Detecting and Linking Events

PAUL WARING
School of Computer Science, University of Manchester

1. INTRODUCTION

The World Wide Web contains billions of pages of information,[1] and this corpus grows in size every day. Many of these pages discuss events, both current and historical, but this information is difficult to extract and analyse due to the unstructured and inconsistent nature of the Web [Dill et al. 2003]. As well as conventional sources, such as stories from large news agencies, there is also a significant amount of material being added through the form of user generated content, particularly via blogs and social networking sites [Ni et al. 2007; Cha et al. 2007]. Furthermore, the growing utilisation of social bookmarking sites allows anyone to categorise content on the Web, as opposed to restricting this ability to central authorities such as editors or document authors [Golder and Huberman 2006].

The overall aim of the DALE project is to create a model of event expressions on the Web, in order to understand how this information can be presented to users in a more structured way, and provide opportunities for serendipitous discovery of new information. This report will examine several areas of related work, which we shall be both building on and contributing to.during the course of the project.

2. 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.[2] 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.

In one of the earliest works in the area of event tracking, Allan et al. [1998] suggest 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. This definition appears to have

[1] Estimates from search engines vary from between two and eight billion pages which are publically accessible (and therefore indexable) on the Web. [Gulli and Signorini 2005]
[2] "e.g. an airplane crash, and earthquake, governmental elections, etc." [Brants and Chen 2003]

been accepted by a number of other researchers,[3] who explicitly refer to it as the definition of "an event" rather than suggesting their own alternative. Makkonen et al. [2002] point out some examples of cases which might not be considered events under this definition, such as those which continue over a long period of time. 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, Allan et al. [1998] state 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. 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. Makkonen et al. [2002] 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.'

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, particularly Scholes [1980] and Vendler [1967], see time as the most important aspect of an event. In addition, most 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:

(1) Something which happened.
(2) The place where it happened.
(3) The time when it happened.

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] a significant amount 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 several items in the literature 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

[3]For example, Makkonen and Ahonen-Myka [2003], Li et al. [2005] and Zhang et al. [2007].
[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

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 of less interest to our work, as we are not initially concerned with 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] agree, 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 thus far, we can suggest that an event might be best described as "something which happened, at a given place and time, and involving a certain set of individuals". Whilst in some cases not all of these event attributes will be explicitly present, this is the definition of event which we shall be applying throughout the rest of this work.

3. 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 sections.

3.1 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 Topic Detection and Tracking (TDT) programme,[7] 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.

[5]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.

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

[7]`http://www.nist.gov/TDT`

If a story discusses a topic which has already been encountered, 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 significant part of their event detection approach, utilising it in a similar way to Allan et al. [2005].

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.

3.1.1 *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 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 themselves 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,[8] 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

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

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].

3.2 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 refer to the process of 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, topic membership is not a requirement in order for a link to be created.

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, there will be a link between them. Each link will have a *link weight* which will be equal to the number of attributes 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 also differs from 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") and happened immediately after one another. We 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.

4. 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, "the task of looking for a known target", can be constrasted with browsing, "the task of 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. 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 suggested by Cove and Walsh [1988],[9] and accepted by various other scholars,[10] which are:

(1) *Search browsing:* Where the goal is already known before browsing begins –

[9]Similar sub-categories, described as *patterns*, can be found in Salomon [1990].
[10]For example, Catledge and Pitkow [1995] and Carmel et al. [1992]

Paul Waring – pwaring@cs.man.ac.uk

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."

These categories cover a wide range of possiblities, from directed browsing informed by search to random browsing where resources are encountered serendipitously. 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. 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], and as such we suggest that it is a useful process to stimulate.

5. DYNAMIC LINK GENERATION

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. Yan et al. [1996] suggest the following reasons for why dynamic link generation may be desirable:

(1) Links can be customised for an individual user, based on the content which she has expressed an interest in 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,[11] dynamic links are the only feasible way to connect pages which mention related events. Attempting to create manual links is both expensive and inefficent and would therefore be unfeasible on a corpus as large as that of the Web [El-Beltagy et al. 2001]. Furthermore, the likelihood of any given URL being available decays over time, and the lifetime of any given URL is limited [Fetterly et al. 2003]. Even in the area of scientific research and publications, where we might expect additional effort to be

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

put into ensuring web references are persistent, surveys have shown that approximately 20% of URL citations are unavailable one year after their creation [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 one possible reason for why so few sites do so at present.[12] Any manual links connecting pages which mention related events would soon fail to accurately reflect the information available on the Web, as new pages are created and existing ones are removed or modified on a daily basis.

In addition to these benefits, dynamic link generation has also been shown to significantly reduce the amount of time required by users in order 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, which suggests 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].[13] 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.

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.

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.

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.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 1–14.

[12] 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.

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

Paul Waring – pwaring@cs.man.ac.uk

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.

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

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

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., and 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, 178–186.

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.

Golder, S. A. and Huberman, B. A. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science 32,* 2, 198–208.

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.

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

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.

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.

Paul Waring – pwaring@cs.man.ac.uk

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.

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.

Ni, X., Xue, G.-R., Ling, X., Yu, Y., and Yang, Q. 2007. Exploring in the weblog space by detecting informative and affective articles. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM, 281–290.

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.

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.

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.

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.

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.

Paul Waring – pwaring@cs.man.ac.uk