Detailed research description

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].

Without the ability to analyse this social expression on the Web, a wealth of opinion and experience is lost. Only by developing a detailed understanding of how events are described and discussed can we propose solutions which will enable us to analyse this information in depth. Therefore, the aim of the AiSC (Automatically Identifying Social Change) project is to create a model of these expressions of events and suggest ways in which this information can be extracted from the inconsistent and unstructured documents on the Web. In order to achieve this, AiSC has three objectives:

(1) Investigate how individuals express their thoughts and opinions about events on the Web in order to develop an understanding of the nature of this method of interaction.

(2) Based on this investigation, create techniques and algorithms to mine the Web for this information and identify social trends.

(3) Design, develop and deploy experimental tools to test and evaluate these techniques and algorithms.

2. HYPOTHESES

Our work is based on three hypotheses:

(1) Solutions can be created to extract event information from unstructured text.

(2) These events can be linked together based on common attributes.

(3) By representing these links between events in a way which users can understand, we can provide additional structure to the Web and allow users to serendipitously discover related information.

[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]

The project will be complete once all of these hypotheses have been tested and conclusive answers have been provided to our research questions (see page 2).

3. BACKGROUND

Event reporting: Events (things which happen at a given time and location, and involve a number of individuals) can be reported and discussed on the Web, both via sites run by large organisations with editorial control (e.g. BBC News) and individuals through blogs and personal Web sites. Extracting and analysing information about events from the Web, particularly news stories, can provide users with useful information, whether this be notice of new events or a collection of stories which discuss the same broad topic. This area has been examined in the past, particularly under the auspices of the Topic Detection and Tracking project,[2] however we will be aiming to build on this existing work and make a unique contribution through the concept of linking events.

4. RESEARCH QUESTIONS AND SIGNIFICANCE

It is clear that changes are occuring to the way in which content is generated for the Web. Furthermore, the nature of this content is changing in such a way that it reflects the views of individuals rather than a single corporate body. These changes are happening at such a rapid pace that significant amounts of information regarding events risks being lost if the problem of analysing this data is not addressed. AiSC seeks to examine ways in which this uninvestigated issue can be overcome.

Before proceeding, we must answer the following novel research questions, which are intrinsically linked to our hypotheses:

(1) Can information about events be extracted from unstructed Web pages?
(2) Can these events be linked based on common attributes?
(3) Can these links be represented in a way which allows users to serendipitously discover related information about the event they are interested in?

By addressing these research questions, we will incorporate additional structure to the Web and improve the browsing experience for users by allowing them to serendipitously discover new information about events which they are interested in.

5. MOTIVATIONS

The principle motivation for addressing the above research questions is to incorporate additional structure into the heterogeneous corpus of information about events which exists on the Web. By adding this structure, we will improve the browsing experience for users by providing them with the facility to serendipitously discover new information about events which they are interested in.

6. GENERALIZABILITY

Although we will initially be applying our methods and algorithms to the problem of dynamically linking pages which mention related events, we expect that our work will be sufficiently extendable to allow other types of content to be linked in a similar way.

[2]In particular: Allan et al. [1998], Yang et al. [1999] and Lavrenko et al. [2002].

Paul Waring – pwaring@cs.man.ac.uk

7. METHODOLOGY

The programme for AiSC consists of several pieces of work, broken down into seperate research packages which build upon one another but can be tackled independently one at a time, in sequence. In addition to the experiments (and their corresponding evaluations), we will also be running a concurrent investigation into how events are detailed and discussed on the Web, in order to inform and improve our work.

7.1 Classifying seed sites

In order to ensure that we select sites which are relevant and useful, we will use the top fifty most popular sites from the Alexa ranking system[3] as our starting points. Once we have selected these sites, we will then place them in categories according to a classification system. We have chosen to use an existing classification system in order to categorise the sites, in order to ensure that the method of classification is not geared towards producing the results we hope to obtain. The system we have chosen is *The connectivity sonar*, as detailed in Amitay et al. [2003], which has been in existence for some time and has been commented on by numerous other researchers.[4]

However, we shall be making a modification to the categories suggested by Amitay et al. [2003] in order to ensure that they are a closer fit for our requirements. Speficially, we will not be including sites from the *Search Engine* and *Web Directory* categories, for two reasons. Firstly, these sites generally do not contain a significant amount of content, and are therefore unlikely to describe events in sufficient detail to enable us to link them to other pages discussing related events. Furthermore, we agree with the assertation made by Lindemann and Littig [2006] that search engines and web directories are used as a way of finding other sites, rather than being the termination point where the user finds the information which she was looking for.

7.2 Identifying events

After selecting our seed pages, we will download each one and attempt to manually extract any information about events which is contained within them. In doing so, we will be trying to formalise the way in which events are identified, e.g. how can we tell that a particular word or phrase refers to a specific location or time?

7.3 Manually generating queries

Once we have a set of seed pages and the event attributes contained within them, we will create queries based on these event attributes. These queries will be fed into a search engine such as Google or Yahoo!, and the results obtained will be analysed to see if they contain pages which mention events related to the once used to generate the query. We will gradually refine this process until we have developed a method for translating event attributes, which we already know, into a query which will return several pages containing information about related events.

[3] `http://www.alexa.com/`
[4] Bechhofer et al. [2006]

By building on the efforts of major search engines, we remove the need to deploy significant resources in order to crawl and index sufficient pages to test our hypotheses. We can also leverage the existing algorithms used by these search engines instead of having to develop our own techniques, which would require a significant investment of time and effort.

7.4 Dynamic link generation

After refining our queries to the stage where we can obtain links to pages containing related information, we will extract and display these links alongside the original page so that the user can browse to sites which we believe contain information about related events. Each time the user follows one of our links, we will analyse the page which is requested and dynamically generate a new set of links which correspond to events related to the new page.

7.5 Experimentation

A major part of our work will involve running experiments and evaluating their outcomes. The experiments which we initially expect to run are outlined in the following sections.

7.5.1 *Manual extraction of events.* Our starting point for experimentation will be to generate a fixed number of browsing trails, created by manually performing the steps outlined in sections 7.2 to 7.4 on our seed sites. We will present users with these pre-determined browser trails and analyse how they interact with the links which we have generated to pages which we feel contain related information. In particular, we will be looking at which links users click and how long users spend on a page after they have clicked a link (more time spent on a page is an indicator that the user has found something of interest to read). In order to obtain this information, we expect to have a proxy between the user and the sites which they are visiting, which will log access times, resources fetched and other data which can be analysed offline after the experiment has been completed.

7.6 Evaluation

After performing each of our experiments, we shall evaluate them against a set of criteria, which will differ for each experiment. These criteria are outlined below.

7.6.1 *Manual extraction of events.* When evaluating our manual methods for extracting events, we will be looking to identify factors which suggest that users have found the links presented to them useful, which will help support our hypotheses. In particular, we will be basing our evaluation on criteria such as the time spent on a page – this can be a useful indicator of whether the reader is interested in the content or not. If the user immediately returns to the previous page or clicks another link, then it is likely that the page contains nothing of interest to them. In addition, we also hope to allow users to rate a link based on how relevant they believe it to be, on a sliding scale from 'not relevant at all' to 'highly relevant', which we can then use to evaluate how relevant our results are from the users' point of view.

7.7 Ethical issues and approval

As part of this project, experiments will be run which involve human participants. As a result, AiSC will require ethical approval from the relevant University ethics committee for these areas of work. As the experiments will not put the users at risk or require the disclosure of any personal information, obtaining ethical approval is expected to be a straightfoward process. In accordance with general ethical research guidelines, we will ensure that partipants are allowed to withdraw from an experiment at any time, without giving a reason, and have their data destroyed. Furthermore, any data which is collected as a result of experiments run as part of this project will be anonymised so that no data can be identified as originating from a named individual.

8. DISSEMINATION

The ongoing work and results from the project will be disseminated through a number of channels, including publications through journals and conference proceedings and presentations to appropriate national and international meetings. Conferences for which the subject material of AiSC would be relevant include ACM Conference on Hypertext and Hypermedia (HT), Joint Conference on Digital Libraries (JCDL) and the International World Wide Web Conference (WWW).

9. FUTURE WORK

Social expression: Individuals can express their thoughts and opinions on the Web in a growing number of ways. One of the most popular ways is by keeping an online diary, often referred to as a 'weblog', in which they present their thoughts about (usually recent) events which have affected them in some way, whether through direct involvement or some other connection [Ni et al. 2007]. Further opportunities for individual expression can be found via sites which allow users to apply 'tags' to items of content, such as photographs,[5] videos,[6] and interesting web pages.[7]

Social bookmarking, through sites such as del.icio.us, can also be seen as a way of annotating the web through a social mechanism [Wu et al. 2006]. Furthermore, individuals can also express themselves through social networking sites such as Facebook[8] and MySpace,[9] which allow users to create a profile page detailing their interests, activities, and what they are currently doing through mechanisms such as 'status updates' and 'mini feeds'.

REFERENCES

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.

[5] E.g. flickr, `http://www.flickr.com/`
[6] E.g. YouTube, `http://www.youtube.com/`
[7] E.g. del.icio.us, `http://del.icio.us/`
[8] `http://www.facebook.com/`
[9] `http://www.myspace.com/`

Amitay, E., Carmel, D., Darlow, A., Lempel, R., and Soffer, A. 2003. The connectivity sonar: detecting site functionality by structural patterns. In *HYPERTEXT '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*. ACM, New York, 38–47.

Bechhofer, S., Harper, S., and Lunn, D. 2006. Sadie: Semantic annotation for accessibility. In *The Semantic Web - ISWC 2006*. Springer, 101–115.

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.

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.

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.

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.

Lindemann, C. and Littig, L. 2006. Coarse-grained classification of web sites by their structural properties. In *WIDM '06: Proceedings of the 8th annual ACM international workshop on Web information and data management*. ACM, New York, 35–42.

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.

Wu, X., Zhang, L., and Yu, Y. 2006. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*. ACM, 417–426.

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.

Paul Waring – pwaring@cs.man.ac.uk