and then with the addition of dynamically generated links to sites containing similar content. The linking facility reduced the amount of time taken to complete the task by 28% in one case and 55% in another, demonstrating that the addition of such links can have significant benefits for users. Whilst dynamic links do have some disadvantages, such as links appearing in the wrong context due to inexact specifications, and their rapidly changing nature which can cause links to be as volatile as the data to which they point (Ashman 2000), it could be argued that these are more than offset by their advantages.

Furthermore, several studies have demonstrated that following links is by far the most common way by which users navigate to new pages, and this has consistently been the case over the ten year period which separates the earliest and latest studies (Catledge & Pitkow 1995, Tauscher & Greenberg 1997, Weinreich et al. 2008).[13] As a result, hyperlinks are considered 'an essential, if not the most important feature of the World Wide Web' (Bry & Eckert 2005), and we therefore suggest that presenting related events in the form of links to the pages which discuss them is a sensible method to use.

2.4.1 Existing Similar Sites Functionality

Two of the most popular search engines, Google and Yahoo!, offer 'similar sites' functionality as part of their advanced search options. The exact criteria are not revealed, but Google's help pages suggest that sites with similar content or which

[12]Over a period of six years, the size of the publicly indexable Web is estimated to have grown to 300 million (Lawrence & Giles 1998), 800 million (Lawrence & Giles 1999), and 11.5 billion (Gulli & Signorini 2005) pages.

[13]Actual studies took place in 1994, 1995/6 and 2004/5 respectively.

would be found using similar sets of keywords would be considered related for this purpose.[14]

Although this functionality already exists, our work differs in that we will be concentrating only on one specific type of information – content relating to events – whereas the similar sites feature of Google and Yahoo! takes into account the whole content of the page, including any inbound and outbound links. The functionality offered by the search engines does not always identify related pages, for example searching Google for pages which are related to a news story about super-casinos,[15] the results returned include the BBC News politics section, the UNISON home page (even though UNISON is not mentioned anywhere in the text of the news story) and a story about a prominent politician being caught smoking on a train. Whilst these pages may be related based on some metric used by Google, they are clearly not related based on the events discussed in the original news story, so our work will differ from this.

In addition to the search engine functionality, BBC News also offers a feature which allows visitors to view a list of links to other news sites which are covering the same story. This is only available on selected pages, and some of the external pages linked to require registration in order to read the news story. Although this system aims to 'identify content from other news websites that relates to a particular BBC story', the relation is based on matching text, and therefore discovers the same story being reported on different news sites.[16] This is different to our work, in which we are aiming to provide visitors with links to related, but not identical stories, which could include similar stories on the same news site.

2.5 Conclusions

Having examined the existing literature in the areas of event detection, user browsing strategies and dynamic linking, we can see that there is already a substantial corpus of research. However, we have identified several gaps within the existing literature where we believe areas for research have been overlooked or neglected. These gaps are examined in more detail in the following sections.

[14] *Google Web Search Features*, `http://www.google.com/intl/en/help/features.html` (Accessed: 2 March 2008).

[15] *Super-casino proposal is ditched*, `http://news.bbc.co.uk/1/hi/uk_politics/7264143.stm` (Accessed: 28 February 2008).

[16] *BBC links to other news sites*, `http://news.bbc.co.uk/1/hi/help/3676692.stm` (Accessed: 30 October 2008).

2.5.1 Focus on Fixed Corpora

Much of the existing research in the area of connecting events has focused on small fixed corpora of text, usually consisting of several hundred news stories. Whilst some successes have been achieved in this area, a fixed corpus of news stories is not representative of the large and hetereogenous collection of information on the Web for two reasons. Firstly, a small corpus of news stories is unlikely to be representative in content of the estimated 11.5 billion publically indexable pages available on the Web (Gulli & Signorini 2005). Secondly, we know from previous work that content on the Web changes continuously (Cho & Garcia-Molina 2000), so any fixed corpus is likely to be out of date within a few months at most, and possibly much sooner if it deals with time-sensitive content such as news stories.

2.5.2 Focus on Precision and Recall

As mentioned previously, a significant amount of the existing research is evaluated on the technical measures of precision and recall, with the aim of generating algorithms with slightly higher values for one or both of these measures under a given set of conditions (e.g. Feng & Allan (2007)). In particular, this focus on precision and recall has meant that the human factors perspective has largely been ignored, despite the importance placed on this issue (Saracevic 1997, Ford 2000, Hayes & Dekhtyar 2005).

In conclusion, we can see that there are several areas which have not been provided for by the existing literature. We will fill some of these gaps with our work. The next chapter, *Understanding Events*, outlines how we will address these gaps and discusses the development of our algorithm for doing so.

Chapter 3

Understanding Events

In this chapter we outline solutions to the gaps in existing research highlighted in the previous chapter, whilst providing the rationale for the novelty of our work. We also discuss the development of our algorithm for extracting keywords from a news Web page and producing links to related events. The algorithm isolates proper nouns within the text of a story, using a combination of techniques including part-of-speech tagging, lexicons and word expansion based on previous occurrences within the text. This provides us with a set of keywords which we can feed into a search engine to obtain links to related events.

3.1 Filling the Gaps in Previous Work

In Chapter 2, we examined the existing literature in the areas of event detection, user browsing strategies and dynamic linking. Whilst a wide variety of work has already been undertaken in these areas, we have identified two main gaps in the literature where we believe areas for research have been overlooked or neglected. Our solutions to these gaps are described in more detail in the following sections.

3.1.1 Dynamic Generation of Links

Much of the existing research in the area of connecting events has focused on small fixed corpora of text, often consisting of only a few hundred news stories. However, by utilising existing tools in a novel way we can create a dynamic solution which overcomes the problems inherent with a fixed corpus of information. We create links on-the-fly each time the user loads a page, ensuring that the links

presented are always as up to date as possible and overcoming the problem of pages changing continuously over time.

3.1.2 Focus on Human Factors

Until now, the majority of existing research has focused on treating the challenge of linking events as a problem in the information retrieval domain, with an overall aim of generating algorithms with slightly improved values for precision or recall under a given set of conditions, such as a fixed corpus of text – e.g. Feng & Allan (2007). However, as explained in Chapter 1, the human factors perspective has been largely ignored, despite previous research drawing attention to the importance of this issue (Saracevic 1997, Ford 2000, Hayes & Dekhtyar 2005).

In order to address this problem, we have placed the user at the centre of our work, and measure our success by the user's perception of whether the events discussed in the story are related. In order to test this theory, we have also designed and performed two experiments with human participants – as opposed to the automated tests used with a fixed corpus – to discover whether users find the results produced by our algorithm to be useful. The results of these experiments are presented and discussed in Chapters 4 and 5.

3.2 Natural Language Processing

As our research required the processing of unstructured text, we considered the possibility of using an existing Natural Language Processing[1] solution to assist in the extraction of keywords from the content of news stories and to take account of context to resolve any ambiguities. However, this possibility was eventually abandoned in favour of writing our own parser for a number of reasons, which we explain below.

Firstly, there are numerous NLP solutions available, from generic implementations to those which focus on a narrow domain. A small selection of such toolkits is shown in Table 3.1, many more can be found through a simple Web search and a number of academic institutions have produced their own software for this

[1]'Natural language processing (NLP), is the attempt to extract a fuller meaning representation from free text. This can be put roughly as figuring out who did what to whom, when, where, how and why.' (Kao & Poteet 2007)

purpose. The limited amount of time available for the HuCEL project did not allow for an evaluation of the various NLP solutions available.

Table 3.1: Natural Language Processing Solutions

Name	**Programming Language**	**URL**
Natural Language Toolkit	Python	`http://www.nltk.org`
Stanford NLP Software	Java	`http://nlp.stanford.edu/software/`
GATE	Java	`http://gate.ac.uk`
Apache UIMA	Java / C++	`http://incubator.apache.org/uima/`
OpenNLP	Java	`http://opennlp.sf.net`
NLP	Java	`http://www.mii.ucla.edu/nlp/`

Secondly, there is no guarantee that an existing NLP solution will integrate smoothly with other components in a system – in this case the Kain Proxy and JTidy. Furthermore, considerable effort may need to be made in order to adapt an existing parser with a specific domain, such as news stories, and this is not always apparent until part-way through the process.

As a result of the time required to evaluate, and possibly modify, existing NLP solutions, we concluded that developing our own algorithm was more likely to result in a working parser in the short timeframe allocated for the HuCEL project, as we would fully understand its implementation and would be able to quickly adapt it to incorporate any changes required. However, we do consider that the use of existing NLP solutions could, with additional time and resources, be used to improve the results of our parser, and we return to this topic in Section 6.5.11.

3.3 Development of our Algorithm

In order to fill the gaps in the existing work discussed in Section 3.1, an algorithm was developed for analysing event-related information with the goal of creating links between stories which discuss related events. This algorithm selects keywords from a news story and formulates a query which can be used to obtain links via the API of a search engine.

Throughout the development of our experimental prototype, our goal was to produce a lightweight algorithm which was 'good enough', i.e. one which could generate search queries that return links to related events. As we were not approaching this task as a natural language processing problem, we were not concerned with tagging each part-of-speech correctly or analysing every single part of the text, but merely to extract sufficient information to find related events. Furthermore, our goal was not to produce an algorithm with high precision and recall, but one which produced results that were of use and interest to users. Whilst these two goals are not necessarily mutually exclusive, we focused on the utility of links to users without measuring precision and recall.

In order to illustrate the development of our algorithm, we have taken a sample page from the Guardian Website, which was used in our experiments, and will use this as a running example for demonstrating how each step of the algorithm is displayed. A reduced screenshot of this page can be seen in Figure 3.1.

3.3.1 Extracting the Main Content from a Web Page

Our algorithm for detecting event-related information assumes that it is working on a plain piece of English text, which only contains the story that we wish to analyse. However, as most Web pages contain a variety of information beyond the main content, such as advertisements and navigation menus, they are not suitable for our algorithm without some form of pre-processing. In order to overcome this problem, we investigated using the SADIe project[2] to separate the content from the rest of the page. Broadly speaking, SADIe offers the following three pieces of functionality (Harper et al. 2006*b*):

1. *De-fluffing:* The removal of unnecessary content which is primarily visual, including 'banners, blank images that provide visual spacing, and advertisements.'

[2] `http://hcw.cs.manchester.ac.uk/research/sadie/`

Figure 3.1: Sample Guardian Web Page

2. *Toggle menus:* The grouping together of all menus, to be placed at the top or bottom of a page (controllable by the user).

3. *Re-order:* Important document items are brought up to the top of the page.

The combination of these three pieces of functionality allows us to utilise SADIe to fetch only the interesting content of the page – what we shall refer to as the *main content* – excluding any navigation elements, advertising etc. Since events are likely to be only mentioned in the main content, extracting just this information prevents our parser from becoming confused by irrelevant material (e.g. advertisements) and text which is the same on every page of a site and therefore unlikely to describe events (e.g. navigation elements).

Whilst SADIe was originally designed to improve the accessibility of Web sites for blind users without requiring any changes by the page author, we have applied the functionality to a different domain. Due to automated tasks, such as our parser, only requiring the main content of a page, the ability to extract this information is just as valuable for machine understanding of Web pages as it is for improving the accessibility of sites.

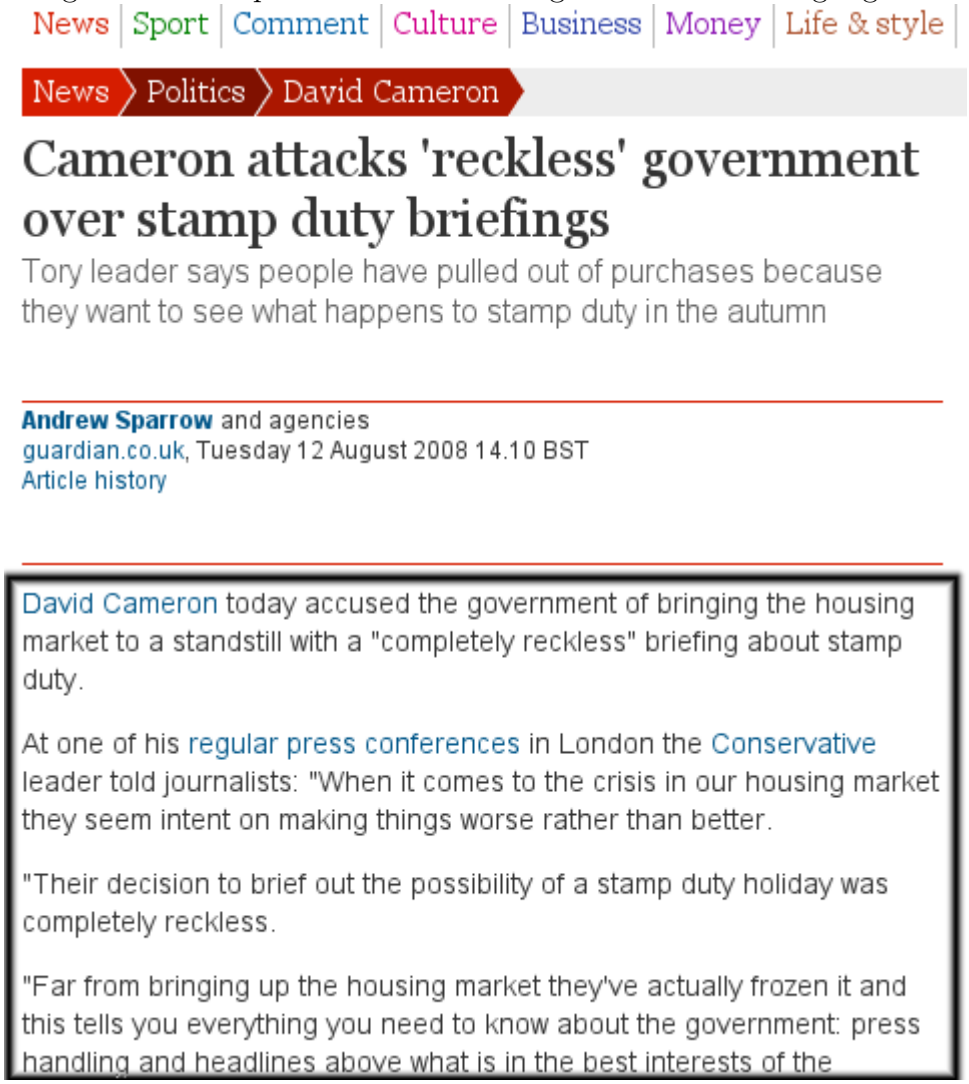
However, whilst a successful initial integration with the SADIe transcoding engine was achieved, the achilles heel of the platform is the need to create an ontology for each individual site (Harper et al. 2006*a*). Ontologies 'can be hard to build, hard to maintain, and hard to use' (Bechhofer et al. 2006), and, as demonstrated in the manual for building ontologies for SADIe, this process is both time consuming and laborious (Lunn 2008). As a result, we ultimately decided to develop a simplified and lightweight version of the process by using a simple XML file[3] to map to the content on sites, rather than creating a separate ontology for each site.[4] Whilst this solution is not as powerful or flexible as the functionality offered by SADIe, it was sufficient for our purposes as we were only interested in one part of the page, and were not performing additional tasks such as reordering elements or toggling menus.

Applying the content mapping functionality to our running example results in the following text being highlighted. A partial screenshot can be seen in Figure 3.2 – due to size constraints, only the top section of the content is highlighted.

[3]Details of the structure and contents of this file can be found in Waring (2009*a*).

[4]The possibility of automatically detecting the content area is discussed further in Section 6.5.5.

Figure 3.2: Sample Guardian Web Page with Content Highlighted



After stripping out the HTML markup, we are left with the following text (of which only two paragraphs are shown):

> David Cameron today accused the government of bringing the housing market to a standstill with a "completely reckless" briefing about stamp duty.
>
> [...]
>
> Cameron called the press conference on his return from his holiday in Devon. He has got political engagements this week before resuming his summer break with a holiday in Turkey next week. Gordon Brown is still on holiday in Scotland.

This running example will be used for the rest of this chapter, to illustrate how various parts of our algorithm work. The full text has been omitted for reasons of space and brevity.

3.3.2 Event Attributes

Based on previous work we proposed a definition of an event as having the following attributes (p.23):

1. Something which happened.
2. The place where it happened.
3. The time when it happened.
4. The set of individuals involved.

Each of these attributes has the potential to be used in constructing queries for finding related events. In order to utilise any particular attribute there are three requirements which must be met:

1. Our parser has to be able to automatically extract the attribute from the text of the story.
2. Our parser has to be able to translate the attribute into a canonical form.
3. The attribute must be demonstrably of use when searching for related events.

There are two possibilities for translating an attribute into a canonical form. For entry names this involves expanding an abbreviation, such as being able to convert a reference to a person by their surname into their full name. For temporal references some arithmetic may be required in order to transform a relative reference into an absolute one (e.g. 'yesterday' might be transformed into 'Sunday, July 12 2009').

In order to evaluate these attributes, we tested each one to discover whether it met the three criteria described above. The results of these investigations are described in the following sections.

3.3.2.1 What Happened

A simple way to identify the 'what happened' attribute of an event is to parse the headline for a news story and assume that any verbs within the headline refer to this attribute. With the assistance of a lexical database such as WordNet,[5] basic part-of-speech recognition is a trivial task. The accuracy of such an approach is reduced by ambiguous words which can be recognised as more than one word type, depending on context - for example, the word 'hand' can be used as a noun or a verb (Yngve 1955). However, given the brevity of headlines (usually six words or less) and their often limited vocabulary, simply picking out verbs from the headline would probably be sufficient for our purposes.

As discussed in Section 3.3.3, the description of 'what happened' in an event can vary from one report to another, with different wording used in each case. This factor means that using the 'what happened' attribute in a search query would miss out any stories which do not use the same word for describing the event. Furthermore, the word used to describe the event is often so broad (e.g. 'crash') that it often returns significant numbers of unrelated events in response to a search query. As a result of these two factors, we concluded that the 'what happened' attribute of an event did not meet the three criteria set out in Section 3.3.2, and therefore decided not to use this particular event attribute as part of our search queries to find related events. As a result, our parser may miss some links to events which are related only by the 'what' attribute – e.g. other train crashes which happen at a different place and time, and involve other people.

[5] `http://wordnet.princeton.edu/`

3.3.2.2 Location of Event

One of the event attributes which we were interested in identifying was the location of an event – particularly those place names referring to specific geographical areas such as 'London'. Whilst these names are almost always expressed as proper nouns, there is generally no way of telling simply by looking at a name whether it refers to a location or not. In order to overcome this problem, we needed to utilise a lexicon (or gazette) of location names which we could use to determine whether a given name referred to a location or to some other entity.

The lexicon which we chose to use for our work was provided by the National Geospatial Intelligence Agency,[6] which is responsible for maintaining a list of place names on behalf of the US Federal Government. This lexicon was chosen because of its comprehensiveness and also the fact that it was freely available and presented in a simple text format which could easily be parsed by a computer program.

In addition to simple lookups on the lexicon, we also noticed that place names often contained modifiers which were not present in the lexicon, such as '*North* England'. As a result, we ensured that our algorithm attempted to match words with these modifiers too, in a similar way to our word expansion method, in order to ensure that the entire description of a location was matched by our parser.

By applying our lexicon lookup to our running example, we can see that three of the proper nouns refer to locations:

Table 3.2: Proper Nouns after Location Lexicon

Proper Noun	**Frequency**	**Location?**
David Cameron	2	No
Devon	1	Yes
Turkey	1	Yes
Gordon Brown	1	No
Scotland	1	Yes

[6] *NGA: GNS Home*, `http://earth-info.nga.mil/gns/html/` (Accessed: 5 June 2008)

3.3.2.3 Time of Event

The temporal attribute of an event, expressed as a single moment or a short duration, is one which has the potential to be used for finding related events which occur at the same time. However, as explained in Chapter 2, temporal information is notoriously difficult to parse and manipulate (Makkonen et al. 2002, Makkonen et al. 2003, Hobbs & Pan 2004), particularly given that few temporal expressions are specified in full and must therefore be calculated from the surrounding context (Mani et al. 2003, Dale & Mazur 2007). This is made especially difficult with regards to news stories, where the most prominent date on the page will often represent the date the article was originally posted or last updated, rather than the date of the event itself. Furthermore, the ways in which timestamps are presented vary greatly from site to site, as can be seen in the following examples (all taken from the most prominent timestamp on the top news story):

1. **Yahoo! News:** '22 mins ago'
2. **BBC News:** '21:02 GMT, Saturday, 11 April 2009 22:02 UK'
3. **Reuters:** 'Fri Apr 10, 2009 9:19am EDT'
4. **CNN:** 'updated 1 hour, 13 minutes ago'
5. **Guardian:** 'Saturday 11 April 2009 18.22 BST'

Extracting and resolving temporal expressions is a significant project in itself and beyond the scope of our work, and as a result we did not use this attribute in our search queries.[7] As a consequence of this decision, we will not be able to produce links to related events based on the temporal attribute (e.g. events which occur before or after the original event), which may limit the number and quality of the final set of links.

3.3.2.4 People Involved

As with locations, the people involved in an event are often easier to extract than the other event attributes, due to the fact that they are almost always expressed as proper nouns. Whilst we encountered the problem of names generally being abbreviated after their first mention – therefore making a simple frequency

[7] Temporal attributes are discussed further in Section 6.5.10.

analysis difficult – after employing our word expansion technique described in Section 3.3.4 we were able to extract the names of individuals and other entities in full.

In addition to the relative ease of extracting the names of people, compared with other event attributes, these names also proved useful in building search queries which returned links to related events. This is possibly due to the fact that, unlike the 'what happened' attribute, names of people will generally be the same across all stories, and therefore will be picked up in a search query.

3.3.3 Focusing on Proper Nouns

One of our reasons for focusing on proper nouns, as opposed to any other keywords, is that they generally remain constant across news stories, regardless of who has written them. For example, news stories about a train crash might use different words to describe the event itself (e.g. 'crash', 'derailment', 'accident') but information such as the location of the event and the people involved will remain the same – possibly because proper nouns refer to concrete entities rather than abstract concepts, and are therefore more likely to have an identity. As a search engine was used to find related events, which utilise the keywords on a page to build their indexes, we needed to pick words which would appear in the original story and in related stories, as these are likely to return links to related events.

When extracting proper nouns, we were aware that many could be composed of several words – common examples being the full name of a person or location. In order to catch these instances, whenever we found a proper noun we would look at the next word in the sentence and see if it was also a proper noun – if so then we considered it to be part of the same phrase. This 'greedy' process (matching as many proper nouns as possible) continued until no further proper nouns were encountered or the end of the sentence was reached, whichever was first. The pseudocode for this section of the algorithm is shown below:

Listing 3.1: Pseudocode for Matching Proper Nouns with Word Expansion

```
proper_nouns = new list

foreach word in sentence
  if current_word.word_type = 'proper noun' AND NOT stop_word(
      current_word)
```

```
    proper_noun = current_word
    while ( words left in sentence )
      if current_word.word_type = 'proper noun'
        proper_noun = proper_noun + ' ' + current_word
      end if
    end while
    proper_nouns.add(proper_noun)
  end if
end foreach
```

Returning to our running example, after running our above algorithm the following output would be generated (words in bold are those proper nouns highlighted by our parser):

> **David Cameron** today accused the government of bringing the housing market to a standstill with a "completely reckless" briefing about stamp duty.
>
> [...]
>
> **Cameron** called the press conference on his return from his holiday in **Devon**. He has got political engagements this week before resuming his summer break with a holiday in **Turkey** next week. **Gordon Brown** is still on holiday in **Scotland**.

Assuming that these two paragraphs were the complete text, we would have the following table of proper nouns identified by our algorithm:

Table 3.3: Proper Nouns Identified by Algorithm

Proper Noun	**Frequency**
David Cameron	1
Cameron	1
Devon	1
Turkey	1
Gordon Brown	1
Scotland	1

3.3.3.1 Filtering out Stop Words

Whilst developing our algorithm to recognise proper nouns, we observed that a number of common words (e.g. 'the', 'at') were being tagged as proper nouns, usually because they occurred at the beginning of a sentence. Such words, often referred to as 'stop words' as they are a sign to stop indexing, are unlikely to be useful in search queries due to the sheer number of stories which use them. As a result, we built a short list of stop words which our algorithm ignores if they are a proper noun on their own. However, stop words are not ignored if they are in the middle of a proper noun consisting of more than one word.

3.3.4 Word Expansion

In the process of deploying our prototype algorithm, we observed that many news stories will refer to a proper noun by its full name in the first instance, and then use an abbreviated version for the remainder of the text. This was most noticeable for the names of individuals, but also occured for other entities, such as company names. For example, a news story might mention 'Gordon Brown' in the first instance, but would refer to him as 'Brown' from there on.

Human readers can of course expand these abbreviations by looking at the words surrounding them, or the individual characters in the abbreviation (Terada et al. 2002). In other cases of automatic text processing, we would need to rely heavily on contextual awareness to disambiguate these abbreviated references and expand them to their full meaning, in the same way as a human reader (Rowe & Laitinen 1995, Larkey et al. 2000). However, as news stories tend to be short – less than 500 words in many cases (McLachlan & Golding 2000) – and focus on one particular topic, we observed that it was sufficient to disambiguate abbreviated versions of words based on a previous mention anywhere within the text.

In order to facilitate word expansion, we examined each proper noun as we encountered it. We compared the noun to a list of all the proper nouns encountered previously in the text. If the current word(s) matched the end of a noun which was already in our list (e.g. 'Brown' would match against 'Gordon Brown' but not 'Brown University'), then we simply incremented the frequency count of that item by one. If no match was found, we added the current word(s) to

our list as a new entry. Whilst this did not catch all abbreviations of proper nouns (e.g. we did not match 'BBC' as being the same as 'British Broadcasting Corporation'), we found that it provided a sufficiently accurate measure for our purposes. Supporting abbreviations which either do not appear in the text as readers are assumed to know what their meaning is, or abbreviations which are not formed by chopping off part of a word, would be an area for future work which could potentially improve the ability of our algorithm to obtain links to related events.

An example of word expansion in action can be seen in a segment of our running sample:

> **Cameron** called the press conference on his return from his holiday in **Devon**. He has got political engagements this week before resuming his summer break with a holiday in **Turkey** next week. **Gordon Brown** is still on holiday in **Scotland**.

When our algorithm reaches the word 'Cameron', it will look through all of the previously encountered proper nouns to see if there is a partial match. As the phrase 'David Cameron' has already been added to the list of proper nouns at this point, the algorithm will assume – correctly in this case – that both words refer to the same entity, and will increment the frequency count of the 'David Cameron' entry by one. Applying this to our running example would give us the following table:

Table 3.4: Proper Nouns after Word Expansion

Proper Noun	**Frequency**
David Cameron	2
Devon	1
Turkey	1
Gordon Brown	1
Scotland	1

As can be seen from Table 3.4, the reference to 'Cameron' has disappeared and the frequency count for 'David Cameron' has increased by one. The same

effect would be achieved if the article writer had referred to David Cameron by his full name in every instance in the text.

3.3.5 Frequency Count

As mentioned previously, as part of our algorithm we kept a list of proper nouns along with the frequency they occured within the text, after expansion had been taken into account. Given the short length of most news stories, we observed that the proper nouns with the highest frequency (usually occuring three or more times in any one text) would generally indicate the event attributes which were most relevant in describing the story.

3.3.6 Building and Executing Queries

After some experimentation with various versions of our algorithm, we found that by utilising the two most frequent proper nouns from a given story (usually entity names or locations) we could obtain a search query which would generally return a set of links to related events. If we run our algorithm on the entire text of our example story, we find that the two most frequent proper nouns are 'David Cameron' with eight references, and 'Alastair Darling' with two references.[8]

After some experimentation with different combinations of the two most frequent popular nouns in a variety of stories, we found that more related links were returned when using these two nouns as two distinct phrases with the boolean 'AND' operator. Our search query for our running example therefore becomes: `'David Cameron' AND 'Alastair Darling'`. At this point the parser side of our algorithm is complete.

In order to obtain links to pages containing information about related events, we had to pass our query to a search engine. We chose to use the Yahoo! Developer Network[9] for this purpose, as it has a freely available API which can be used to programmatically obtain results from search queries. At the time of implementation, Google had deprecated its original API and the new version strictly

[8] Although in this particular example both of the two most frequent proper nouns refer to people – i.e. the 'who' attribute of an event – in the sample of twenty Web pages used in our experiments, 45% of queries generated by our parser included a location as one of the two most frequent proper nouns.

[9] `http://developer.yahoo.com/`

prohibited automated searching and 'permanently storing any search results',[10] and so we could not use it in our prototype. However, the way in which our parser is set up means that it is a trivial matter to plug in a different search engine at a later date, or even multiple search engines, as discussed in Section 6.5.4.

3.3.7 Adding Dynamic Links to Pages

Having developed our experimental prototype algorithm into a lightweight implementation of a parser, we needed to devise a method for displaying the links returned by Yahoo! as a result of processing our search queries. Based on our initial research in Section 2.4, we concluded that the best way to provide this functionality was through the addition of dynamic links to existing Web pages. In most cases, it is not possible to insert markup in documents which we do not own or control (Davis 1995), however, there are three different ways in which we can circumvent this barrier to add dynamic links to existing Web pages.

The first option is to implement a solution on the server side, in which links are added to each page before the response is sent to the browser, using a similar approach to that of Yesilada et al. (2007). Unfortunately, this requires a separate solution for each individual site, as it is likely that most sites will generate the pages which are sent to the browser using different systems. Furthermore, implementing such a solution would require access to the servers from which the site is delivered – access which we generally do not have.

The second option is to add the links on the client side, which would usually entail the use of a browser plugin. The major disadvantage to this approach is that every user has to install the plugin before they can benefit from the functionality which it provides. This barrier is particularly acute for our initial experiments, as it would be even more difficult to get users to install a plugin before they can see what benefits it will offer to them. In addition to this, there is also a substantial development overhead required to create the plugin, which is compounded by the wide variety of browsers in existence, each of which provides a different plugin architecture which is generally not compatible with other browers. At a bare minimum, any plugin would need to be available for both Internet Explorer and Mozilla Firefox if a significant number of users could be expected to install and

[10] *Google AJAX APIs Blog: Google Code Labs and the SOAP Search API*, `http://googleajaxsearchapi.blogspot.com/2009/03/google-code-labs-and-soap-search-api.html` (Accessed: 11 April 2009).

use it. The need to install a plugin, and the lack of ability to do so in some cases (e.g. where a browser requires administrative priviledges in order to install a plugin) can also cause difficulties (Graham et al. 1999).

Another way in which dynamic links can be added to existing Web pages is through the use of a proxy. This software sits between the server and the client, and intercepts all of the requests which pass between them. When intercepting the requests, the proxy has the opportunity to alter the information before it is passed on, for example it can change the Web page before presenting it to the end user. The principal disadvantage of using a proxy is the delay which it introduces between a request being issued and a response being received, as it effectively adds another step between the server and the client.

In our final solution, we opted to use a proxy, which sits in between the server and the client, for several reasons. First and foremost, the need to have the smallest barrier possible to using our software was a strong argument for using a proxy, as it does not require any software to be installed on the server or the client. Secondly, a proxy allowed us to add support for new sites immediately, as no updates would be required to the client or server for this to happen. Finally, we discovered that there was already an existing proxy server, the Kain Proxy,[11] produced in collaboration with Sun Microsystems, which was used as part of the COHSE project[12] and could easily be adapted to utilise our algorithm.

In addition to a method of adding the code for dynamic links to Web pages, we require a way of displaying these dynamic links when a page is requested. Our method is similar to that of Yesilada et al. (2007),[13] which implements the following steps in order to create reverse links for the Hypertext 2007 (HT'07) Web site:[14]

1. Parse server logs to obtain list of remote page referrers for a specific local page.

2. Create an XML file for the local page, which contains the results obtained in the previous step and has the same name (but different extension) to the local page (e.g. `about.html` maps to `about.xml`).

[11] `https://cohp.dev.java.net/`
[12] `http://cohse.cs.manchester.ac.uk/`
[13] Related experiment code and data can be found in Yesilada, Lunn & Harper (2008).
[14] `http://www.sigweb.org/ht07/`

3. When a user requests a page on the HT'07 site, a specific AJAX[15] component fetches the corresponding XML file.

4. A link box is dynamically generated to display the XML linkbase.

Whilst our approach works in the same way for steps 3 and 4, we differ from the procedure used in reverse linking in two ways. Firstly, our links are sourced from the results of a query to the Yahoo! search engine, via its developer API, as opposed to parsing Web server logs. However, the format in which the results are saved (a series of `link` nodes containing `title` and `url` elements) is identical, allowing us to re-use the code for loading in a list of links from the HT'07 web site.

Secondly, we generate the XML files used to populate the link boxes on the fly – i.e. each time a page is requested, the corresponding XML file will be created.[16] This differs to the approach taken in reverse linking, in which the XML files are generated in advance for all of the pages on the HT'07 site. Whilst it might be useful to generate the files in advance in some situations, the corpus we are working with is so large[17] that to generate files for all possible pages would be unfeasible given our limited resources. Furthermore, attempting to guess the pages which a user might request would likely result in a significant amount of wasted effort in generating files for pages which may never be accessed.

As a result of basing our queries on the content of the page, we can, in theory, create a list of links for any Web page – provided that we have created an entry in the content mapping XML file for that particular site – because all of the information we require is publicly accessible. This is in contrast to reverse linking and the dynamic linking discussed in Yan et al. (1996). These techniques cannot be applied to most third party sites as Web server logs are generally not available to anyone other than system administrators.

[15] AJAX (Asynchronous JavaScript and XML) is a technique used to fetch additional information after the original Web page has loaded. This information can be used to dynamically update the Web page on the client side, without requiring the entire page to be reloaded.

[16] These files are cached for 24 hours in order to reduce the number of queries, as there is a limit to the number of requests which can made to the Yahoo! API in any given time period.

[17] As mentioned previously on page 31, recent estimates of the size of the publicly indexable web suggest over 11 billion documents are available (Gulli & Signorini 2005).

3.4 System Architecture

The architecture of our system is described below and illustrated in Figure 3.3.

1. User requests Web page via proxy.
2. Proxy fetches page from Web server and passes the Document Object Model representation of the page to JTidy.
3. JTidy extracts the content element of the page using the content mapping file, strips out all the unwanted elements (e.g. forms) and passes the result to SentParBreaker.
4. SentParBreaker marks up the resulting text and saves it to a file.
5. Our parser reads in the file, breaks the text into sentences and words and attempts to identify proper nouns . Each proper noun is also tested against our lexicon to ascertain whether it is a reference to a location.
6. Our parser takes the two most frequent proper nouns and queries the Yahoo! API to retrieve a set of links, which are saved to an XML file
7. The proxy inserts a reference to the Overlib library into the DOM.
8. The page is returned to the user.
9. The Overlib library reads in the links from the XML file and displays them in the top right corner of the screen.

As can be seen from the architecture diagram, the system consists of a number of components, each of which can be replaced without significantly affecting the rest of the system. For example, the Yahoo! API component could be replaced with a different search engine or multiple search engines.[18] The architecture also has the ability to make use of Linked Data[19] – for example, the location lexicon, which is currently a text file consisting a list of possible locations, could be replaced by a request to a web service, such as GeoNames,[20] that performs the

[18]Section 6.5.4 discusses the possible utilisation of multiple search engines in more detail.

[19]'Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets.' (Bizer et al. 2009)

[20]`http://www.geonames.org/`

Figure 3.3: Architecture Diagram

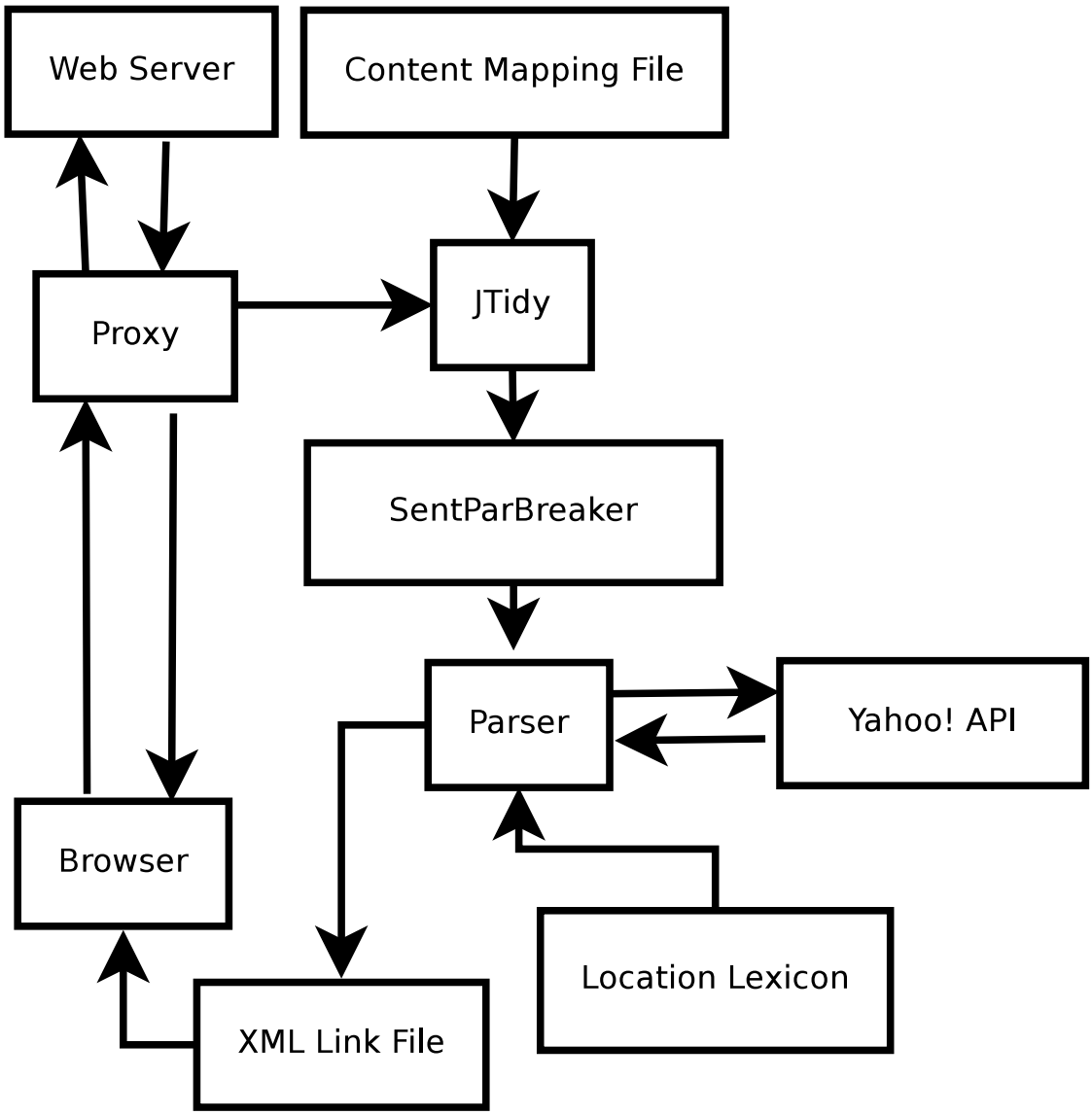


Figure 3.4: Screenshot of Links to Related Events

task of checking whether a given word corresponds to a place name. Although the architecture does not currently expose any of its data in a standard format such as RDF, it would be possible to take the links generated from the Yahoo! API and publish the 'related events' relationship between these links and the original Web page URI.

A sample screenshot of the system in action, taken from one of the BBC News pages used in our experiments and reduced to show the 'related events' box, can be seen in Figure 3.4.

3.5 Conclusions

Having analysed the various attributes of an event, we found that the use of entity and location names, expressed as proper nouns, was effective in creating a search query which would return links to related events. We have created an algorithm which can automatically extract the two most frequent proper nouns from a Web page and construct a query which, when passed through an external search engine API, can retrieve links to related events. These links can then be presented alongside the original Web page to provide the user with the opportunity to

serendipitously discover information about related events. The next chapter, *Methodology and Results*, outlines how we designed two experiments for testing whether these links were considered to be related by users, and presents the results of these experiments.

Chapter 4

Methodology and Results

In this chapter we discuss two experiments conducted in order to increase our understanding of the keywords which can be used to find related events, and to produce data which can be used to evaluate our algorithm described in the previous chapter. We demonstrate how, even with the undue influence of memory bias, users still rate our keywords as related to the story in question, and in some cases display a marked preference for our keywords. We can also clearly see how users find the links produced by our queries to be related to the original story. Finally, even though users often state that they prefer their keywords in our first experiment, the results of our second experiment clearly demonstrate that our algorithm actually produces more related links.

In order to ensure a robust set of results, we created a series of experiments designed to discover the keywords which users associate with news stories and to evaluate our solution by testing it against links produced by these keywords.

In our first experiment, we presented users with a set of Web pages, one at a time, and asked them to provide keywords which they felt described the main story on the page (these keywords had to be taken from the text of the page). Users were then provided with the same set of Web pages and asked to rate two sets of keywords, and indicate which set of keywords they preferred.

In our second experiment, users were presented with a set of Web pages and asked to provide keywords which could be used for finding related events using a search engine. As with the first experiment, these keywords had to be taken from the text of the page. Users were then provided with the same set of Web pages and asked to rate two sets of keywords, and also to express a preference.

In our third and final experiment, users were presented with a set of Web

pages, each with three links listed. One link was the result of a Yahoo! search based on the keywords supplied in our first experiment, another was from the keywords provided in our second experiment, and the final link was from our parser. Participants were asked to rate each link on its relatedness to the main story of the Web page, indicate which link they would follow, and invited to provide additional qualitative feedback through a free-type text box.

During the course of our second experiment it became clear that there would be an insufficient number of participants ((seven in total who managed to complete all of the tasks) for any substantiated conclusions to be drawn from it. As a result, we have not included this experiment in our analysis.

The results for all of our experiments are based upon data obtained from a technical evaluation and user study, run over the period of August to December 2008. We have released this data with an accompanying technical report for each experiment (Waring 2009*c*, Waring 2009*b*, Waring 2009*d*). The data does not contain any personal information about the participants, save for some minor demographic details which cannot be used to identify individuals. All three repository items include the following:

1. The full source code to the experiment (written in PHP), with all accompanying dependancies such as templates and SQL statements to recreate the database.

2. An SQLite database containing all of the data collected in the course of the experiment.

3. Instructions on how to set up and re-run the experiment.

The free availability of this code allows anyone to either repeat our experiments with an empty database, using the same structure and data collection methods, or to reanalyse the data which we collected.

4.1 Experiment Design

Several parts of our methodology were applied across our experiments, including the techniques used for selecting the corpus of pages which were presented to participants. These generic methods from our experiment designs are described in the following sections.

4.1.1 Site Selection

When analysing sites for events, we need to ensure that we cover a number of sites which users are likely to visit. One possible way to achieve this is by selecting sites which are 'popular' amongst the majority of Web users – i.e. those sites which receive more traffic than the rest. Fortunately, the basic data pertaining to popular sites is freely available from several sources on the Web.

4.1.1.1 Overview of Alexa

One site which offers ranking data for popular sites is Alexa,[1] which has been tracking traffic to Web sites for several years and is now owned by Amazon.com. The statistics are obtained by users downloading a toolbar for their browser, which sends data such as the URL accessed and the amount of time spent on a page to Alexa. This data is then processed to obtain rankings for individual Web sites.

One reason why Alexa is a useful source for ranking Web sites is that its data is freely available, both in terms of ease of access and cost. The list of the top 500 sites globally is available on the Alexa site without registration or a fee, as are lists of the top 100 sites for each country, language and category – including news. Although Alexa does charge a fee for some data services, such as customised reports and access to its API, all of the sources which we will be using are available free of charge. This factor will allow other researchers to verify our results and extend our work at a later date.

In addition to the Alexa data being freely available, it also accessible in a machine-readable format. Each site listing in the top 500 results shares a common format, so it is a trivial matter to extract this data programatically. The same applies to the top 100 sites listed by country, language or category, which means that we can always obtain the latest rankings by running a simple script to extract this information from the relevant page on Alexa's website.

Alexa also offers global statistics, whereas some alternative sites, such as Compete,[2] only track statistics for visitors in the US. As a result, Alexa largely avoids any bias which might be introduced as a result of country-specific interests, and also includes sites which are popular to a global audience, such as BBC News.[3]

[1] `http://www.alexa.com`
[2] `http://www.compete.com`
[3] `http://news.bbc.co.uk`

Finally, the Alexa ranking system has already been utilised by a number of researchers in a wide variety of areas, including transforming Web pages to become standards-compliant (Chen & Shen 2006), segmenting Web pages for mobile devices (Hattori et al. 2007), measuring privacy loss and protection (Krishnamurthy et al. 2007) and testing the reliability of the Domain Name System (Ramasubramanian & Sirer 2004).

4.1.1.2 Filtering Alexa Rankings

Whilst we initially attempted to utilise all of the ten most popular news sites provided by Alexa, we found that in several cases we were unable to integrate our solution due to aspects of the sites which were outside our control. As a result, we filtered out several sites from the list provided by Alexa, for reasons which we discuss below, and eventually settled upon five sites which we believed represent pages which users were likely to encounter on a regular basis.

Our first filtering criterion was based on the validity of the HTML used by the news sites. Whilst we did not enforce any compliance with the relevant HTML standards, some sites contained so much invalid HTML that JTidy,[4] the tool used for creating a DOM tree, could not parse the page. As a result, we were forced to remove these sites from our study as it was not technically feasible to automatically extract the content element from individual pages. In doing so, we also encountered the same effect observed by Chen et al. (2005), in that sometimes the appearance of a Web page can be altered after a tool such as Tidy has been run over it.

Our second filtering criterion involved the removal of any sites which were not original news outlets, but simply aggregated news stories from a variety of different sources. Google News[5] is one prominent example of this type of site. Our reasoning for removing such sites was that they contain the same content as other news sites and that to include them would simply clutter up the results returned to the user. Furthermore, in many cases Google News only provides brief summaries of news stories, which may not contain all of the event-related information which is required by our parser in order to generate relevant search queries.

[4] JTidy is a Java implementation of the HTML Tidy software, which takes the HTML of a Web page and attempts to transform it into a valid DOM tree, whilst optionally tidying up the HTML to include spacing etc. JTidy is freely available from: `http://jtidy.sourceforge.net`.

[5] `http://news.google.com/`

Our third filtering criterion was to remove all sites which were using any language other than English, as we do not currently support multiple languages as part of our work.[6] However, as all of the top 10 news sites were written in English, in this case the filtering rule had no effect.

Finally, any sites which require users to login in order to access the majority of content, such as the New York Times,[7] were excluded from our list. Our reasoning for this is down to two factors. First, it may not be possible to fetch a particular page and extract event information for it if some form of authentication, which we cannot guarantee to have, is required. Second, we do not wish to redirect users to pages containing potentially related events if they cannot access this information without logging in.

4.1.1.3 Modified Alexa Rankings

In order to obtain a list of sites for use in our experiment, we downloaded the Alexa top 10 news sites list[8] on 11 August 2008 and applied the filters described in Section 4.1.1.2. The results are shown in Table 4.1.

Table 4.1: Modified Alexa Rankings

Alexa rank	**Domain**	**Reason for exclusion**
1	`news.yahoo.com`	n/a
2	`www.cnn.com`	n/a
3	`news.bbc.co.uk`	n/a
4	`www.nytimes.com`	Login required
5	`news.google.com`	Aggregated news site
6	`www.msnbc.msn.com`	Invalid HTML
7	`www.reuters.com`	n/a
8	`news.aol.com`	Invalid HTML
9	`www.foxnews.com`	Invalid HTML
10	`www.guardian.co.uk`	n/a

[6] The possibility of supporting additional languages beyond English is discussed in Section 6.5.8.

[7] `http://www.nytimes.com/`

[8] `http://www.alexa.com/browse/general/?&CategoryID=8&mode=general&R=False&Start=1&ListingCount=True&SortBy=Popularity`

As a result of our filtering criteria, we were left with a list of the following five sites to select individual stories from.

1. **Yahoo! News:** `news.yahoo.com`
2. **BBC News:** `news.bbc.co.uk`
3. **Reuters:** `www.reuters.com`
4. **CNN:** `www.cnn.com`[9]
5. **Guardian:** `www.guardian.co.uk`

4.1.2 Page Selection

Having decided upon five sites for inclusion in our experiments, we needed to select specific pages which could be used in our experiments. We suggest that twenty pages provides a good balance between ensuring a wide variety of pages, and creating an experiment which is short enough that participants do not exit the experiment before answering all of the questions.[10] In order to select these pages, we visited each of the five sites on one day (13 August 2008) and took a local copy of the four news stories which were the highest up on the front page of the site. Our reasoning for selecting the highest up stories is that these are the ones which users are most likely to see and read when entering the site, due to their prominence – albeit a position dictated by editorial control.

The following pages were used in our experiments.

Table 4.2: Web Pages Used

ID	URL
1	`http://www.reuters.com/article/politicsNews/idUSWAT00989220080812`

[9]At the time of performing our experiment, news stories listed under `www.cnn.com` were redirected automatically to `edition.cnn.com`, however the content of these pages remained unchanged.

[10]Even with twenty pages, we found that 53% (32 out of 62) of participants failed to complete the first experiment, and suggest that using a greater number of pages would have increased the drop out rate further. A full analysis of dropout rates can be found in Section 5.4.

2	http://www.reuters.com/article/politicsNews/ idUSN1030195920080811
3	http://www.reuters.com/article/technologyNews/ idUSN1140611920080811
4	http://www.reuters.com/article/technologyNews/ idUSWNA802020080811
5	http://news.yahoo.com/s/ap/20080812/ap_on_re_eu/ georgia_russia;_ylt=AprHUEAeSFSk20AZg0yHMyes0NUE
6	http://news.yahoo.com/s/ap/20080812/ap_on_el_pr/obama_ untaxing_seniors;_ylt=ApoW7JelujNbN3.aYbR5JJ2s0NUE
7	http://news.yahoo.com/s/ap/20080812/ap_on_sp_ol/oly_ swm_swimming;_ylt=AsrXFeCxEe2shyr3jcb9HDCs0NUE
8	http://news.yahoo.com/s/ap/20080812/ap_on_re_as/ pakistan_bombing;_ylt=AtBYSSQr7jQFXlm1Z0I4YeCs0NUE
9	http://edition.cnn.com/2008/BUSINESS/08/12/skorea. pardons.ap/index.html
10	http://edition.cnn.com/2008/WORLD/africa/08/11/ zimbabwe.talks/index.html
11	http://edition.cnn.com/2008/TECH/science/08/11/bush. endangered.species.ap/index.html
12	http://edition.cnn.com/2008/TECH/space/08/11/nasa. orion/index.html
13	http://news.bbc.co.uk/1/hi/world/europe/7555858.stm
14	http://news.bbc.co.uk/1/hi/business/7555788.stm
15	http://news.bbc.co.uk/1/hi/uk/7555299.stm.html
16	http://news.bbc.co.uk/1/hi/uk_politics/7555691.stm
17	http://www.guardian.co.uk/world/2008/aug/12/russia. georgia
18	http://www.guardian.co.uk/media/2008/aug/12/bbc.radio
19	http://www.guardian.co.uk/environment/2008/aug/11/ kingsnorthclimatecamp.activists
20	http://www.guardian.co.uk/politics/2008/aug/12/ davidcameron.housingmarket

From here on we will refer to a Web page by its ID in order to avoid long URLs within the text.

N.B. As mentioned previously, Web pages in general, and news pages in particular, are prone to be changed, moved or deleted without warning. As such, some of the URLs listed in Table 4.2 may no longer be accessible or contain the same content, which is why we took local copies for our experiments, as discussed in Section 4.1.3.1.

4.1.3 Robustness of Experiments

We have taken a number of steps to ensure that the results from our experiments are robust and not overly influenced by any external factors.

4.1.3.1 Local Copies of Pages

Before releasing the experiment to potential participants, we downloaded all of the Web pages which we were using. By providing local copies, we could ensure that the Web pages would be available whilst the experiment was being run, as news stories can sometimes disappear a short time after being published. Furthermore, by using local copies we could ensure that all participants were viewing exactly the same text as one another, and the same as our parser used when producing the keywords for the page.[11]

4.1.3.2 Hosting the Experiment Online

By running the experiment online, we ensured that there was no facilitator available who could, intentionally or not, guide the user towards the 'correct' answer – i.e. the one which would produce the expected results (Rosenthal 1966). In addition, we mitigated possible feelings of pressure upon participants to perform as they were able to complete the experiment at their own pace.

4.1.3.3 Consistency of Information

In order to ensure a consistent environment across participants, we stablised many of the variables involved in the experiments. The same set of twenty Web pages

[11] As discussed in Section 2.4, Web pages can change or disappear frequently, with 40% of pages on popular sites – such as the ones which we are using – changing on a weekly basis (Cho & Garcia-Molina 2000).

was used for all three experiments, and the keywords and links produced by our parser were generated once at the beginning and then reused in a static context to ensure that all participants were presented with the same information.

4.1.3.4 Randomisation and Memory Bias

As a result of asking participants for pieces of information (in this case, keywords which describe the main story on the Web page) and then later asking them to rate this information, there is likely to be a degree of memory bias inherent within our keywords experiment. This bias could arise as a result of participants remembering that they entered certain keywords during the course of the experiment, and therefore automatically rating them as highly relevant. In order to reduce this bias, we attempted to disrupt the memory of participants by asking them to provide keywords for each individual Web page first, before going on to rate two sets of keywords for each story (specifically, the set of keywords suggested by the user earlier in the experiment and the set of keywords generated by our parser). Furthermore, we randomised the order in which the Web pages were displayed in both parts of the experiment, so it is unlikely that any user will have seen all of the pages in the same order for both parts, further reducing the likelihood of memory bias. However, even though we have taken steps to reduce the possibility of memory bias affecting the results, it is still likely that some participants will have their ratings affected by memory.

In addition to randomising the order of Web pages, we also placed the keywords on a random side of the page – i.e. sometimes the user's keywords would be on the left, and sometimes on the right, with the parser keywords on the other side. This is an attempt to mitigate the possibility of users getting bored part way through the experiment and simply selecting the same option each time.

4.1.3.5 Lack of Source Information

At no point during the keywords experiment did we inform users as to where the sets of keywords they were asked to rate originated. This was intended to ensure that there was no inherent bias when selecting the results. If we had informed the participants of the source of the keywords (i.e. that one set derived from their earlier answers in the experiment, and the other set from our parser), it is possible that they would always have expressed a preference for their keywords (having suggested them) or been biased towards expressing a preference for our

keywords in an attempt to give what they thought was the 'correct' answer. The same principle was applied to our links experiment – participants were not made aware of the origins of the links, or even that different sources were used to obtain the links.

4.2 Keywords Experiment

Our first experiment involved presenting users with a series of Web pages and asking them to provide up to five keywords which they felt described the story on the page. Participants were then asked to rate two sets of keywords for the same series of Web pages and express a preference for one set or the other.

The aim of our experiment was to discover the keywords which users would associate with news stories, and to compare them with the results of running our parser over the same page. We were looking for looking for two outcomes:

1. An indication of how relevant users felt the keywords generated by our parser to be in relation to the main story on the Web page.

2. An indication of how relevant users felt our parser keywords to be in comparison to their keywords.

Furthermore, the keywords obtained from this experiment were processed through Yahoo!'s search API and the results used in our links experiment.

4.2.1 Methodology

The aspects of our methodology which are specific to our keywords experiment are described in the following sections.

4.2.1.1 Participants

As discussed in Section 4.1.3.2, our experiment was made available online and could be completed wherever and whenever the participants wished. Participants were recruited via emails sent to individuals and group mailing lists. Before beginning the experiment, participants were asked for some general demographic information, which was sufficient for our analysis but not specific enough to identify individuals (i.e. all participants were anonymous, so that we could release all the data freely). The following pieces of demographic information were requested:

1. **Gender:** Either male or female.
2. **Age Range:** Possible values were:
   (a) 25 and younger
   (b) 26-35
   (c) 36-45
   (d) 46 and older
3. **Time spent on the Web each week:** Possible values were:
   (a) Less than 1 hour
   (b) 1-5 hours
   (c) 6-10 hours
   (d) 11-20 hours
   (e) More than 20 hours
4. **Native English speaker:** Either yes or no.

Sixty two users took part in our experiment, of whom thirty completed all of the tasks required.[12] Participants reported themselves as spending varying amounts of time browsing the Web – a graphical breakdown of this information can be found in Figure 4.2.

Out of those who completed the experiment, 27 (90%) participants regarded themselves to be native English speakers, with 3 (10%) claiming to be non-native speakers. Participants were able to select their age from four ranges, the most common being 26-35 (50% of participants) and the majority of participants (77%) were 35 or under. A full breakdown of the ages of participants can be found in Figure 4.1.

Participants were neither offered nor received any reward or incentive for taking part in the experiment, and the initial instructions emphasised that there were no right or wrong answers to any of the questions.

[12]In both experiments, we elected to only use data from participants who had completed all of the tasks.

4.2.1.2 Questions

The questions which participants were asked to answer are described in the following sections. All questions were compulsory and participants were not permitted to continue to the next question until they had answered the current one. No time limit was imposed for individual questions or the experiment as a whole. However, we did log the time it took each participant to answer each question, and the total time to complete the experiment (if applicable).

4.2.1.3 User Keywords Question

Participants were presented with a static copy of a Web page and asked to supply a set of up to five keywords which they felt described the story on the page. The only constraint placed on participants was that the keywords entered must be present on the page, so they could not suggest any other keywords which they felt described the story. This constraint was imposed because our parser can only work on the text of the page, and therefore to allow users to enter other keywords would prevent us from being able to make a fair comparison between their keywords and those from our parser. This question was repeated for each of the twenty pages in our corpus, presented in a random order.

4.2.1.4 Keywords Rating Question

After supplying keywords for each Web page, users were shown the same set of Web pages again (in a random order, as described in Section 4.1.3.4). In addition to the page itself, participants were shown two sets of keywords and asked to perform the following tasks:

1. Rate how well the two sets of keywords described the story on the page.
2. Indicate which set of keywords they prefered.

For both keyword ratings, users were asked to select one of the following options, using a Likert scale:[13]

1. Highly irrelevant.

[13] A Likert scale asks participants to indicate their answer to a question, or agreement with a given statement, from an ordered range of options. Any number of options can be offered, though 5-point and 7-point scales are common, and an odd number of options allow participants to express a neutral opinion, e.g. 'Neither relevant or irrelevant'.

2. Slightly irrelevant.

3. Neither relevant or irrelevant.

4. Somewhat relevant.

5. Highly relevant.

No guidance was provided to participants as to the definition of 'relevance', it was left to each individual to interpret this as they saw fit as we did not wish to lead participants towards any particular answer.

For expressing the preference between the two sets of keywords, a similar Likert scale was used, with the following options:

1. Strongly prefer left set of keywords.

2. Slightly prefer left set of keywords.

3. Neutral (i.e. no preference).

4. Slightly prefer right set of keywords.

5. Strongly prefer right set of keywords.

As mentioned previously, participants were not made aware of the source of each set of keywords, so there was no sense of them being able to express a preference between the keywords which they supplied and those generated by our parser. However, in our experiment we did log the source of the keywords on each page and were therefore able to determine which set the participant preferred, without visibly marking them as such.

4.2.2 Results

The results from our keywords experiment are listed in the following sections.

4.2.2.1 Keyword Scores

The following scores were used for rating both user and parser keywords throughout the experiment. These scores were used internally – participants were only shown the description.

Table 4.3: Keyword Scores

Score	Description
1	Highly irrelevant
2	Slightly irrelevant
3	Neither relevant or irrelevant
4	Somewhat relevant
5	Highly relevant

4.2.2.2 Preference Scores

The following scores were used for rating a preference between the user and parser keywords throughout the experiment. These scores were used internally – participants were asked to choose between two sets of keywords and were not told which were provided by our parser and which were provided by themselves.

Table 4.4: Preference Scores

Score	Description
1	Strongly prefer user keywords
2	Slightly prefer user keywords
3	Neutral
4	Slightly prefer parser keywords
5	Strongly prefer parser keywords

The following table lists the modal score for user and parser keywords for each Web page in our study. The numbers in brackets represent the magnitude of the modal value; multiple values are listed for cases where there was more than one modal value. We have chosen to use the modal value as it demonstrates which rating attracted the most support amongst participants, and in most cases the modal value also represents the majority of opinions expressed. Furthermore, as answers were selected from a scale of five discrete values, the modal value is an appropriate measure to use. Web page IDs are based on Table 4.2.

Table 4.5: Keyword and Preference Ratings for Individual Web Pages

Web page ID	**Modal user keywords rating**	**Modal parser keywords rating**	**Modal preference**
1	5 (23)	4 (18)	1 (21)
2	5 (21)	4 (10)	1 (21)
3	5 (27)	4 (16)	1 (18)
4	5 (27)	4 / 1 (9)	1 (23)
5	5 (21)	4 (22)	1 (15)
6	5 (24)	4 (14)	2 (13)
7	5 (25)	4 (9)	1 (25)
8	5 (23)	2 / 4 (11)	1 (20)
9	5 (20)	4 (16)	1 (15)
10	5 (21)	4 (18)	1 (15)
11	5 (25)	4 (13)	1 (19)
12	5 (25)	2 (9)	1 (24)
13	5 (21)	4 (17)	1 (12)
14	5 (25)	4 (12)	1 (18)
15	5 (24)	4 (14)	1 (24)
16	5 (22)	4 (11)	1 (24)
17	5 (24)	4 (11)	1 (19)
18	5 (20)	5 (14)	2 (16)
19	5 (24)	4 (17)	1 (16)
20	5 (23)	4 (14)	1 (21)

As can be seen from the results above, the majority of participants rate the keywords which they supplied as ‘Highly relevant’, whereas for the keywords generated by our parser the modal rating is generally ‘Somewhat relevant’. In almost all cases the modal preference is ‘Strongly prefer user keywords’, and there are no cases where the modal preference indicates that participants prefer our keywords.

If we summarise the results as a whole, we can see how many times a participant rated a set of keywords as having some degree of relevance to the original story (i.e. a rating of 'Somewhat relevant' or 'Highly relevant'). This summary is shown in Table 4.6.

Table 4.6: Relevance Ratings for Keywords

Keywords Source	**Frequency of 'Relevant' Ratings**
User	587 (98%)
Parser	352 (59%)

In other words, even before passing our keywords into Yahoo!'s search API, participants rate our keywords as being either 'Somewhat relevant' or 'Highly relevant' in 59% of cases.

4.2.2.3 Correlation Between Age and Keyword Ratings

The Pearson correlation coefficient[14] (r) was calculated for the age of the participant and the ratings given to the keywords suggested by users and our parser. The results are shown in Table 4.7.

Table 4.7: Correlation Between Age and Keyword Ratings

Keywords Source	**Pearson correlation coefficient (r)**
User	0.06
Parser	-0.29

[14]The Pearson correlation coefficient, often denoted by r, is a measure of the correlation between two variables, x and y. It provides a value of $-1 \leq r \leq 1$ which indicates the level of linear dependence between the two variables. A positive value of r indicates a positive correlation between the variables (i.e. as x increases, y decreases), and a negative value of r indicates a negative correlation (as x increases, y decreases). A value of $r = 0$ implies no linear correlation between x and y.

As can be seen from the values above, there is no correlation between the age of a participant and the ratings they assign to keywords suggested by users, and only a small negative correlation between age and ratings for keywords supplied by our parser. This suggests that the age of the user does not affect the ratings provided to keywords.

4.2.2.4 Correlation Between Time Spent on Web and Keyword Ratings

The Pearson correlation coefficient (r) was calculated for the time a participant spends on the Web each week and the ratings given to the keywords suggested by users and our parser. The results are shown in Table 4.8.

Table 4.8: Correlation Between Time Spent on Web and Keyword Ratings

Keywords Source	**Pearson correlation coefficient (r)**
User	0.26
Parser	-0.01

As can be seen from the values above, there is no correlation between the amount of time a participant spends on the Web and the ratings they assign to keywords from our parser, and only a small positive correlation between time spent on the Web and ratings for user keywords. This suggests that the amoutn of time a participant spends on the Web does not affect the ratings provided to keywords.

4.2.2.5 Correlation Between Native Language and Keyword Ratings

We originally intended to compare the ratings of native English speakers with those participants for whom English was not their native language, particularly to see if their ability to suggest and rate related keywords was affected by this factor. However, due to the small number of participants for whom English was a non-native language (only three out of the thirty participants fitted into this category), we were not able to perform any statistical analysis which would produce relevant results.

4.2.2.6 Correlation Between User and Parser Keyword Ratings

The following table shows the Pearson correlation coefficient of the user and parser keyword ratings broken down by Web page.

Table 4.9: Correlation Between User and Parser Keyword Ratings

Web page ID	**Pearson correlation coefficient (r)**
1	−0.26
2	−0.10
3	−0.02
4	0.16
5	0.02
6	−0.05
7	−0.13
8	−0.23
9	−0.04
10	−0.07
11	−0.15
12	−0.25
13	−0.05
14	0.14
15	0.12
16	0.21
17	0.21
18	0.30
19	−0.19
20	0.38
Overall score:	−0.01
Lowest score:	−0.27
Highest score:	0.38

As can be seen from Table 4.9, for any given Web page w, $-0.27 \leq r_w \leq 0.38$, and the overall value when all keyword ratings are aggregated is -0.01, suggesting that there is no correlation between the rating given to user keywords and the rating given to parser keywords for the same Web page.

4.2.3 Summary

A summary of the results from our keywords experiment follows:

1. Keywords produced by our parser are rated as having some degree of relatedness ('Somewhat relevant' or 'Highly relevant') in 59% of cases.

2. The modal rating for parser keywords is 'Somewhat relevant' or 'Highly relevant' for 19 out of 20 Web pages.[15]

3. Even with the effect of memory biases, participants still assign our keywords a rating equal to or greater than the rating for their keywords in 25% of cases.

Based on the results from this experiment, we can clearly see that our parser is producing keywords which participants consider to be related to the original story, even if such keywords are often rated lower than the ones supplier by the participants. However, as one of the objectives of the HuCEL project is to produce *links* to related events, it is not sufficient for the keywords generated by our parser to be considered related to the original story, as the keywords are only an intermediate step and do not represent the final result which will be presented to users. In order to build on the results obtained in our keywords experiment and evaluate the final result, we took all of the keywords and passed them through the Yahoo! search API. This produced a list of links, which were used in our next experiment where we allowed participants to rate and rank the links in a similar way to the keywords.

4.3 Links Experiment

Our links experiment involved presenting users with a series of Web pages and asking them to indicate how related three links were to the story on the page.

[15] In two of these cases, the most popular rating was tied.

Participants were also asked to indicate which link they would follow if these three were presented to them as the results of a search engine query.

The aim of our links experiment was to investigate how users perceived the relatedness of the links generated from user keywords and the keywords extracted by our parser using the algorithms described in Chapter 3.[16] We were looking for two outcomes:

1. How related the links generated from keywords supplied by our parser were on their own (standalone test).

2. How related the links generated from keywords supplied by our parser were, relative to the links generated from keywords supplied by users (comparative test).

In Chapter 5, *Analysis and Discussion*, we discuss whether these two tests were satisfied.

4.3.1 Methodology

The methodology for our links experiment was similar to that of our keywords experiment. We had to remove one of the Web pages (number 3) as the keywords produced for it by users returned no results when fed through the Yahoo! API. As a result, we could not produce a full set of links for participants to choose from and this would not enable us to make a fair comparison between links for that particular page. However, we left the pages in the same order within the database so that a direct comparison could be made between the two experiments (i.e. page 1 in the keywords experiment is page 1 in the links experiment etc.).

4.3.1.1 Participants

In the same way as our keywords experiment, our links experiment was made available online and could be completed wherever and whenever the participants wished. Participants were again recruited via emails sent to individuals and group mailing lists. Before beginning the experiment, participants were asked for some general demographic infomration, which was sufficient for our analysis but not specific enough to identify individuals (i.e. all participants were anonymous, so

[16]This reflects our decision to put the human factors element at the core of our work, as discussed in Section 3.1.2.

that we could release all the data freely).The demographic information requested is the same as that for our keywords experiment, detailed in Section 4.2.1.1.

Fixty eight users took part in our experiment, of whom twenty one completed all of the tasks required.[17] Participants reported themselves as spending varying amounts of time browsing the Web – a graphical breakdown of this information can be found in Figure 4.4.

Out of those who completed the experiment, 19 (90%) participants regarded themselves to be native English speakers, with 2 (10%) claiming to be non-native speakers. Participants were able to select their age from four ranges, the most common being 26-35 (43%) and the majority of participants (71%) were 35 or under. A full breakdown of the ages of participants can be found in Figure 4.3.

As with our keywords experiment, participants were neither offered nor received any reward or incentive for taking part in the links experiment, and the initial instructions emphasised that there were no right or wrong answers to any of the questions.

4.3.1.2 Questions

Participants were presented with a series of Web pages and asked to rate three links on how related they were to the original story described on the Web page. For each link participants were shown the link title, URL and a short summary of the page[18] and asked to assign one of the following ratings:

1. Same story
2. Closely related
3. Related
4. Distantly related
5. Not related

The following values were assigned to each of the ratings. The higher the value, the more related the link is to the original story.

[17] As with our keywords experiment, only data from participants who completed all of the tasks was used in our results and analysis.

[18] All three pieces of information were taken directly from the results of a query to Yahoo's search API.

Table 4.10: Link Rating Values

Value	Description
1	Not related
2	Same story
3	Distantly related
4	Related
5	Closely related

In addition, for each Web page users were asked which of the three links they would follow to find information about related events if they were presented as the results of a search engine query. Finally, participants were offered the option of leaving qualititative feedback on the links through a free-form text box.

4.3.2 Results

The following table lists the modal rating for the links generated by user keywords and parser keywords for each Web page in our study, based on the values in Table 4.10. The numbers in brackets represent the magnitude of the modal value; multiple values are listed for cases where there was more than one modal value.[19]

Table 4.11: Link Ratings for Individual Web Pages

Web page ID	Modal user link rating	Modal parser link rating
1	4 (8)	3 (9)
2	1 (10)	3 (11)
3	n/a	n/a
4	3 (6) / 1 (6)	2 (12)
5	1 (9)	5 (16)
6	5 (10)	4 (9)
7	5 (9)	5 (12)
8	4 (8)	2 (16)

[19]Web page 3 has no results due to being excluded from our survey (Section 4.3.1).

9	1 (20)	2 (17)
10	5 (15)	2 (19)
11	5 (6) / 1 (6)	2 (19)
12	4 (13)	2 (16)
13	4 (8)	5 (13)
14	3 (9)	3 (8)
15	1 (14)	2 (11)
16	3 (13)	2 (20)
17	4 (7)	5 (16)
18	1 (6) / 4 (6)	2 (16)
19	4 (9)	5 (14)
20	3 (10)	2 (12)

As can be seen from the figures above, the modal parser link rating is often not only the most popular rating, but also accounts for the majority of ratings – i.e. on an initial examination there appears to be a consensus across participants for many of the links generated by either our parser or user keywords. For example, all but one of the participants rate the parser link for Web page 16 as 'Same story', and the same applies to Web page 9 where all but one of the participants rate the user link as 'Not related'.

4.3.2.1 Correlation Between Age and Link Ratings

The Pearson correlation coefficient (r) was calculated for the age of the participant and the ratings given to the links generated from user and parser keywords. The results are shown in Table 4.12.

Table 4.12: Correlation Between Age and Link Ratings

Links Source	**Pearson correlation coefficient (r)**
User	-0.11
Parser	-0.16

As can be seen from the values above, there is a small negative correlation between the age of participants and the ratings they assign to user and parser links. However, we suggest that such a small correlation is insufficient to imply that older participants tend to assign lower ratings to both user and parser links.

4.3.2.2 Correlation Between Time Spent on Web and Link Ratings

The Pearson correlation coefficient (r) was calculated for the time a participant spends on the Web each week and the ratings given to the links generated from keywords suggested by users and our parser. The results are shown in Table 4.13.

Table 4.13: Correlation Between Time Spent on Web and Link Ratings

Links Source	**Pearson correlation coefficient (r)**
User	0.02
Parser	0.05

As can be seen from the values above, there is no correlation between the amount of time a participant spends on the Web and the ratings they assign to links generated from either user or parser keywords.

4.3.2.3 Correlation Between Native Language and Link Ratings

We originally intended to compare the ratings of native English speakers with those participants for whom English was not their native language, particularly to see if their ability to rate related links was affected by this factor. However, due to the small number of participants for whom English was a non-native language (only two out of the twenty one participants fitted into this category), we were not able to perform any statistical analysis which would produce relevant results.

4.3.2.4 Correlation Between User and Parser Link Ratings

The following table shows the Pearson correlation coefficient of the user and parser link ratings broken down by Web page.

Table 4.14: Correlation Between User and Parser Link Ratings

Web page ID	**Pearson correlation coefficient (r)**
1	0.22
2	0.27
3	n/a
4	−0.16
5	−0.20
6	0.06
7	−0.13
8	0.23
9	−0.08
10	−0.75
11	0.06
12	0.27
13	0.15
14	0.62
15	−0.31
16	0.00
17	−0.04
18	−0.04
19	−0.14
20	0.04
Overall score:	0.05
Lowest score:	−0.75
Highest score:	0.62

As can be seen from Table 4.14, there are two outlying cases where the correlation coefficient is displaced from the other values, specifically pages 10 ($r = -0.75$) and 14 ($r = 0.62$). These results suggest that there may be a correlation between the ratings given to user and parser links. However, on closer examination of the individual ratings we can see that the majority of ratings have no correlation, and that there are only two or three cases where there does

appear to be a strong positive or negative correlation between the ratings. With a small dataset (twenty one pairs of ratings per page), these cases are sufficient to skew the final result. If we remove pages 10 and 14 from our results then we can see that for any given Web page w, $-0.31 \leq r_w \leq 0.27$, and the overall value when all keyword ratings are aggregated is 0.05, suggesting that there is no correlation between the rating given to user keywords and the rating given to parser keywords for the same Web page.

4.3.3 Summary

A summary of the results from our links experiment follows:

1. Participants state that they would click the links generated by our parser's keywords over the ones generated by human suggested keywords in 63% of cases.

2. Users rate the links generated by our parser's keywords as equal to or better than the ones generated by human suggested keywords in 60% of cases.

Having tested the keywords generated by our parser, and the links generated from those keywords, our set of experiments at this point is complete.

4.4 Conclusions

In the next chapter, *Analysis and Discussion*, we will examine the results of our experiments in more detail, suggesting why users may prefer our links over the ones provided by keywords from other users and how widely their ratings are distributed, both amongst Web pages in general and from specific sites.

Figure 4.1: Breakdown of Keyword Experiment Participants by Age Range

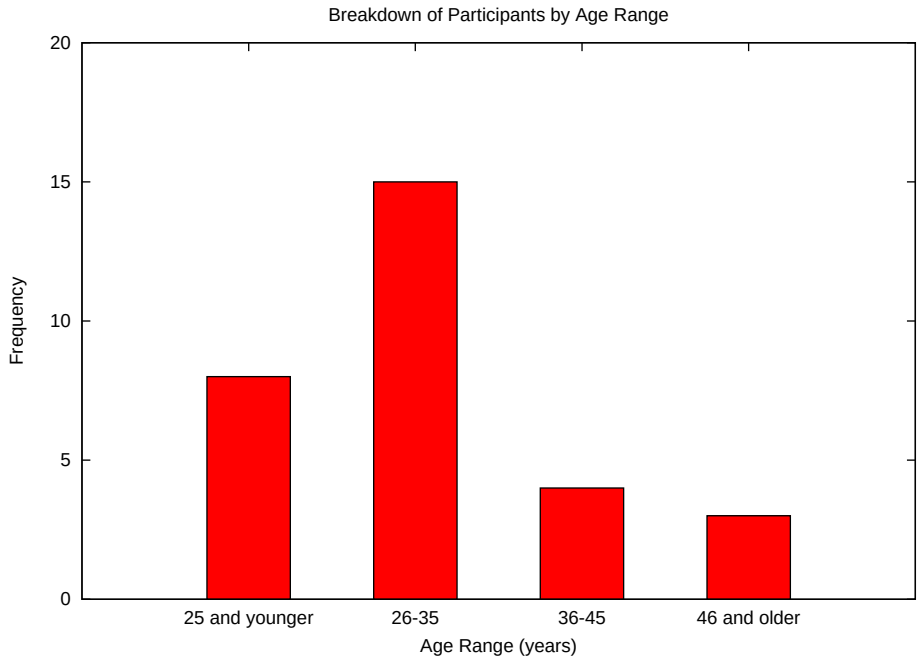


Figure 4.2: Breakdown of Keyword Experiment Participants by Time Spent on the Web

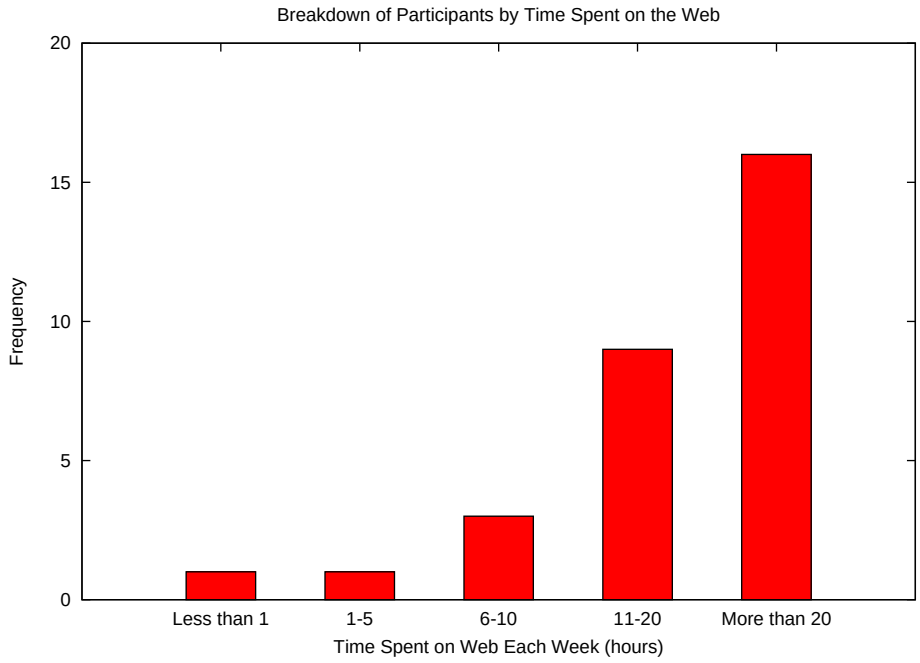


Figure 4.3: Breakdown of Link Experiment Participants by Age Range

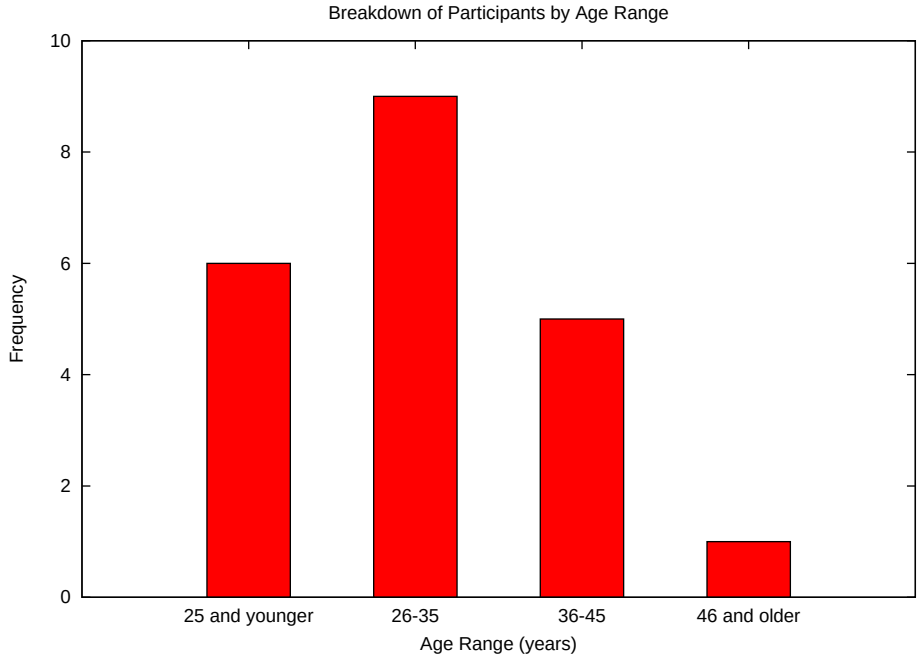
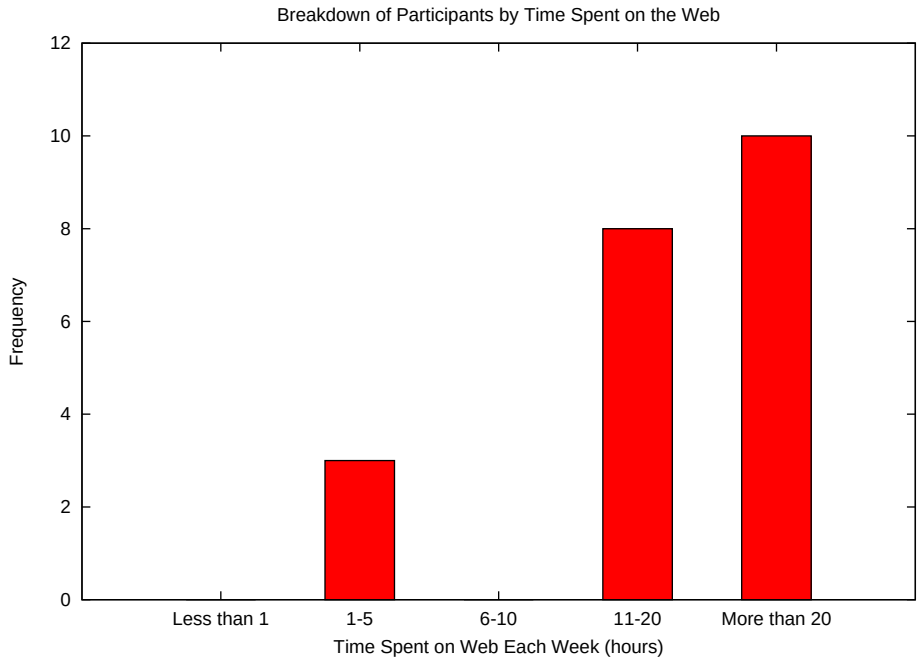


Figure 4.4: Breakdown of Link Experiment Participants by Time Spent on the Web



Chapter 5

Analysis and Discussion

This chapter presents a detailed analysis of the results from our experiments. The analysis of keyword preferences shows that users display a weaker preference for their keywords than we would expect, suggesting that our keywords are making users think again about what they entered (Section 5.1.2). Furthermore, results which indicate that users prefer the links produced by our algorithm suggest that user keywords are returning links to the same event, whilst our results are focused on links to related events (Section 5.2.2).

Our analysis is split into four parts. First, we examine the results from our keywords experiment, to see if there are any trends within users or across Web pages. Second, we analyse the results from our links experiment, again looking for trends and correlations. Third, we combine the results from the two experiments, as they were based on the same data sets, to demonstrate the perception mismatch between the keywords which users prefer and the links which they prefer. Finally, we make some general observations on statistics such as participant dropout rates, which may have applications beyond our specific area of research.

5.1 Analysis of Keyword Experiment Results

As described in Section 4.2, our keywords experiment involved two stages. First, we presented participants with a series of Web pages and asked them to provide keywords which they felt described the story on the page. Second, we presented participants with the same series of Web pages and for each one asked them to rate two sets of keywords (one provided by the user, the other by our algorithm) and indicate which set they preferred.

5.1.1 Individual Web Pages

The first observation which we can make about these results (shown in Table 4.5) is that users clearly prefer their own keywords in the majority of cases, both directly by expressing an explicit preference for them and implicitly by assigning a higher score to them. This is not surprising, as we would expect users by default to rate their own keywords higher than ones provided from elsewhere,[1] and suggest that the barrier required to overturn this preference is a high one.

However, the strong preference for user keywords in these results is not universal, and for each Web page there were between five and eighteen participants (17-60% of the total) who expressed a preference of 'slightly prefer' user keywords or less (i.e. not a strong preference for user keywords). This suggests to us that the keywords from our parser were sufficiently related to make users think again about the keywords which they provided, and in some cases accept that our keywords were at least as good as theirs.

Out of 600 preferences, 61 (10%) were neutral and on 23 (4%) occasions users expressed a direct preference for the keywords generated by our parser. Furthermore, even when users did express a preference for their keywords, in the majority (59%) of cases the keywords generated by our parser were rated as 'relevant' (Table 4.6).

5.1.2 Parser Keywords Compared to User Keywords

As mentioned previously, there is a notable difference between the ratings given to user keywords and those given to parser keywords. In order to obtain an indication of how far the ratings differed, as opposed to merely confirming that a difference exists, we performed the first step of a Wilcoxon Matched Pairs Signed Rank test[2] (Wilcoxon 1945) by comparing the difference between the ratings given to parser keywords and the ratings given to user keywords. This test is suitable for our data as it compares dependent variables within subjects (in this case, two

[1] Although we did not inform users where each set of keywords originated, in such a short experiment we would expect them to remember their keywords even after the randomisation of pages (see Section 4.1.3.4).

[2] The Wilcoxon Matched Pairs Signed Rank test is applied by calculating the difference between a pair of matched values (in this case two ratings given by the same participant for keywords from the same Web page) by subtracting one value from the other. These results are then ranked from highest to lowest, and the sum of these differences can be used to determine whether the positive differences cancel out the effect of the negative differences.

keyword ratings provided by the same participant). The following variables were used in our calculations:

- w is the Web page ID.
- r is the rating number.
- u_r is the rating assigned to the keywords suggested by the participant earlier in the experiment.
- p_r is the rating assigned to the keywords generated by our parser.
- $d_r = p_r - u_r$ - i.e. the difference between two ratings from the same participant for the same Web page (where $-4 \leq d_r \leq 4$).
- D_w is the sum of all the rating differences, d, for a given Web page.

The difference in ratings is used as a measure of success as this ties in with our original assertation in Section 1.2 that the ultimate test of any results is whether users find them to be useful. We treat all instances of $d_r \geq 0$ as a success, because in these cases the keywords from our parser have been given a rating equal to or greater than the keywords suggested by the participant in the first part of the experiment – i.e. our keywords are performing as well as or better than those suggested by the user. Table 5.1 shows the differences in ratings for each table and the number of ratings where $d_r \geq 0$ and $d_r < 0$.

Table 5.1: Differences Between User and Parser Keyword Ratings

Web page ID (w)	$d_r \geq 0$	$d_r < 0$	D_w
1	7	23	-41
2	7	23	-54
3	8	22	-39
4	3	27	-67
5	9	21	-30
6	13	17	-27
7	6	24	-54
8	5	25	-53

9	8	22	-43
10	9	21	-33
11	6	24	-56
12	3	27	-66
13	12	18	-29
14	4	26	-50
15	6	24	-44
16	3	27	-58
17	4	26	-56
18	20	10	-13
19	12	18	-27
20	4	26	-42
Total	149	451	-882
Minimum	3	10	-67
Maximum	20	27	-13

As can be seen from the results above, $D_w < 0$ for all Web pages, which means that overall, participants prefer their keywords for every single page in our experiment. Whilst the keywords from our parser are occasionally rated equally or higher than the keywords suggested by the participant (25% of cases), the difference in rating is more than offset by the cases where the keywords suggested by the participant are rated higher than the keywords generated by our parser.

A particularly notable fact about these results is the magnitude of the participants' preference for the keywords which they suggested. Even for the one page ($w = 18$) where participants rate the parser keywords equally or higher than their own keywords in the majority of cases, the overall sum of the difference in ratings is still negative ($D_{18} = -13$). This means that participants are not only rating their keywords higher than the keywords from our parser, but that the degree of preference is particularly strong, suggesting that participants may have remembered the keywords they submitted and thus rate them highly.

In addition to this, the magnitude of the differences also changes, even when the split between positive and negative differences remains the same. For example, for pages 1 and 2 the number of cases where the difference in ratings is

positive ($d_r \geq 0$) is the same, yet the sum of the differences (D_w) differs. This suggests that the individual Web page has a noticeable effect on the perceived relevance of the keywords, and also supports our results in Section 4.2.2.6, which indicate that there is no correlation between the ratings given to the two sets of keywords.

5.1.3 Keyword Preferences

As discussed in Section 4.2, users were asked to indicate which set of keywords they preferred, in addition to the ratings for each set of keywords. In 132 cases (22%), participants rated our keywords the same as the keywords which they had provided. Table 5.2 below shows the preferences expressed by participants when they provided the same rating for both the user and parser keywords.

Table 5.2: Actual Preference Frequencies

Preference	**Frequency**
Strongly prefer user keywords	25
Slightly prefer user keywords	55
Neutral	42
Slightly prefer parser keywords	9
Strongly prefer parser keywords	1

If participants have genuinely given the same rating to both sets of keywords, we would expect the majority of preference scores to be neutral, with a similar number of slight preferences either side. However, as can be seen in Table 5.2, the distribution of preferences is skewed towards a preference for the user keywords. This suggests to us that participants may be remembering the keywords they entered earlier in the experiment and expressing a slight or even strong preference for them, even though they have given both sets of keywords the same rating.

5.2 Analysis of Link Experiment Results

As detailed in Section 4.3, our links experiment involved presenting participants with three links and asking them to rate each one, indicate which they preferred and provide free-form feedback on the links. These links were generated using the keywords from our previous experiments, though participants were not made aware of this.

5.2.1 Individual Web Pages

From these results (shown in Table 4.11) we can see that there is a significant variation across different Web pages as to the ratings which participants supplied for links generated by both user and parser keywords. Neither sets of keywords produce related results every time, and it would appear that the individual Web page has a significant influence on this – perhaps unsurprisingly given that the keywords are taken from the text of the page.

A further observation is the difference between the opinions of different participants. For example, on page 11 six participants rated the links provided by user keywords as 'not related' and the same number rated them as 'closely related'. Another example of this can be found with the links provided by our parser's keywords on page 15. Eleven participants rated the link provided (the same for all users) as 'same story', whereas ten participants rated it as 'closely related'. These two figures together accounted for all of the ratings expressed, demonstrating an even split of different opinions. This suggests that it may be beneficial to provide links on a per-user basis rather than showing all users the same links.[3]

Finally, we observed that there was a cluster of 'same story' ratings for the parser links on pages 9-12. As the four pages from each site were listed sequentially in our database, this suggests that there may be an issue with extracting keywords from that particular site. We examine this result in more detail in Section 5.2.3.

[3]Utilising user preferences in discussed in more detail in Section 6.5.9.

5.2.2 Parser Links Compared to User Links

As mentioned previously, there is a notable difference between the ratings given to the links generated from user keywords and those produced by our parser. In order to obtain an indication of how far the ratings differed, as opposed to merely confirming that a difference exists, we performed the first step of a Wilcoxon Matched Pairs Signed Rank test by comparing the difference between each pair of link ratings, in the same way as we analysed pairs of keyword ratings in Section 5.1.2. The following variables were used in our calculations:

- w is the Web page ID.
- r is the rating number.
- u_r is the rating assigned to the link generated from user keywords.
- p_r is the rating assigned to the link generated from our parser's keywords.
- $d_r = p_r - u_r$ - i.e. the difference between two ratings from the same participant for the same Web page (where $-4 \leq d_r \leq 4$).
- D_w is the sum of all the rating differences, d, for a given Web page.

We treat all instances of $d_r \geq 0$ as a success, because in these cases the links from our parser have been given a rating equal to or greater than the links generated from the keywords suggested by participants in the first experiment. As with our keywords analysis, the difference in ratings is used as a measure of success because ratings are subjective and this ties in with our original assertation in Section 1.2 that the ultimate test of any results is whether users find them to be useful. Table 5.3 shows the differences in ratings for each table and the number of ratings where $d_r \geq 0$ and $d_r < 0$.

Table 5.3: Differences Between User and Parser Link Ratings

Web page ID (w)	$d_r \geq 0$	$d_r < 0$	D_w
1	12	9	-7
2	19	2	24
3	n/a	n/a	n/a

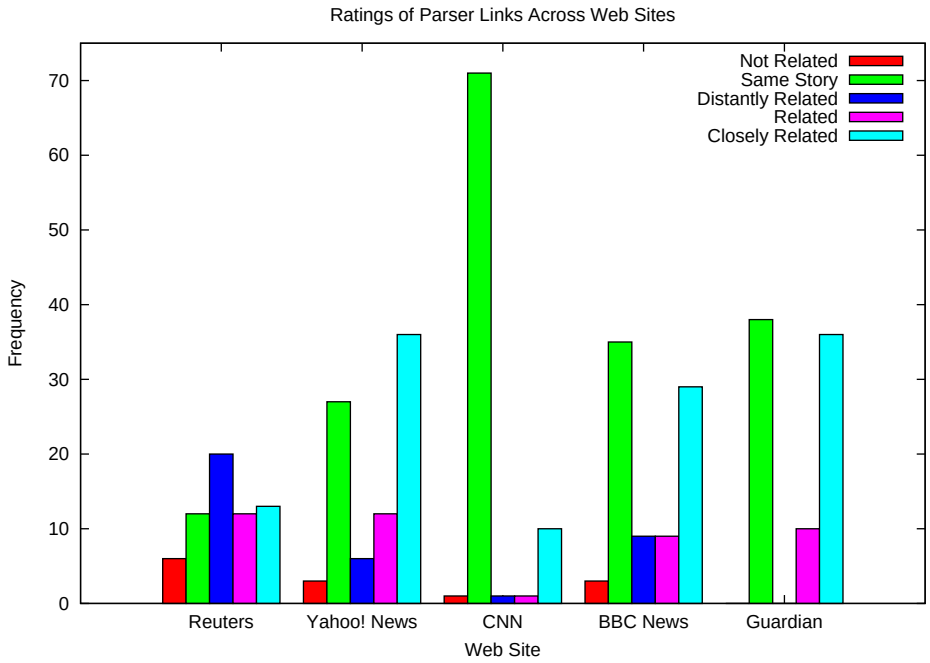
4	11	10	5
5	18	3	36
6	11	10	-4
7	13	8	0
8	11	10	-12
9	20	1	26
10	2	19	-45
11	8	13	-19
12	9	12	-11
13	18	3	15
14	17	4	-1
15	16	5	35
16	2	19	-31
17	20	1	32
18	9	12	-12
19	15	6	8
20	9	12	-3
Total	240	159	26
Minimum	2	1	-45
Maximum	20	19	36

The differences are often large, in most cases the difference is ± 15 , suggesting that our links are performing extremely well on some pages and poorly on others, but there are only a few pages where the sum of the difference in scores is close to zero. As with the modal ratings, discussed in Section 5.2.1, there is a cluster of negative ratings in Web pages 10-12, which are all from the same site (CNN).

5.2.3 Grouped Web Pages

In addition to individual Web pages, we have also analysed the results of pages grouped together from the five sites which we used in our experiments. This was motivated by our observation in Section 5.2.1, where pages 9-12 all received similar scores for the links generated through the keywords from our parser,

Figure 5.1: Ratings for Parser Links in Grouped Web Pages



suggesting that there may be a correlation between the ratings given to keywords on pages and the sites from which they were taken.

From the data shown in Figure 5.1, we can see that the results have a spread over the various different ratings, though 'same story' and 'closely related' make up the majority of ratings for Yahoo! News, BBC News and the Guardian.[4] Pages from Reuters had the most even spread of ratings.

However, the result which immediately stands out from all the others is the number of links generated by our parser from CNN stories which are rated as 'same story'. This suggests three possibilities:

1. Our parser is not capable of picking up proper noun keywords from CNN stories, or;

2. CNN stories do not provide many of the proper noun keywords which our algorithm focuses on, or;

3. CNN stories do contain proper nouns which are picked up by our parser, but these keywords are so specific that when passed into a search engine

[4]As one of the Reuters pages was excluded from our experiments, for reasons explained in Section 4.3.1, its overall rating frequencies are lower.

the only results obtained lead to the same story.

Of these possibilities, the first is unlikely, given that we can clearly see from running our parser over the pages from CNN that it is picking up some keywords from these stories. The same fact also makes the second possibility unlikely, as running our parser over CNN pages clearly produces a list of potential keywords, even if we only choose to use the two most frequently occuring proper nouns in our search engine query. Therefore, we would suggest that the most likely possibility is that the keywords extracted from CNN are so specific that they only return the same story when passed through a search engine.

The specific keywords extracted by our parser from the CNN pages used in our experiments can be seen in Table 5.4. In two of the cases, pages 10 and 11, our parser has extracted a long proper noun phrase which consists of four words. Whilst this is the result we would want and expect – particularly in the case of page 11 where the phrase extracted is the full name of the entity, rather than a name with a title attached – in order for this to be useful in a search engine query the same phrase would need to be used on other Web pages. We suggest that the longer a phrase is, the less likely it is that it will be referred to in the same way across a variety of Web sites. It is perhaps more likely that an abbreviated phrase would be used, for example.

In the cases of pages 9 and 12, it would appear that our parser has extracted generic keywords, so it seems surprising that when these keywords are fed back into a search engine they produce links to either the same story on another Web site, or even exactly the same page (i.e. the same story on CNN). There are three possibilities for why this effect may have occured. One possibility is that CNN writers often use phrases which are not used by writers on other Web sites, and so feeding these phrases into a search engine will often return the same page. Another possibility is that the specific phrases selected, whilst seemingly generic, are only found on pages describing those particular stories. The final possibility is that these two particular pages just happen to use keywords which have not been used on other sites, and this is not representative of CNN pages in general. A more detailed study running a wider selection of pages through our parser would be able to ascertain whether this was indeed the case.

Table 5.4: Parser Keywords from CNN Pages

Web page ID	Proper nouns extracted
9	“Liberation Day”, “South Korean”
10	“Zimbabwe President Robert Mugabe”, “Change MDC”
11	“Species Act”, “National Marine Fisheries Service”
12	“NASA”, “September”

As we consider any ratings which involve some degree of relatedness as a success (the level depending on how related users judge the links to be), we can observe that for all sites other than CNN, our parser successfully produces related links in the majority (62%) of cases. This data is outlined graphically in Figure 5.2.

Even though our parser links perform well in the majority of cases for all sites other than CNN, there is still a noticeable difference between the individual sites, as shown in Table 5.5:

Table 5.5: Related Results for Web Sites

Web Site	Results Rated as ‘Related’
Reuters	71%
Yahoo! News	64%
BBC News	55%
Guardian	55%
CNN	14%

These results suggest that our parser is more effective at picking up keywords on Reuters and Yahoo! News pages than on BBC News and Guardian pages. Alternatively, the keywords which are being picked up are more generic and likely to lead to related pages – or possibly a combination of these two factors. There is clearly some future work which could be performed on our algorithm to improve the keywords selected from pages from BBC News, the Guardian and CNN to ensure that the links returned are rated as highly as those from Yahoo! News and Reuters.

5.2.4 Qualitative Feedback

In addition to the quantitative feedback generated through explicit link ratings and preferences, we also offered participants the option to provide further details regarding how related they found the links to be, through the addition of a box at the bottom of each page into which free-style text could be entered. This was intended to gather additional information which could not be derived from the numerical values obtained through the selection of ratings from a restricted set of options. For example, we can tell from the quantitative data that a participant prefered one link above the others, but not why they chose that link. In theory the reason should be because it appeared to lead to an article describing a related event, but we cannot assume that this would be the case.

The majority of the comments entered simply reflected the ratings and preferences already expressed, for example criticising some of the links suggested (which is accounted for in the rating of 'not related' or 'same story'). Some comments focused on personal preferences, such as 'I don't like the way the quote was used, it is a bit too tabloid style for my preferences' (participant 34, page 15). Given the vagueness and scarcity of these comments, we cannot draw any firm conclusions as to whether our links were useful or not in these cases. However, as some participants have stated a preference for certain types of related links, which may different from other participants, this is an area which could be explored further in the future.[5] Other comments covered factors outside of our control, such as 'I prefer the two line summary in the third link' (participant 34, page 4). As page summaries are influenced by the content creator and the way in which Yahoo! automatically creates a summary, we cannot change this infomration without implementing our own page summary generater.[6]

Finally, whilst a number of comments were submitted, we can see that the majority of comments were left by a small number of users. In total, 52 comments were left out of a possible 399, and these comments came from only 12 participants.[7]

[5] A more detailed discussion of this area can be found in Section 6.5.9.

[6] A detailed study of the effect of result descriptions on searching can be found in Lewandowski (2008).

[7] Full details of all the qualitative feedback can be found in Appendix A, raw data is available in Waring (2009*d*).

5.2.5 Additional Results

As well as the fact that the links generated from keywords produced by our parser often equalled or outperformed those generated from user keywords, there is also the fact that our parser always produced links, whereas this was not the case for every set of user keywords. Out of 600 sets of user keywords, 166 (28%) sets did not produce any results at all when fed through the Yahoo! search API. In the case of Web page 3, every set of user keywords suggested for that page failed to produce a single link.[8]

5.3 Analysis of Combined Results

As we had 30 participants in our keywords experiment but only 21 for our links experiment, we interpolated the results by multiplying each difference (D_w in Table 5.1) by a factor of $\frac{30}{21}$ in order to ensure that we had comparable values. This linear interpolation assumes that had 9 more participants completed the links experiment, they would have provided answers similar to the average of answers given by the 21 other participants overall.

The following variables were used in our calculations:

- w is the Web page ID.
- K_w is the sum of all keyword rating differences for a given Web page - i.e. D_w from Table 5.1.
- L_w is the sum of all link rating differences for a given Web page - i.e. D_w from Table 5.3, after applying our scaling factor.
- P_w is the percentage change from K_w to L_w.

Table 5.6: Differences Between Combined Result Ratings

Web page ID (w)	K_w	L_w	$\lvert L_w - K_w \rvert$	P_w
1	-41	-10	31	76%
2	-54	34.29	88.29	164%
4	-67	7.14	74.14	111%

[8]See Section 4.3.1 for full details.

5	-30	51.43	81.43	271%
6	-27	-5.71	21.29	79%
7	-54	0	54	100%
8	-53	-17.14	35.88	68%
9	-43	37.14	80.14	186%
10	-33	-64.29	31.29	-195%
11	-56	-27.14	28.86	52%
12	-66	-15.71	50.29	76%
13	-29	21.43	50.43	174%
14	-50	-1.43	48.57	97%
15	-44	50	94	214%
16	-58	-44.29	13.71	24%
17	-56	31.43	87.43	156%
18	-13	-17.14	4.14	-132%
19	-27	11.43	38.43	142%
20	-42	-4.29	37.71	90%

As can clearly be seen from the results above, there is often a large gap between the sum of the keyword difference ratings and the sum of the link difference ratings. In particular, on all but two pages $L_w > K_w$, and for all but three pages $P_w > 50\%$. This suggests a perception mismatch between the initial keywords and the end result (links) – in our keywords experiment participants are giving higher ratings to their keywords than those produced by our parser, yet in the links experiment the differences in ratings shift considerably.

5.4 General Discussion

In addition to the results directly relating to our work, we also discovered some further interesting results which may have applications beyond our specific area of research.

One statistic which was particularly noticeable amongst the information gathered during the course of our experiments was the rate at which participants exited the survey before completing all of the questions.

Table 5.7: Participant Dropout Rates

Experiment	Total participants	Completed participants	Dropout Rate
Keywords	62	30	53%
Links	58	21	64%

As we can see from the results in Table 5.7, over half the participants in our keywords experiment dropped out before completing all of the questions, and nearly two thirds of the participants in the links experiment failed to complete all sections.

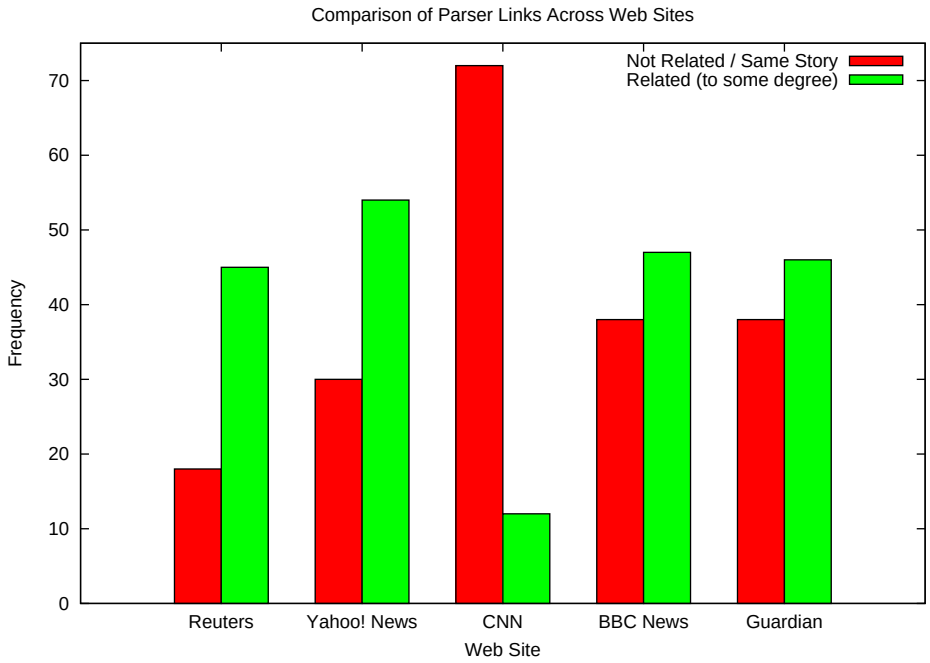
One possible reason for the significant dropout rate could be that participants lost interest part way through the experiment and, having no incentive to continue, decided to stop. However, this does not explain why a higher percentage of users managed to complete the keywords experiment, which had twice as many steps as the links experiment and arguably involved more cognitively complex tasks, given that users had to supply their own data as well as answering pre-determined questions. Furthermore, the keywords experiment had a median completion time of 38 minutes, whereas the median completion time for the links experiment was 33 minutes – approximately 15% lower. In addition, the distribution of completion times for both experiments is similar up until a point, as we can see from Figure 5.3.

In both experiments, we can clearly see some completion times which are considerably higher than the majority of results. We suggest that these outlier results were caused by participants taking a break (perhaps through an unexpected interruption) from the experiment and returning to it later. This tendancy to distraction is an issue which does not arise with observed experiments, where the participant is only able to focus on the tasks in hand, and an interesting piece of future research would to repeat the experiment in supervised conditions to see if this affected the results.

5.5 Conclusions

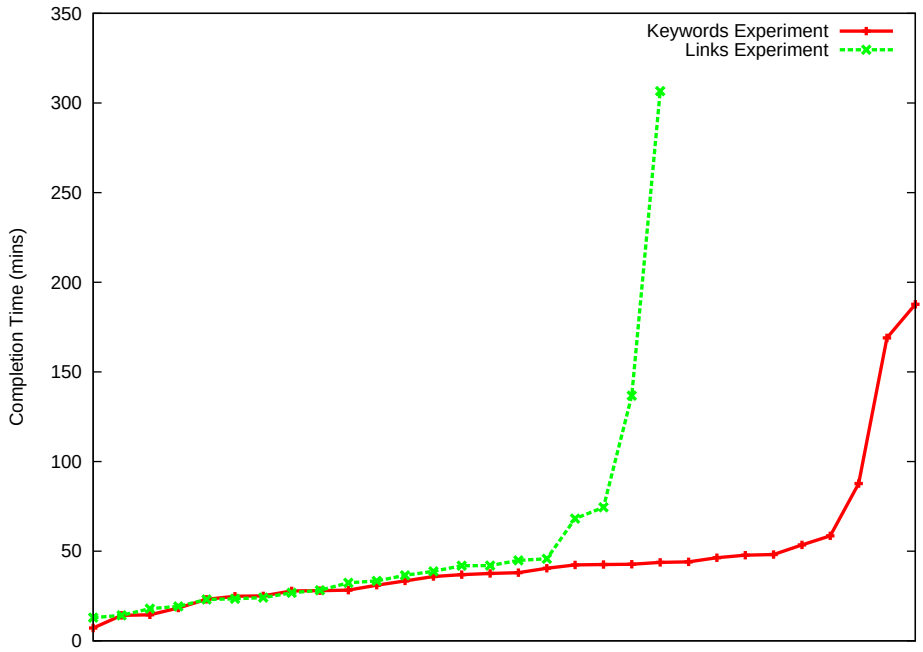
Having analysed the results from our experiments, we can see that our algorithm produces related links in the majority of cases, both in absolute terms and in

Figure 5.2: Comparison of Parser Links Across Web Sites



comparison with links generated from user keywords. In the next chapter, we will summarise our final conclusions and discuss areas of future work which could follow on from the research outlined in this thesis.

Figure 5.3: Completion Times for Keywords and Links Experiments



Chapter 6

Conclusions and Future Work

In this chapter we discuss our overall conclusions and summarise our unique contributions to research. We also present a number of avenues for future work which build upon the research presented in this thesis.

6.1 Summary of Conclusions

Through the analysis of results from our keywords and links experiments, we have been able to reach a number of conclusions based on our research. Firstly, we observed that participants often assigned high rankings to the keywords they had suggested, both in absolute terms and relative to the keywords generated by our algorithm. This suggests that there exists a degree of memory bias when asking participants to provide information which they are later asked to rate or comment on, and that requesting this information can result in skewed preferences in favour of such data.

Secondly, we observed that the ability of our algorithm to produce links to related events varied from page to page, according to ratings from participants. We suggest that the content on a page, both in terms of its language and structure, has a noticeable effect on the performance of any generic algorithm which attempts to extract and manipulate the text. We also observed that this effect appears to be consistent within sites, suggesting that the style of writing used by a given news outlet is part of the effect, in addition to the specific text used on each individual page.

Finally, we observed that there was a considerable difference between the ratings which participants assigned to their keywords and those generated by our

algorithm and the ratings assigned to the links produced by those keywords. We suggest that this indicates a perception mismatch between the keywords which users prefer and their opinions of the links generated from those keywords, and that users may obtain more related results through the use of an algorithm which has been specifically designed to find links to related events.

6.2 Summary of Research Contributions

The research described in this thesis was motivated by our initial observation in Section 1.1 that much of the information describing events on the Web is poorly connected, and as a result opportunities for users to discover information about related events are lost. Our aim was to aid users in their navigation of the vast information space on the Web, and stimulate opportunities for serendipitous discovery of related events.

Throughout this research we have focused on the human factors perspective, as opposed to the standard information retrieval measures of precision and recall. This focus was based on the assertation that the ultimate test of whether retrieved elements (in our case, links to other Web pages) are related is the user's perception of the results. Our results are therefore based on subjective ratings for keywords and links from participants in our experiments.

Our principal contribution is an algorithm which, for five high profile news sites, produces three results which back up our research. First, our algorithm produces links which users would choose over their own in nearly two thirds of cases and which users rate as being at least as related as their own in 60% of cases (Section 4.3.2). This suggests that our algorithm is replicating the manual process employed by users when producing keywords, and on many occasions is improving on the results.

Second, our algorithm produces links which, with the exception of CNN pages, are rated by users as being related to the original story in 55-71% of cases. This suggests that our algorithm is not only performing as well as or better than users when producing links to related events, but that the links are also related in their own right. Our algorithm has therefore satisfied both the standalone and comparative tests outlined in Section 4.3.

Finally, our algorithm always produces at least one link for every page selected for our investigation. This is in contrast to user supplied keywords which only

produced links in 72% of cases (Section 4.3.1). We suggest that this may be because not all users are experts at using search engines (Hölscher & Strube 2000), and so may not be able to produce keywords which return links to related events. However, our algorithm has the advantage of being developed with expert knowledge, effectively providing all users with the same level of search expertise.

In addition to the these three strong results, our algorithm, combined with the system using the Kain Proxy, is fully automated, and will attempt to produce links for the page which the user is viewing with no further interaction required. We have therefore researched and produced a solution which performs at least as well as individual users in producing links to related events (by the subjective measure of whether users find these links to be related) but can operate without any input from the user.

6.3 User Perceptions

As shown in Figure 5.1, the majority of ratings for our parser links on all sites other than Reuters consisted of 'same story' and 'closely related', and these two ratings often have similar frequencies. This raises the question of whether participants were rating links as 'closely related' when the two stories were actually the same, or possibly rating links as 'same story' when in fact they were similar. One objective method of measuring whether a story really was the same as another would be to compare the text of both, although the difficulty of defining at what point should a story be considered 'different' then arises. If two stories contained exactly the same text it would be reasonable to label them as 'same story' – for example, in the case where both copies of the story are taken from the same source such as the Press Association. However, if a news site has edited the text even slightly then the question of whether two stories are the 'same' becomes difficult to answer, as even changing a few words might make the story sufficiently distinct to be considered a different story by some readers.

6.4 Use of Proper Nouns

As discussed in Section 3.3.2, after examining the four attributes of an event we decided to focus on the *who* and the *where*, as the *when* attribute was difficult to extract in a canonical form and the *what* attribute could be vague and ambiguous.

The results of our experiments suggest that these two attributes were sufficient to produce links to related events, as participants preferred the links generated by our parser over the links generated by user keywords in 63% of cases, though these figures may be improved upon through a more advanced parser which can extract the *when* and *what* attributes of an event.

6.5 Future Work

The work presented in this thesis has several applications for future work. In this section, we examine some possibilities for future areas of research which would build upon the work contained within this thesis.

6.5.1 Detecting Multiple Events

One limitation of the work as it currently stands is the assumption made by our algorithm that there is only one event discussed on each page which we parse. Whilst this is a reasonable assumption to make in many cases, particularly for news stories which remain 'on topic', longer articles may discuss multiple events which our algorithm will not detect. As events which are discussed on the same page could have a high probability of being related, it might be useful if we could identify and isolate each event as an individual entity. However, automatically detecting all attributes (*who*, *what*, *where* and *when*) of a single event is a difficult process in itself, and it is not clear how much automation could be employed to detect the boundaries where the text moves from describing one event to another, particularly if these events are closely connected.

6.5.2 Brief Mentions of Related Events

Following on from the detection of multiple events on a single page, there is also a tendancy on some sites to mention related events towards the end of a report on the main story of interest.[1] Often this is in the form of a reference to a similar event in the past, e.g. when a train derailment occurs, it will often be compared – usually in terms of the number of casualties – to previous incidents of a similar nature. Even though these brief mentions may not provide us with enough information to construct a complete description of the event, having the

[1] This effect was particularly noticeable on CNN pages.

ability to detect them would provide us with a starting point when searching for related events. Furthermore, as these events have been selected by the author of the story, they may be more likely to be considered 'related' by other users than the events detected by an automated algorithm trying to pick out keywords from the text.

6.5.3 Reliability of Sites

In our work so far, we have treated all sites the same in terms of analysing them for events. However the quality – in terms of accuracy and reliability – of information found on sites can vary dramatically. Whilst the freedom to access, create and modify the global knowledge base that is the Web has tremendous advantages, it brings with it the problem of inaccurate, misinformed or even deliberately misleading information being made available at the same level as peer-reviewed publications. Another potential issue is the popularity of satirical sites, such as The Onion,[2] which at first glance appear to be reporting real events, but on closer inspection are revealed to be exaggerating or even inventing stories for comedic effect. Unfortunately, no guarantees can be offered as to the accuracy of any information on the Web (Yin et al. 2007), so methods are required to make judgements as to whether a particular piece of material can be trusted to provide correct information.

6.5.4 Querying Multiple Search Engines

One possibility for extending our existing work would be to implement support for multiple search engines when executing queries to discover associated events. Previous work has demonstrated that search engines generally only index a small portion of the Web, and often these slices do not contain extensive amounts of overlap (Bharat & Broder 1998). These studies were performed in the 1990s, and search engines have improved, appeared and consolidated since. However, more recent research suggests that there is still only a small amount of overlap in search results (though not necessarily indexes) returned by the major search engines, and that by using only one search engine a user may be prevented from finding the best results for their query (Spink, Jansen, Blakely & Koshman 2006, Spink, Jansen, Kathuria & Koshman 2006).

[2] `http://www.theonion.com/`

A second reason for using multiple search engines is that the algorithms used for each one are likely to differ, and concentrate on providing results in a particular order. This can easily be demonstrated by entering a simple query on each of the three major search engines - Google, Yahoo! and Bing. Whilst there may be some overlap in the results, it is likely that the results and their ordering will differ between the three sites, as they will not be using identical algorithms for retrieval and ranking.[3] As a result, one search engine might produce more relevant results for one particular query, or type of query, but no search engine will necessarily provide the most relevant results for all queries.

In addition to utilising multiple generic search engines, the possiblity also exists to query other Web services as they become available. For example, the image sharing service Flickr[4] could be used to display photographs related to the events under discussion. Some of these services have already been utilised successfully in the COHSE project (Yesilada, Bechhofer & Horan 2008), which part of our work is based on. Furthermore, the design of our parser is such that the keyword extraction process is separated from the link generation process, so the ability to query multiple search engines and other services with the same keywords used in our Yahoo! search already exists. However, some work would be required to combine the results into one set of links to display to the user, ensuring that duplicates were removed and that the final results were sorted in order of relevance (e.g. by following links to pages, as discussed in Section 6.5.6).

6.5.5 Automatically Detecting Content Blocks

At present, our lightweight platform relies upon a configuration file to detect the main content block of a page and to strip out unnecessary blocks within the content such as embedded advertising. Adding a new site is a simple matter of visiting the site and inserting some lines of XML into the existing configuration,[5] and this single entry will generally cover an entire news site. However, whilst this task is less effort than the creation of an ontology required by SADIe, it is still a somewhat laborious and manual process which has to be performed whenever new sites appear or existing sites change the markup of their pages. Ideally, we

[3] If search engines were using identical algorithms, the main method of competition in the search engine market would be focused on index size and retrieval speed – both factors which can be significantly influenced by purchasing large data centres and network capacity.

[4] `http://flickr.com/`

[5] Full technical details are available in Waring (2009*a*)

would like to be able to automatically detect the content block on any page, extract the text and remove any unnecessary information (e.g. advertisements) with no human intervention or hints. Some work has already been conducted in this area (Lin & Ho 2002, Gupta et al. 2003, Mantratzis et al. 2005, Mantratzis & Cassidy 2005) but there is room for improvement.

6.5.6 Following Links to Pages

At present, the results from our search queries are presented to users in the order which they are returned by Yahoo!'s API. Whilst this approach enables us to leverage the existing precision and recall of the Yahoo! search engine and build upon the work of a significant number of engineers, the rankings provided are not necessarily focused on event-related information. As a result, it is sometimes the case that the most relevant pages from our perspective – i.e. those which contain the most information about related events – are not returned in the top five results which we present to the user. In addition, even when related events are present in the results, they are not necessarily ranked in order of event-related information.

One way of improving the ranking system would be to use the first twenty results returned by Yahoo (or multiple search engines if we followed the approach suggested in Section 6.5.4) in response to our search query as a starting point. The next step would be to fetch each of these pages and analyse them for event-related information. Based on our initial analysis of the original page which the user was browsing, we could then apply an additional algorithm to reorder the links using a measure of how related the events on the result pages were to the original event.

6.5.7 Eliminating Duplicate Descriptions of Events

In a number of cases, we observed that results returned by our search queries contained links to pages with identical, or highly similar, content – a problem which is a continuing issue for any work involving infomration retrieval (Cooper et al. 2002). This was particularly noticeable with stories from Reuters which syndicates the text of its news stories in addition to publishing them on its own Web site. As a result, some news sites simply contain the same news story as the Reuters site, occasionally with some minor modifications. As the individual pages

are not identical, we would need to develop a method for detecting similar content with a given threshold, or implement an existing solution – e.g. by applying fuzzy logic (Koberstein & Ng 2006), looking at content (Manku et al. 2007) or by examining URLs (Bar-Yossef et al. 2007).

6.5.8 Support for Multiple Languages

Due to the time constraints involved in this project, we decided to focus exclusively on sites which were written in the English language. In addition, as mentioned in Section 4.1.1.2, all of the top ten news sites provided by Alexa were actually written in English. However, whilst English is the most popular language on the Web, there is still a significant and growing amount of information out there which is written in other languages (Grefenstette & Nioche 2000) and therefore it would be useful to extend HuCEL to support languages other than English. Furthermore, research has already been undertaken and tools are already available for identifying the language of a given document with varying degrees of accuracy (Martins & Silva 2005), so it would be possible for pages returned by search engines to be checked to determine whether they were composed in one of the supported languages. In addition, these existing tools could also be used to determine the language of the Web page which the user is currently browsing and execute a search query which only returned results in the same language.

6.5.9 Remembering User Preferences

At present, our system does not take into account any user preferences – all users will be shown the same links if looking at a page at the same time. However, analysis of the qualitative feedback from our links experiment in Section 5.2.4 demonstrated that, perhaps unsurprisingly, some users have different preferences for the links which they would like to see displayed. These preferences could be based on reading levels, political persuasions or other measures.

6.5.10 Incorporating Temporal Information

Whilst our existing work focuses on the people and places attributes of events, we took a concious decision not to utilise the aspect of time when extracting events (Section 3.3.2.3). Some existing work has already been undertaken on the problem of specifying and annotating temporal expressions in text, most notably as part of

the TimeML project,[6] but this appears not to have been followed up with further research. Automatically annotating pages with TimeML is acknowledged, even by its authors, as a hard problem, particularly due to the fact that 'many temporal expressions are not fully specified and require additional information from other temporal expressions to provide their full value' (Boguraev et al. 2007). The alternative, manually annotating pages, even with the assistance of tools, requires a significant investment of time and does not keep up to date with changes made to the pages after annotation (Takagi et al. 2002).

6.5.11 Utilising Existing NLP Toolkits

As discussed in Section 3.2, we made a concious decision early on in our research to write our own lightweight parser, as opposed to utilising an existing NLP solution. Our reasons for this were primarily related to the amount of time required to evaluate the applicability of the numerous existing solutions to our work, balanced against the short timeframe of the HuCEL project. However, as our parser already successfully extracts the main content of a Web page and can break this text up into sentences, there is no reason why this text could not be passed to a NLP toolkit for parsing. Some work would then be required to analyse the results obtained and integrate them with the rest of our system. This solution could also be compared with our parser by running the links experiment again with data produced by our parser and the NLP solution.

6.6 Final Conclusions

In the process of this research, we have developed an algorithm which produces links to related events, by extracting keywords from the text of a news Web page. Through our experiments, we have shown that users find these links to be related, both in their own right and in comparison to the links generated by user keywords. We have also extended an existing project to provide a framework which allows our results to be automatically presented to the user whenever a supported Web site[7] is visited, with no manual intervention required.

[6] `http://www.timeml.org/`

[7] A supported Web site is one for which we have created a content mapping entry, as described in Waring (2009*a*).

Finally, the research described within this thesis has produced multiple avenues for further work, ranging from incremental improvements such as querying multiple search engines, to new areas of research which could form an entire project in themselves (e.g. incorporating temporal information when searching for related events).

Bibliography

Allan, J. (2002), Introduction to topic detection and tracking, *in* J. Allan, ed., 'Topic Detection and Tracking: Event-based Information Organization', The Kluwer International Series on Information Retrieval, Kluwer Academic Publishers, pp. 1–16.

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S. & Amstutz, P. (2005), Taking topic detection from evaluation to practice, *in* 'Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)', Vol. 4, IEEE Computer Society.

Allan, J., Papka, R. & Lavrenko, V. (1998), On-line new event detection and tracking, *in* 'SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, pp. 37–45.

Ashman, H. (2000), 'Electronic document addressing: dealing with change', *ACM Computing Surveys* **32**(3), 201–212.

Ashman, H., Garrido, A. & Oinas-Kukkonen, H. (1997), Hand-made and computed links, precomputed and dynamic links, *in* 'Proceedings of Hypermedia–Information Retrieval–Multimedia '97 (HIM '97)', pp. 191–208.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press.

Bar-Yossef, Z., Broder, A. Z., Kumar, R. & Tomkins, A. (2004), Sic transit gloria telae: towards an understanding of the web's decay, *in* 'WWW '04: Proceedings of the 13th international conference on World Wide Web', ACM, New York, pp. 328–337.

Bar-Yossef, Z., Keidar, I. & Schonfeld, U. (2007), Do not crawl in the dust: different urls with similar text, *in* 'WWW '07: Proceedings of the 16th international conference on World Wide Web', ACM, New York, pp. 111–120.

Bechhofer, S., Yesilada, Y., Horan, B. & Goble, C. (2006), Knowledge-driven hyperlinks: Linking in the wild, *in* 'Adaptive Hypermedia and Adaptive Web-Based Systems', Vol. 4018 of *Lecture Notes in Computer Science*, Springer, pp. 1–10.

Bharat, K. & Broder, A. (1998), 'A technique for measuring the relative size and overlap of public web search engines', *Computer Networks and ISDN Systems* **30**, 379–388.

Bizer, C., Heath, T. & Berners-Lee, T. (2009), 'Linked data - the story so far', *International Journal on Semantic Web and Information Systems* **5**(3), 1–22.

Boguraev, B., Pustejovsky, J., Ando, R. & Verhagen, M. (2007), 'Timebank evolution as a community resource for timeml parsing', *Language Resources and Evaluation* **41**(1), 91–115.

Brants, T. & Chen, F. (2003), A system for new event detection, *in* 'SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval', ACM, New York, pp. 330–337.

Bry, F. & Eckert, M. (2005), Processing link structures and linkbases in the web's open world linking, *in* 'HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia', ACM, New York, pp. 135–144.

Carmel, E., Crawford, S. & Chen, H. (1992), 'Browsing in hypertext: a cognitive study', *IEEE Transactions on Systems, Man and Cybernetics* **22**(5), 865–884.

Carr, L., Hall, W., Bechhofer, S. & Goble, C. (2001), Conceptual linking: ontology-based open hypermedia, *in* 'WWW '01: Proceedings of the 10th international conference on World Wide Web', ACM, New York, pp. 334–342.

Catledge, L. D. & Pitkow, J. E. (1995), 'Characterizing browsing strategies in the world-wide web', *Computer Networks and ISDN Systems* **27**(6), 1065–1073.

Chakrabarti, D., Kumar, R. & Tomkins, A. (2006), Evolutionary clustering, *in* 'KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 554–560.

Chen, B. & Shen, V. Y. (2006), Transforming web pages to become standard-compliant through reverse engineering, *in* 'W4A: Proceedings of the 2006 international cross-disciplinary workshop on Web accessibility (W4A)', ACM, New York, pp. 14–22.

Chen, S., Hong, D. & Shen, V. Y. (2005), An experimental study on validation problems with existing HTML webpages, *in* 'Proceedings of the 2005 International Conference on Internet Computing', pp. 373–379.

Cho, J. & Garcia-Molina, H. (2000), The evolution of the web and implications for an incremental crawler, *in* 'VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases', Morgan Kaufmann Publishers Inc., San Francisco, pp. 200–209.

Cooper, J. W., Coden, A. R. & Brown, E. W. (2002), Detecting similar documents using salient terms, *in* 'CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management', ACM, New York, pp. 245–251.

Cove, J. F. & Walsh, B. C. (1988), 'Online text retrieval via browsing', *Information Processing & Management* **24**(1), 31–37.

Dalal, Z., Dash, S., Dave, P., Francisco-Revilla, L., Furuta, R., Karadkar, U. & Shipman, F. (2004), Managing distributed collections: evaluating web page changes, movement, and replacement, *in* 'JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries', ACM, New York, pp. 160–168.

Dale, R. & Mazur, P. (2007), The semantic representation of temporal expressions in text, *in* 'AI 2007: Advances in Artificial Intelligence', Vol. 4830 of *Lecture Notes in Computer Science*, Springer, pp. 435–444.

Davis, H. (1995), 'To embed or not to embed...', *Communications of the ACM* **38**(8), 108–109.

Delgadillo, R. & Lynch, B. P. (1999), 'Future historians: Their quest for information', *College and Research Libraries* **60**(3), 245–259.

Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M. & Schilling, L. M. (2003), 'Going, going, gone: Lost internet references', *Science* **302**(5646), 787–788.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A. & Zien, J. Y. (2003), SemTag and seeker: bootstrapping the semantic web via automated semantic annotation, *in* 'WWW '03: Proceedings of the 12th international conference on World Wide Web', ACM, pp. 178–186.

El-Beltagy, S. R., Hall, W., Roure, D. D. & Carr, L. (2001), Linking in context, *in* 'HYPERTEXT '01: Proceedings of the twelfth ACM conference on Hypertext and Hypermedia', ACM, New York, pp. 151–160.

Feng, A. & Allan, J. (2007), Finding and linking incidents in news, *in* 'CIKM '07: Proceedings of the sixteenth ACM Conference on information and knowledge management', ACM, pp. 821–830.

Fetterly, D., Manasse, M., Najork, M. & Wiener, J. (2003), A large-scale study of the evolution of web pages, *in* 'WWW '03: Proceedings of the 12th international conference on World Wide Web', ACM, New York, pp. 669–678.

Fogelson, R. D. (1989), 'The ethnohistory of events and nonevents', *Ethnohistory* **36**(2), 133–147.

Ford, N. (2000), 'Improving the darkness to light ratio in user-related information retrieval research', *Journal of Documentation* **56**(6), 624–643.

Foster, A. & Ford, N. (2003), 'Serendipity and information seeking: an empirical study', *Journal of Documentation* **59**(3), 321–340.

Gibson, D. (2004), The site browser: catalyzing improvements in hypertext organization, *in* 'HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext and hypermedia', ACM, New York, pp. 68–76.

Graham, T., Stewart, H., Kopaee, A., Ryman, A. & Rasouli, R. (1999), A world-wide-web architecture for collaborative software design, *in* 'Proceedings of Software Technology and Engineering Practice (STEP '99)', pp. 22–29.

Grefenstette, G. & Nioche, J. (2000), Estimation of English and non-English language use on the WWW, *in* 'Recherche d'Information Assiste par Ordinateur (RIAO)'.

Gulli, A. & Signorini, A. (2005), The indexable web is more than 11.5 billion pages, *in* 'WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web', ACM, New York, pp. 902–903.

Gupta, S., Kaiser, G., Neistadt, D. & Grimm, P. (2003), DOM-based content extraction of HTML documents, *in* 'WWW '03: Proceedings of the 12th international conference on World Wide Web', ACM, New York, pp. 207–214.

Halpin, H. & Moore, J. D. (2006), Event extraction in a plot advice agent, *in* 'ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL', Association for Computational Linguistics, New Jersey, pp. 857–864.

Harper, S., Bechhofer, S. & Lunn, D. (2006*a*), Sadie:: transcoding based on css, *in* 'ASSETS '06: Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility', ACM, New York, pp. 259–260.

Harper, S., Bechhofer, S. & Lunn, D. (2006*b*), Taming the inaccessible web, *in* 'SIGDOC '06: Proceedings of the 24th annual ACM international conference on Design of communication', ACM, New York, pp. 64–69.

Hattori, G., Hoashi, K., Matsumoto, K. & Sugaya, F. (2007), Robust web page segmentation for mobile terminal using content-distances and page layout information, *in* 'WWW '07: Proceedings of the 16th international conference on World Wide Web', ACM, New York, pp. 361–370.

Hayes, J. H. & Dekhtyar, A. (2005), Humans in the traceability loop: can't live with 'em, can't live without 'em, *in* 'TEFSE '05: Proceedings of the 3rd international workshop on Traceability in emerging forms of software engineering', ACM, New York, pp. 20–23.

Hobbs, J. R. & Pan, F. (2004), 'An ontology of time for the semantic web', *ACM Transactions on Asian Language Information Processing* **3**(1), 66–85.

Hölscher, C. & Strube, G. (2000), 'Web search behavior of internet experts and newbies', *Computer Networks* **33**, 337–346.

Jul, S. & Furnas, G. W. (1997), 'Navigation in electronic worlds', *SIGCHI Bulletin* **29**(4), 44–49.

Kao, A. & Poteet, S. R. (2007), Overview, *in* 'Natural Language Processing and Text Mining', Springer.

Koberstein, J. & Ng, Y.-K. (2006), Using word clusters to detect similar web documents, *in* 'Knowledge Science, Engineering and Management', Vol. 4092 of *Lecture Notes in Computer Science*, Springer, pp. 215–228.

Krishnamurthy, B., Malandrino, D. & Wills, C. E. (2007), Measuring privacy loss and the impact of privacy protection in web browsing, *in* 'SOUPS '07: Proceedings of the 3rd symposium on Usable privacy and security', ACM, pp. 52–63.

Kumaran, G. & Allan, J. (2004), Text classification and named entities for new event detection, *in* 'SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, pp. 297–304.

Lam, W., Meng, H. M. L., Wong, K. L. & Yen, J. C. H. (2001), 'Using contextual analysis for news event detection', *International Journal of Intelligent Systems* **16**(4), 525–546.

Larkey, L. S., Ogilvie, P., Price, M. A. & Tamilio, B. (2000), Acrophile: an automated acronym extractor and server, *in* 'DL '00: Proceedings of the fifth ACM conference on Digital libraries', ACM, New York, pp. 205–214.

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V. & Thomas, S. (2002), Relevance models for topic detection and tracking, *in* 'Proceedings of the second international conference on Human Language Technology Research', Morgan Kaufmann Publishers Inc., pp. 115–121.

Lawrence, S. & Giles, C. L. (1998), 'Searching the world wide web', *Science* **280**(5360), 98–100.

Lawrence, S. & Giles, C. L. (1999), 'Accessibility of information on the web', *Nature* **400**(6740), 107–109.

Lewandowski, D. (2008), 'The retrieval effectiveness of web search engines: considering results descriptions', *Journal of Documentation* **64**(6), 915–937.

Li, Z., Wang, B., Li, M. & Ma, W.-Y. (2005), A probabilistic model for retrospective news event detection, *in* 'SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, pp. 106–113.

Liestman, D. (1992), 'Chance in the midst of design: approaches to library research serendipity', *RQ* **31**(4), 524–532.

Lin, S.-H. & Ho, J.-M. (2002), Discovering informative content blocks from web documents, *in* 'KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, pp. 588–593.

Liu, B. (2005), *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Data-Centric Systems and Applications, Springer.

Lunn, D. (2008), 'Building Ontologies For The SADIe Transcoder', Technical Report, University of Manchester, `http://hcw-eprints.cs.man.ac.uk/63/`.

Lynch, P. & Horton, S. (1997), 'Imprudent linking weaves a tangled web', *Computer* **30**(7), 115–117.

Makkonen, J. & Ahonen-Myka, H. (2003), Utilizing temporal information in topic detection and tracking, *in* 'Research and Advanced Technology for Digital Libraries', Lecture Notes in Computer Science, Springer, pp. 393–404.

Makkonen, J., Ahonen-Myka, H. & Salmenkivi, M. (2002), Applying semantic classes in event detection and tracking, *in* R. Sangal & S. M. Bendre, eds, 'Proceedings of International Conference on Natural Language Processing (ICON 2002)', Mumbai, India, pp. 175–183.

Makkonen, J., Ahonen-Myka, H. & Salmenkivi, M. (2003), Topic detection and tracking with spatio-temporal evidence, *in* 'Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003. Proceedings', Lecture Notes in Computer Science, Springer, pp. 251–265.

Mani, I., Schiffman, B. & Zhang, J. (2003), Inferring temporal ordering of events in news, *in* 'NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology', Association for Computational Linguistics, pp. 55–57.

Mani, I. & Wilson, G. (2000), Robust temporal processing of news, *in* 'ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, pp. 69–76.

Manku, G. S., Jain, A. & Sarma, A. D. (2007), Detecting near-duplicates for web crawling, *in* 'WWW '07: Proceedings of the 16th international conference on World Wide Web', ACM, pp. 141–150.

Mantratzis, C. & Cassidy, S. (2005), DOM-Based XHTML Document Structure Analysis Separating Content from Navigation Elements, *in* 'CIMCA '05: Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce', IEEE Computer Society, Washington, pp. 632–637.

Mantratzis, C., Orgun, M. & Cassidy, S. (2005), Separating XHTML content from navigation clutter using DOM-structure block analysis, *in* 'HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia', ACM, New York, pp. 145–147.

Markwell, J. & Brooks, D. W. (2002), 'Broken links: The ephemeral nature of educational www hyperlinks', *Journal of Science Education and Technology* **11**(2), 105–108.

Martins, B. & Silva, M. J. (2005), Language identification in web pages, *in* 'SAC '05: Proceedings of the 2005 ACM symposium on Applied computing', ACM, New York, pp. 764–768.

McLachlan, S. & Golding, P. (2000), Tabloidization in the British Press: A Quantitative Investigation into Changes in British Newspapers, 1952-1997, *in* C. Sparks & J. Tulloch, eds, 'Tabloid Tales: Global Debates over Media Standards', Rowman & Littlefield, Oxford, pp. 75–90.

Nakahira, K. T., Matsui, M. & Mikami, Y. (2007), The use of xml to express a historical knowledge base, *in* 'WWW '07: Proceedings of the 16th international conference on World Wide Web', ACM, New York, pp. 1345–1346.

Nallapati, R., Feng, A., Peng, F. & Allan, J. (2004), Event threading within news topics, *in* 'CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management', ACM, New York, pp. 446–453.

Ntoulas, A., Cho, J. & Olston, C. (2004), What's new on the web?: the evolution of the web from a search engine perspective, *in* 'WWW '04: Proceedings of the 13th international conference on World Wide Web', ACM, New York, pp. 1–12.

Petras, V., Larson, R. R. & Buckland, M. (2006), Time period directories: a metadata infrastructure for placing events in temporal and geographic context, *in* 'JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries', ACM, New York, pp. 151–160.

Ramasubramanian, V. & Sirer, E. G. (2004), The design and implementation of a next generation name service for the internet, *in* 'SIGCOMM '04: Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications', ACM, New York, pp. 331–342.

Rattenbury, T., Good, N. & Naaman, M. (2007), Towards automatic extraction of event and place semantics from flickr tags, *in* 'SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York.

Rosenthal, R. (1966), *Experimenter Effects in Behavioral Research*, Century Psychology Series, Appleton-Century-Crofts, New York.

Rowe, N. C. & Laitinen, K. (1995), 'Semiautomatic disabbreviation of technical text', *Information Processing & Management* **31**(6), 851–857.

Salomon, G. B. (1990), Designing casual-user hypertext: the CHI'89 InfoBooth, *in* 'CHI '90: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, New York.

Saracevic, T. (1997), 'Users lost: reflections on the past, future, and limits of information science', *SIGIR Forum* **31**(2), 16–27.

Scholes, R. (1980), 'Language, narrative, and anti-narrative', *Critical Inquiry* **7**(1), 204–212.

Sellen, A. J., Murphy, R. & Shaw, K. L. (2002), How knowledge workers use the web, *in* 'CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, New York, pp. 227–234.

Smith, D. A. (2002), Detecting and browsing events in unstructured text, *in* 'SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 73–80.

Spinellis, D. (2003), 'The decay and failures of web references', *Communications of the ACM* **46**(1), 71–77.

Spink, A., Jansen, B. J., Blakely, C. & Koshman, S. (2006), 'A study of results overlap and uniqueness among major web search engines', *Information Processing & Management* **42**(5), 1379–1391.

Spink, A., Jansen, B. J., Kathuria, V. & Koshman, S. (2006), 'Overlap among major web search engines', *Internet Research* **16**(4), 419–426.

Stelmaszewska, H. & Blandford, A. (2004), 'From physical to digital: a case study of computer scientists behaviour in physical libraries', *International Journal on Digital Libraries* **4**(2), 82–92.

Takagi, H., Asakawa, C., Fukuda, K. & Maeda, J. (2002), Site-wide annotation: reconstructing existing pages to be accessible, *in* 'Assets '02: Proceedings of the fifth international ACM conference on Assistive technologies', ACM, New York, pp. 81–88.

Tauscher, L. & Greenberg, S. (1997), 'How people revisit web pages: empirical findings and implications for the design of history systems', *International Journal of Human-Computer Studies* **47**(1), 97–137.

Teevan, J., Alvarado, C., Ackerman, M. S. & Karger, D. R. (2004), The perfect search engine is not enough: a study of orienteering behavior in directed search, *in* 'CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, New York, pp. 415–422.

Terada, A., Tokunaga, T. & Tanaka, H. (2002), 'Automatic expansion of abbreviations by using context and character information', *Information Processing & Management* **40**(1), 31–45.

Thakor, M. V., Borsuk, W. & Kalamas, M. (2004), 'Hotlists and web browsing behavior–an empirical investigation', *Journal of Business Research* **57**(7), 776–786.

Vendler, Z. (1967), *Linguistics in Philosophy*, Cornell University Press.

Waring, P. (2009*a*), 'Deploying the HuCEL Java Servlet', Technical Report, University of Manchester, `http://hcw-eprints.cs.man.ac.uk/111/`.

Waring, P. (2009*b*), 'HuCEL: Keywords Experiment II Manual', Technical Report, University of Manchester, `http://hcw-eprints.cs.man.ac.uk/89/`.

Waring, P. (2009*c*), 'HuCEL: Keywords Experiment Manual', Technical Report, University of Manchester, `http://hcw-eprints.cs.man.ac.uk/88/`.

Waring, P. (2009*d*), 'HuCEL: Links Experiment Manual', Technical Report, University of Manchester, `http://hcw-eprints.cs.man.ac.uk/90/`.

Wei, C.-P. & Lee, Y.-H. (2004), 'Event detection from online news documents for supporting environmental scanning', *Decision Support Systems* **36**(4), 385–401.

Weinreich, H., Obendorf, H., Herder, E. & Mayer, M. (2008), 'Not quite the average: An empirical study of web use', *ACM Transactions on the Web* **2**(1), 1–31.

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', *Biometrics Bulletin* **1**(6), 80–83.

Wilkinson, R. & Smeaton, A. F. (1999), 'Automatic link generation', *ACM Computing Surveys* **31**(4), 1–4.

Wren, J. D. (2004), '404 not found: the stability and persistence of urls published in medline', *Bioinformatics* **20**(5), 668–672.

Yan, T. W., Jacobsen, M., Garcia-Molina, H. & Dayal, U. (1996), 'From user access patterns to dynamic hypertext linking', *Computer Networks and ISDN Systems* **28**, 1007–1014.

Yang, Y., Ault, T., Pierce, T. & Lattimer, C. W. (2000), Improving text categorization methods for event tracking, *in* 'SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, pp. 65–72.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B. & Liu, X. (1999), 'Learning approaches for detecting and tracking news events', *Intelligent Systems and Their Applications* **14**(4), 32–43.

Yang, Y., Pierce, T. & Carbonell, J. (1998), A study of retrospective and on-line event detection, *in* 'SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, pp. 28–36.

Yesilada, Y., Bechhofer, S. & Horan, B. (2008), Dynamic linking of web resources: Customisation and personalisation, *in* 'Advances in Semantic Media Adaptation and Personalization', Vol. 93 of *Studies in Computational Intelligence*, Springer, pp. 1–24.

Yesilada, Y., Lunn, D. & Harper, S. (2007), Experiments toward reverse linking on the web, *in* 'HT '07: Proceedings of the 18th conference on Hypertext and hypermedia', ACM, New York, pp. 3–10.

Yesilada, Y., Lunn, D. & Harper, S. (2008), 'Experiments toward reverse linking on the web', Experiment, University of Manchester, `http://hcw-eprints.cs.man.ac.uk/91/`.

Yin, X., Han, J. & Yu, P. S. (2007), Truth discovery with multiple conflicting information providers on the web, *in* 'KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, pp. 1048–1052.

Yngve, V. H. (1955), Syntax and the problem of multiple meaning, *in* W. N. Locke & A. D. Booth, eds, 'Machine Translation of Languages', Wiley, pp. 208–226.

Yu, B.-M. & Roh, S.-Z. (2002), 'The effects of menu design on information-seeking performance and user's attitude on the world wide web', *Journal of the American Society for Information Science and Technology* **53**(11), 923–933.

Zacks, J. M. & Tversky, B. (2001), 'Event structure in perception and conception', *Psychological Bulletin* **127**(1), 3–21.

Zhang, K., Zi, J. & Wu, L. G. (2007), New event detection based on indexing-tree and named entity, *in* 'SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, pp. 215–222.

Appendix A

Qualitative Feedback

Table A.1: Qualitative Feedback from Links Experiment

Web page ID	**Participant ID**	**Feedback**
2	10	None of these links really seem relevant
11	10	The first link looks to be the same page as the one shown, so I've chosen the second link
10	10	Second link appears to be exactly the same as the page shown, so have chosen first link
1	10	None of the links are particularly relevant
16	10	Have chosen 2nd link even though it's not really relevant, because first link appears to be pointing to the same page that's shown above
17	11	The original story seems to indicate an ending, but the second link, I feel, is probably closest to the content of the original article.

2	11	Although I feel none of these directly relate to the original article, the third link is the closest in similar phrases (e.g. Beijing, Sudan, Olympics, China)
13	11	The first link, although not the same article, seems the closest to the Guardian report
1	11	Although I thought all three links were distantly related (seeing as they all mentioned Barack Obama) I would pick the first link over the others as it seems the closest distantly related one!
10	11	It seems that the third link is the same story - it has the same sub-heading. The second link seems like one that is related albeit that the message is different (i.e. that the power-sharing element was contested more).
15	11	The second link seems to be the same story. The others seem random.
16	11	The first link seems to be the exact match to the original article.
8	11	Although the second link seems not related at all, if one was to click on it it might present the story as Top News.
7	11	I would find it hard to choose between the two closely related links. At a pinch, I'd choose the first.

18	11	I'd probably follow the first link to discover more about the UK composition
6	11	I'd probably check out the first link
12	11	I'd probably follow the first link to update knowledge about developments in Science
14	11	I would follow the second link to learn more about the effects on food and consumers
19	11	I'd follow the first story to learn more of the arrests of the protesters.
10	14	it depends what I was searching for! I don't think that's a fair question to have to chose which link to find more...
15	14	first 2 links were same
19	14	1st two are same link
14	14	wouldn't follow any
5	14	last 2 links identical
2	19	First and second links are identical
19	19	Links two and three are identical
12	19	First and second links are identical
8	25	I'd probably check two AND three.
14	25	Any / all of them to be honest.
17	25	All or any to be honest!
2	25	I wouldn't follow any of them!
9	29	Not very good links

2	29	Not very related links, interesting though
4	34	If I wanted additional information I would choose the third link as the second link essentially contains the same information. I prefer the two line summary in the third link.
14	34	I would choose the third link as it relates to the UK news whereas the others seem to relate to other countries.
15	34	I would choose the first link as it is the most relevant but if there was another link to the same story I might choose that as I don't like the way the quote was used, it is a bit too tabloid style for my preferences.
15	36	first 2 stories are the same
18	37	links 1 and 2 the same
7	37	links 1 and 2 the same
14	46	No real preference. All about economy/inflation, but not in UK.
20	52	I feel like I'm doing an English GCSE!
19	52	Are the first and last links meant to be the same or is that to see if I'm paying attention to what I'm doing?
9	52	Are any of those related or is it just me?

7	54	There was a lot of text to scan quickly!
17	54	First link predates article, second link follows up
4	54	First link appears to add to the information in the article
2	54	I probably wouldn't have followed any of the links.
18	54	Link 1 is a follow-up article
16	54	2nd and 3rd links were the same
15	54	I probably wouldn't have followed up
12	54	Probably wouldn't have followed any link
9	54	Probabl wouldn't have followed any link