Initial research discussion

PAUL WARING
School of Computer Science, University of Manchester

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 and 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 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 and 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. Others 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].

[1] "e.g. an airplane crash, and earthquake, governmental elections, etc." [Brants and Chen 2003]

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 and Allan 2007; Zhang et al. 2007].

The definition provided by Yang et al. [1998] is slightly more specific than that of Allan et al. [1998], as it brings in the concept of an event being 'non-trivial', and, as might be expected, this definition raises the question of 'how does one define non-trivial?' This is a difficult point to address as the measure of triviality is purely subjective, and what is trivial to one researcher might be of the utmost importance to another. One would hope that within a group there would be some form of agreement as to which events are 'obviously trivial' and which are of importance, and that the subjectivity factor would only come into play for the events which fit in between these two extremes.[2] The introduction of subjectivity also potentially creates problems for the automatic identification of events, because this human bias needs to be transformed into some form of machine weighting. For example, any event involving the keyword 'death' might have its balance automatically weighted heavily towards importance and away from triviality.

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 [1980] 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 [1980] 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 [1989] 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 possibly 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 and Ahonen-Myka [2003] state that 'clearly, an event as well as the news-stream itself are intrinsically sensitive to time.' Furthermore, whilst the

[2]However, even the labels of 'obviously trivial' and 'obviously important' contain a degree of subjectivity within the group of researchers who apply them to events.

Paul Waring – pwaring@cs.man.ac.uk

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 [1967] 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' [Vendler 1967, p.144] and thus do not exist in space and cannot be said to occur in a particular location. Rattenbury et al. [2007] also 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 and 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 [2002] suggests that this enables users 'to browse document collections by the common and well-understood dimensions of time and space.' Smith [2002] 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 [2002] specifically draws attention to this, stating that 'despite the definition of an event, however, as occuring in a certain place, most TDT[3] systems do not directly take geographical location into account.'

Looking at the work from previous scholars so far, 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, most scholars also mention the location

[3]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/`

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, the 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,[4] for the purposes of identifying social change 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 significant number of papers which do 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 and Lee [2004] agree, declaring that a news story generally reports event properties including 'when the even 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 perhaps of less interest to our work, as we are not initially interested in connecting cause and effect with regards to events, the other four element 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.' Furthermore, they suggest that each of these attributes can be represented as a semantic class, namely *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.

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

Paul Waring – pwaring@cs.man.ac.uk

1.1 Types of event

Broadly speaking, events on the Web can be divided into two types, *structured* and *unstructured*. Structured events are those which, perhaps unsurprisingly, have a well-defined and labelled structure. For example, all events on Upcoming[5] 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 events consist of natural language phrases. Unstructured descriptions of events can be ambiguous, especially when taken out of context [Feng and 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 4th December 2007") [Makkonen et al. 2002; 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 relative dates to their absolute equivalents [Mani and Wilson 2000], in much the same way as a web browser converts relative URLs into absolute ones so that it can fetch a resource. Unfortunately, few temporal expressions within text are fully specified and therefore need to be computed from the context of the surrounding text [Dale and 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 and 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.

1.2 Topics vs 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 and Allan [2004] and Yang et al. [2000] being the most obvious contrasting opinions. Kumaran and 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] is supported more by other scholars[6] than the method of treating topics and events as synonyms.

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

[6] 'An event (e.g., SPSS acquiring NetGenesis) is an instance of a specific event topic (e.g., business merger).' [Wei and Lee 2004]

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.

2.1 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,[7] 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.[8]

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] also consider clustering to be 'a major component of our event detection approach', and [something here]

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.

2.1.1 *Hierarchical clustering.* Hierarchical clustering is an extension of the general concept of clustering. Instead of clusters being interspersed amongst one another with no form or structure, clusters are instead arranged in a hierarchy according to how close (i.e. similar) two clusters are to one another.

Broadly speaking, there are two main algorithms used for hierarchical clustering, namely *agglomerative* and *divisive*. Agglomerative clustering begins by placing each data item in its own individual cluster. The two clusters which are nearest (i.e. most similar) to each other are then merged into a single cluster. This step is repeated until all of the data items have been merged into a single process. This process of iteratively merging clusters creates a hierarchy.

Divisive clustering, in constrast to agglomerative clustering, employs a top down approach to creating a hierarchy. Initially, all of the data items are contained in a single cluster. This cluster is then split into a set of child clusters, which are them-

[7] One practical use of this might be to remove duplicate Web pages from search results, by only returning one result from a cluster of identical documents.
[8] For example, if we wish to separate documents which refer to different events [Smith 2002].

selves divided further, until each cluster only contains a single data item. Whilst both algorithms work in a similar way (one is the reverse of the other), agglomerative algorithms are generally considered to be more computationally efficient than their divisive counterparts [Liu and Kellam 2003, p.233], and are therefore more popular overall [Liu 2005, p.132].

One use of hierarchical clustering is to construct a topic hierarchy from a group of text documents [Liu 2005, p.135]. A well-known example of hierarchical clustering being used to create such a topic hierarchy is the Yahoo! Directory,[9] which is a human-edited list of Web pages organised under a range of subdirectories. Each subdirectory acts like a cluster, in that it contains links to sites relating to the same topic. In addition, clusters also contain links to sub-clusters, which contain links to Web pages related to sub-topics, creating a hierarchy.

Whilst hierarchical clustering has many uses, its major flaw exists in the fact that clusters can only be connected by a single attribute of the data items contained within the clusters. For example, in the Yahoo! Directory clusters are connected in a hierarchy based on the topic of the Web pages contained within the clusters. However, there is no way to connect clusters based on other attributes. Once a user has found a Web page which is of interest, he can only use the hierarchical structure of the directory to discover sites with the same broad topic, but not sites which may be related by some other criteria, e.g. being written by the same author. Furthermore, a hierarchical structure is intrinsically limited as it can only represent attributes which naturally fit into a hierarchy. Even with topics, which arguably fit this criterion, it is sometimes difficult to create a hierarchy which represents all of the possibilities, and categories often end up being duplicated across the directory. Finally, the precision required by a directory structure may frustrate users who are looking for an unexpected or serendipitious connection [Catledge and Pitkow 1995].

2.2 Linking

Some confusion can arise with the use of the word 'linking', as it has several meanings. Perhaps the most common example of this is the use of linking to mean creating a hyperlink between two Web pages. This is not what we are aiming to achieve, although it may be the case that after we have performed our link detection, we present the results as a series of hyperlinks.

The TDT project also 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 aiming to link events based on topic membership.

Instead, our aim is to create connections (links) 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. For example, if two events occur at the same location, then there will be a link between them. Each link will have a *link weight* which will be equal to the number of attributes

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

which are shared by the two events. The links between events will be bi-directional, but not transitive.

The issue of linking events has been discussed previously in Feng and 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. Firstly, 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. In addition to this, we will also be aiming to expand the concept of event detection and linking beyond the limited area of news stories and onto the Web in general.

Our form of linking is different to clustering in that it connects events based on whether *some* attributes have common values, rather than *all* attributes. For example, if we have two events, we would compare them as follows:

(1) If the events have common values for all of their attributes, then they are *identical* (for our purposes) and should be placed in the same cluster.

(2) If the events have some, but not all, attributes in common, then they are *related* and the clusters which they are in should be linked.

(3) If the events have no attributes in common, they are *unrelated* and their clusters should not be linked.

The advantage which linking gives us over clustering is that it allows us to serendipitiously discover related events, whereas clustering only allows us to discover 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. 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 relevant information for the user, which would otherwise not have been presented to them.

A secondary benefit of building linking on top of clustering instead of simply linking events directly to one another is the degree of scalability which this system provides. Without clustering, adding a new event would involve comparing it to each event already in the collection and seeing if the two events are identical or related. This would involve a number of comparisons equal to the size of the collection, i.e. this algorithm would be linear in performance. However, if we cluster events in our collection, then a new event will only need to be compared

to each cluster, and in most cases there will be far fewer clusters than there are events. Even in the worse case, where each event is unique and therefore has its own cluster, this method will still only require the same number of comparisons as a system not using clustering. Furthermore, when adding a new event to a cluster, the event will automatically inherit all of the properties of that cluster - i.e. its links to other clusters.

2.3 Previous work

Whilst some research does exist on the subject of connecting events, all 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 done on 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'. In addition to this, the detection of events seems to be largely limited to collections of online news stories, with the exception of Smith [2002] (though even this is a fixed corpus with a degree of structure), and there appears to be little or no work on the subject of detecting and linking events on the Web in general, as opposed to a specific and well-defined corpus of text.

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 and 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 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 and Walsh 1988], as opposed to searching, where the goal is known beforehand (citation needed). However, whilst there are differences between the two techniques, searching and browsing are not mutually exclusive – the two methods may be considered complementary [Jul and Furnas 1997], and both are often employed in the user's quest for the information she seeks [Catledge and 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 sug-

gested by Cove and Walsh [1988][10] and accepted by various other scholars [Carmel et al. 1992; Catledge and Pitkow 1995], 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.

Although knowledge workers appear to split their time equally amongst all three tasks,[11] 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.

3.1 Serendipitous browsing

In our work, we will be concentrating on serendipitous browsing as well

Serendipity can be considered to be an essential aid to the process of discovery across disciplines, both in the humanities [Delgadillo and Lynch 1999] and the sciences [Foster and Ford 2003].

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 (*precomputed links*).
(3) Links created when needed as the results of a computation (*dynamically computed links* or *dynamic links*).

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

[10] Similar sub-categories, described as *patterns*, can be found in Salomon [1990].
[11] 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].

Paul Waring – pwaring@cs.man.ac.uk

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,[12] 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]. 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*[13] between nine months and one year [Fetterly et al. 2003; Bar-Yossef et al. 2004; Dalal et al. 2004; Ntoulas et al. 2004]. Even 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 and 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, which is a possible reason for why so few sites do so at present.[14] Furthermore, the lifetime of any given URL is limited [Fetterly et al. 2003], so any links created manually must be constantly checked and updated in case they no longer point to the original material.

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

[12] A study by Cho and Garcia-Molina [2000] found that 40% of pages on over 200 popular sites changed on a weekly basis.
[13] The average time it takes for 50% of pages to become unavailable.
[14] Over a period of six years, the size of the publicly indexable Web is estimated to have grown to 300 million [Lawrence and Giles 1998], 800 million [Lawrence and Giles 1999], and 11.5 billion [Gulli and Signorini 2005] pages.

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.

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 and Pitkow 1995; Tauscher and Greenberg 1997; Weinreich et al. 2008].[15] As a result, hyperlinks are considered "an essential, if not the most important feature of the World Wide Web" [Bry and Eckert 2005].

As a result, we suggest that presenting related events in the form of links to the pages which discuss them is a sensible method to use.

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 would be found using similar sets of keywords would be considered related for this purpose.[16]

Although this functionality already exists, our work will be different 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 recent news story about super-casinos,[17] 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 Charles Kennedy 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. However, we will be comparing the results obtained from our work with those from the similar search feature to see exactly what the differences are between the two.

5. RELIABILITY OF INFORMATION ON THE WEB

The Web offers a largely decentralised and deregulated environment into which almost anyone can place information about any topic imagineable for little or no cost (citation would be useful). Whilst this freedom to access, create and modify a global knowledge base has tremendous advantages (e.g. ...), it also brings forth the problem of inaccurate, misinformed or even deliberately misleading information

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

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

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

Paul Waring – pwaring@cs.man.ac.uk

being made available at the same level as peer-reviewed publications. Another potential issue is the popularity of satirical sites, which at first glance appear to be reporting true events, but on closer inspection are deliberately inventing or exaggurating events 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 on whether a particular piece of material can be trusted as providing correct information.

Broadly speaking, there are two ways by which we can measure the reliability of the information on a page, namely *popularity* and *textual analysis*. These two methods are discussed in detail in the following sections.

5.1 Popularity

The popularity of a given page can be used as a measure for determining its reliability as a source of information. Popularity can be measured either by how many pages link to a given resource, and how trustworthy these pages are considered to be (this is broadly how search engines rank pages), or by how many users have tagged a particular resource on one of the many social bookmarking sites which are available. The rationale behind this is that sites which are popular are likely to offer reliable information, as users are unlikely to link to or tag resources which offer incorrect or misleading information.[18] However, this is not always the case, as many popular sites make .

5.1.1 *Search engines.* As mentioned previously, search engines generally assign a rank/score of some description to each page in their index, so that popular sites (which, by definition, are more likely to be what a user is looking for in most cases) appear higher up in search results.[19]

However, whilst search engines are often good at finding authoritative sites, the sites which are considered most reliable using this metric may not actually provide the most accurate information on a given subject [Yin et al. 2007]. As a result, search engine rankings, whilst a useful indicator of reliability, cannot be the only factor whici is taken into account when determining the accuracy of information on a web site.

5.1.2 *Social bookmarking.*

5.2 Textual analysis

6. GENERAL OBSERVATIONS

The majority of work in the field of event detection is related to or concerned with the detection of events in online news stories – particularly because a significant amount of the existing research involves the TDT corpus, which is basically a collection of news stories. As far as can be discerned from the literature surveyed to date, no one appears to have expanded the detection of events beyond a narrow application to take into account the rest of the Web. In addition to this, it is inter-

[18]This does of course require users to be capable of distinguishing between correct and incorrect information.

[19]It should be noted that the rank/score of a particular page is not the only factor which influences the position of a resource in the results of a query, but it is a significant one.

esting to note two points which apply to the sources of the news stories used in the papers examined so far. Firstly, the sources consist almost exclusively of American media outlets (CNN, Reuters etc.). There are sometimes good reasons for this,[20] and the choice may be at least partially explained by the fact that the TDT project is sponsored by a collection of US agencies, but it still demonstrates a potential bias in the way in which events are reported. Secondly, all of the news sources are ultimately owned by commercial organisations, who report events in order to derive revenue (and hopefully profit). Not only does this restrict the type of events which are reported (anything which is unlikely to increase circulation may well go unreported), but it also affects the way in which events are reported. Furthermore, there are now a significant number of weblogs, run by both professional authors writing for money and amateurs writing for enjoyment, which potentially contain a significant amount of information about events. The Web places very few restrictions on who can publish material and what they can publish, so it is surprising to see that these sources have been largely ignored in the literature.

REFERENCES

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

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., and 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., and 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, 37–45.

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

Bar-Yossef, Z., Broder, A. Z., Kumar, R., and 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, 328–337.

Brants, T. and 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, 330–337.

Bry, F. and 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, 135–144.

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

Catledge, L. D. and 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., and Tomkins, A. 2006. Evolutionary clustering. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 554–560.

Cho, J. and 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, 200–209.

[20] Nallapati et al. [2004] justify using CNN as 'stories from this source tend to be short and precise and do not tend to digress or drift too far away from the central theme.'

Paul Waring – pwaring@cs.man.ac.uk

Cove, J. F. and 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., and 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, 160–168.

Dale, R. and Mazur, P. 2007. The semantic representation of temporal expressions in text. In *AI 2007: Advances in Artificial Intelligence*. Lecture Notes in Computer Science, vol. 4830. Springer, 435–444.

Delgadillo, R. and 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., and Schilling, L. M. 2003. Going, going, gone: Lost internet references. *Science 302,* 5646, 787–788.

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

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

Fetterly, D., Manasse, M., Najork, M., and 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, 669–678.

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

Foster, A. and 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, 68–76.

Gulli, A. and 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, 902–903.

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

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

Kumaran, G. and 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, 297–304.

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

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., and 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., 115–121.

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

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

Li, Z., Wang, B., Li, M., and 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, 106–113.

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

Liu, X. and Kellam, P. 2003. Mining gene expression data. In *Bioinformatics: Genes, proteins & computers*, C. Orengo, D. Jones, and J. Thornton, Eds. BIOS Scientific Publishers.

Paul Waring – pwaring@cs.man.ac.uk

Makkonen, J. and 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, 393–404.

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

Makkonen, J., Ahonen-Myka, H., and 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, 251–265.

Mani, I., Schiffman, B., and 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, 55–57.

Mani, I. and 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, 69–76.

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

Nakahira, K. T., Matsui, M., and 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, 1345–1346.

Nallapati, R., Feng, A., Peng, F., and 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, 446–453.

Ntoulas, A., Cho, J., and 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, 1–12.

Petras, V., Larson, R. R., and 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, 151–160.

Rattenbury, T., Good, N., and 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.

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.

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

Sellen, A. J., Murphy, R., and 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, 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, 73–80.

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

Tauscher, L. and 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., and 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, 415–422.

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

Paul Waring – pwaring@cs.man.ac.uk

Wei, C.-P. and 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., and Mayer, M. 2008. Not quite the average: An empirical study of web use. *ACM Transactions on the Web 2,* 1, 1–31.

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., and 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., and 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, 65–72.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. 1999. Learning approaches for detecting and tracking news events. *Intelligent Systems and Their Applications 14,* 4, 32–43.

Yang, Y., Pierce, T., and 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, 28–36.

Yesilada, Y., Lunn, D., and 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, 3–10.

Yin, X., Han, J., and 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, 1048–1052.

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

Zhang, K., Zi, J., and 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, 215–222.