Report on the 2nd Web Dynamics Workshop, at WWW'2002

Mark Levene and Alexandra Poulovassilis
School of Computer Science and Information Systems
Birkbeck College, University of London
Malet Street, London WC1E 7HX, U.K.
{m.levene, ap}@dcs.bbk.ac.uk

1 Introduction

The World-Wide-Web is a ubiquitous, global tool, used for finding information, communicating ideas, carrying out distributed computation, and conducting business. The web is highly dynamic in the quantity and nature of the information that it encompasses. Thus, there is a need to understand how the information content and usage of the web change, and to develop techniques for organising and processing information which handle and exploit its inherent dynamics. Access to the web may be from a variety of devices and interfaces, different users at different locations, and at varying times. There is thus also is a need for techniques which dynamically adapt information presentation to the mode of access and to the specific user requirements.

In January 2001 we hosted the first Web Dynamics workshop, in conjunction ICDT'01 in London, to explore these issues. The papers presented are available on-line [1] and a review of the workshop appears in [2]. The papers in that first workshop fell into three main areas: web information retrieval (search, navigation, querying), dynamic applications (web site management, mobile computation, active rule languages) and adaptive hypermedia. Following on from the workshop, we organised a special issue of *Computer Networks* [3] and the papers in this special issue covered XML data management issues, mathematical models of the web, web searching, and web navigation.

The aim of the 2nd Web Dynamics Workshop was to continue this momentum and to provide a forum for discussion of current research into these key areas. The papers accepted to the workshop fell into four tracks which we briefly report on below. The proceedings of the workshop are available on-line at `http://www.dcs.bbk.ac.uk/webDyn2/`.

2 Web Structure and Evolution

The first paper on "Web structure, age and page quality" was given by R. Baeza-Yates and co-authored by F. Saint-Jean and C. Castillo. The authors collected data from the Chilean web domain during two time points: the first half of 2000 and the second half of 2001. Regarding the structure of the web graph, their results show that older sites are more likely to be in the strongly connected component of the web while, on average, newer sites tend to be disconnected from the main web component. The part of the web that links into the main component is split between new sites and older sites that have not become popular, and the part of the web that is reached from the main component has sites which are, on average, newer than sites in other components. The authors also investigated the connection between link-based ranking schemes, such as Google's pagerank, and age of web pages. They found that new pages, as well as very old pages, have low pagerank, and that the highest pagerank is possessed by pages which are 1.6 months old. They propose a variant of pagerank which adds to it an age function which decreases with age. In this way the bias of pagerank against new pages can be decreased.

The second paper, "A steady state model for graph power law", was given by J. Wang and co-authored by D. Eppstein. In the past few years there have been substantial advances in our understanding of complex and evolving networks, the largest one being the World-Wide-Web. An important result, which may have a big impact on web algorithmics, is that the degree distribution of the web graph follows a power law. In order to understand the emergence of these power laws researchers have created models that generate such graphs. These models include two crucial ingredients, preferential attachment and incremental growth. Eppstein and Wang present an algorithm that produces a power law distribution through preferential attachment without recourse to incremental growth. Their algorithm consists of repetitively removing and adding edges from a random sparse graph. They have validated their model by simulation, although it remains an open problem to prove analytically that a power law emerges.

The third paper on "A multi-layer model for the web graph" was given by L. Laura and co-authored by S. Leonardi, G. Caldarelli and P. De Los

Rios. As in the previous paper, the authors investigate a stochastic model of the evolution of the web graph, the point of departure being the division of the graph into regions which are local subsets of the graph. Regions could, for example, be viewed as cyber-communities. Individual web pages may belong to more than one region but linking is localised to pages within a region. By assigning web pages preferentially to regions, in addition to the usual preferential attachment of links, a multi-layer graph is formed. As in the previous paper, the model has been validated by simulation and, as predicted, the power law distribution of incoming links emerges.

3 XML Technologies

The first paper in this track, "Modelling adaptive hypermedia with an object-oriented approach and XML" by M. Cannataro, A. Cuzzocrea, C. Mastroianni, R. Ortale and A. Pugliese, could not be presented due to unforeseen personal circumstances. Adaptive hypermedia concerns the adaption of information presentation according to the user's needs and the client device being used. This paper presents a 'data-centric' model for adaptive hypermedia. Heterogeneous data sources are represented by XML meta descriptions. These descriptions are organised into a directed graph for navigational purposes and into object classes for semantic purposes. Adaption is along three orthogonal dimensions: the user's browsing behaviour, the external environment, and the client device. The first two of these drive the generation of page content and page links, while the third drives the format of pages. A 'profile view' defines the set of objects accessible by users belonging to each user profile. The final pages to be displayed on the client device are dynamically generated by applying an XSL document, either at the server or at the client. The authors conclude by describing a three-tier modular architecture which supports their model, which they are currently implementing.

The second paper on "A logico-categorical semantics of XML/DOM" was given by C. H. C. Duarte. The author argues that XML lends itself naturally to a categorical logical semantics, and after a brief overview of XML and DOM, outlines a formal semantics for both. For XML he first develops a formal interpretation for individual XML documents, and then discusses how the composition of documents can be captured by means of a pushout operation. For the DOM semantics, temporal features are used to capture the dynamic properties of DOM. The author's aim with this work is to provide a formal semantics which allows formal reasoning about distributed systems that use XML for communication of messages and other information, and facilitates the development of automated tools for the verification of such systems.

The third paper on "XML structure compression" was given by M. Levene and co-authored by P.Wood. XML is becoming a standard means of moving large volumes of information over the web, and hence there is a pressing need for compression techniques for XML data. This paper presents an algorithm for compressing XML documents which conform to a DTD. Knowledge of the DTD allows separation of the structure of a document from its content, so that the encoded version does not contain any information that may be inferred from the DTD. Moreover, the textual parts of data can be compressed separately by conventional means. It is shown that under the assumption of independence within DTD rules, the resulting encoding is minimal and, on average, equal to the entropy of the document. Ongoing work on validation of the algorithm involves implementation and comparison with other XML compressors.

4 Web Information Retrieval

The first paper on "Criteria for evaluating information retrieval systems in highly dynamic environments" was given by J. Bar-Ilan. The thesis of this paper is that current evaluation methods of search engines, such as precision and recall, are insufficient due to the fact that the growth of the web is much faster than the growth of a search engine's index. The author addresses methods for evaluating the freshness of the search engine's results and its stability over time. The evaluation involved submitting a query to several search engines over a period of time and examining how the answer set changed. The results show that web pages that appeared in earlier results quite often got lost, despite the fact that these web pages were still valid, and that the overlap in the answer sets was not as high as it should be.

The second paper on a "Query language for structural retrieval of deep web information" was given by S. Mueller and co-authored by R.-D. Schimkat and R. Muller. The 'hidden' or 'deep' web refers to information that resides in databases and is accessible only through dynamic script pages or servlets. The authors suggest access to deep web information by making meta-data available to the search engine in the form of Entity-Relationship models. They argue that a query interface to deep web information that is based on a graphical representation pertaining to the Entity-Relationship model will improve the usability and allow more precise access to these hidden resources.

The third paper on "Exploration versus exploitation in topic driven crawlers" was given by G. Pant and co-authored by P. Srinivasan and F. Menczer. Topic driven crawlers, also known as focused crawlers, have been suggested as a means of limiting the search space of these software robots to specific queries or categories. In this paper metrics are suggested to evaluate focused crawlers that combine an exploration of the information space by following links which may appear sub-optimal with an exploitation strategy which involves only following the most relevant links. Through experimentation with different crawling strategies the authors conclude that exploration is important to discover highly relevant pages whereas too much exploitation using greedy strategies can be harmful.

5 Dynamic Applications

The first paper on "WebVigil: an approach to just-in-time information propagation in large network-centric environments" was given by S. Chakravarthy and co-authored by J. Jacob, N. Pandrangi and A. Sanka. This paper is concerned with efficient detection and notification of change in web pages. It discusses the use of event-condition-action (ECA) rules for this purpose, and in particular the need to extend ECA rule functionality from a centralised context to handle distributed, heterogeneous events, and also the need for selective notification of changes to users depending on their interests or profiles. The paper begins with an overview of related work, a review of the push and pull paradigms, and a discussion of possible architectures for the push paradigm. The paper then describes the architecture and ongoing implementation of WebVigil, a system for detecting and notifying changes in web pages. Users specify by means of a 'sentinel' the kinds of changes they are interested in with respect to specific pages and the frequency with which they wish to be notified of such changes. Information is then automatically extracted from this sentinel for use by the various components of the system. Future work will include assessing the scalability and flexibility of the WebVigil architecture, and extending the expressiveness of sentinels to allow different change 'windows' (not just changes between the previous and current version of a page) and personalised change summaries.

The second paper on "Caching schema for mobile web information retrieval" was given by R. Lee and co-authored by K. Goshima, Y. Kambayashi and H. Takakura. It discusses how traditional LRU caching is not suitable in mobile environments, and that new algorithms are needed to determine the priority of items in the cache and also to determine what information to cache. They propose the use of the semantic relationships between locations to determine the likelihood of a user moving from one location to another, and hence the likely usefulness of caching information relating to already visited and new locations. Also, due to the small cache size, it may be necessary to cache URLs and other meta-data rather than actual web page contents. An experimental travel guide application is described which consists of two phases, a 'planning phase' before the user starts travelling and a 'retrieval-and-guide phase' during travel. A new web page priority ranking algorithm is proposed which utilises the semantic relationships between locations, and the results of some simulation experiments are presented. One problem that arises with their algorithm is that pages containing many geographical keywords tend to get ranked higher, and that a deeper semantic analysis is needed of the actual meaning of web page contents.

6 Conclusions

At the end of the paper presentations, there was a brief discussion session on the major current research themes in Web Dynamics and directions for continuing the momentum of the workshops. Mark Levene started the discussion by noting some areas where more theoretical and practical research is needed. Pedro Domingos expressed the opinion that understanding the growth of the web is becoming less relevant due to the infinite nature of the 'deep' web, and accessing and utilising this is now of the essence. Ricardo Baeza-Yates said that we are now at the 'big bang' of the web, and techniques for predicting future content, structure, and usage are needed. Judit Bar-Ilan said that there is a need for more sophisticated analysis of visits to pages, in that visits may not just be due to links. Sharma Chakravarthy said it was important to have a precise focus to the Web Dynamics theme, and make clear the distinctive research directions it entails. Alex Poulovassilis noted that the original theme of the first workshop was "change in the information content, topology and usage of the web" and that by this second workshop a number of clear research tracks had indeed emerged.

In conclusion, several people thought that this was the right time to put together a book of collected works, and that a third workshop on Web Dynamics should be held in 2003. We very much enjoyed organising and hosting this second workshop and we do plan to pursue the possibility of a book and a third workshop in 2003.

References

[1] *Proceedings of the 1st Web Dynamics Worskhop*, January 2001. www.dcs.bbk.ac.uk/webDyn.

[2] M. Levene and A. Poulovassilis. Web dynamics. *Software Focus*, 2:60–67, 2001.

[3] *Computer Networks*, volume 39(3), June 2002. Special issue on Web Dynamics, www.elsevier.nl/locate/comnet.

[4] *Proceedings of the 2nd Web Dynamics Worskhop*, May 2002. www.dcs.bbk.ac.uk/webDyn2.