Report on the 3rd Web Dynamics Workshop, at WWW'2004

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

1 Introduction

The web is highly dynamic in both the content and quantity of the information that it encompasses. In order to fully exploit its enormous potential as a global repository of information, we need to understand how its size, topology, and content are evolving. This then allows the development of new techniques for locating and retrieving information that are better able to adapt and scale to its change and growth. The web's users are highly diverse and can access the it from a variety of devices and interfaces, at different places and times, and for varying purposes. Thus, new techniques are being developed for personalising the presentation and content of web-based information depending on how it is being accessed and on the individual user's requirements and preferences. New applications in areas such as e-business, sensor networks, and mobile and ubiquitous computing need to be able to detect and react quickly to events and changes in web-based information. Traditional approaches using query-based 'pull' of information to find out if events or changes of interest have occurred may not be able to scale to the quantity and frequency of events and changes being generated, and new 'push'-based techniques are being deployed in which information producers automatically notify consumers when events or changes of interest to them occur. Semantic Web and Web Service technologies are being developed and adopted, with the aim of providing standard ways for web-based applications to share and personalise information.

Motivated by these issues and research challenges, we organised the 3rd Web Dynamics Workshop at WWW'2004 [3], following on from the 1st and 2nd Web Dynamics Workshops in 2001 and 2002 [1, 2]. The papers accepted to the 3rd Web Dynamics Workshop were presented in four sessions, each of which we report on below. We also refer the interested reader to a recent book on Web Dynamics which we co-edited [4].

Session 1: Web Search and Navigation

A paper on "Dynamics of Search Engine Rankings - A Case Study" was given by Judit Barl-Ilan and Mazlita Mat-Hassan, and co-authored with Mark Levene. The objective of this study was to characterise changes in the top-n results of major search engines over time and to compare the ranking between these search engines. Three methods were used to assess the changes in rank-

ing over time: (1) the size of the overlap, (2) Spearman's rho and (3) an extension of Spearman's footrule. The rankings of the top-ten results of the search engines Goole and AlltheWeb were compared on identical queries for two three-week periods during October 2003 and January 2004. The results show that the rankings of AlltheWeb were very stable over each period, while the rankings of Google underwent constant minor changes, with occasional major ones. Changes over time can be explained by the dynamic nature of the web or by fluctuations in the search engines' indexes, especially when frequent switches in the rankings are observed. The top-ten results of the two search engines have surprisingly small overlap (occasionally only a single URL). The task of comparing their rankings is therefore extremely challenging, and more refined methods for comparison need to be developed.

A paper on "Local Methods for Estimating PageRank Values" was given by Yen-Yu Chen and co-authored by Qingqing Gan and Torsten Suel. The PageRank of a web page is a global metric used by Google to rank pages according to their authority independently of their content. The computation of PageRank values is expensive and normally done offline due to the size of the web, currently exceeding four billion pages. However, there are situations when global computation of PageRank values is not possible and a quick estimate of the PageRank is needed for a collection of pages. For such situations the authors propose an efficient algorithm which approximates PageRank values for individual pages, based on local methods which expand a small subgraph in the vicinity of the pages of interest. It is shown that a reasonable approximation of PageRank, with slightly below 10% error, can be obtained by visiting between a few dozen and a few hundred pages.

A paper on "Finding Relevant Web Pages Through Equivalent Hyperlinks" was given by Simon Courtenage and co-authored by Steven Williams. The authors present a web search technique based on link analysis, which finds links similar to the one a user is currently traversing. The algorithm first finds web pages with similar content to the destination page of the current link. The back links to these similar pages are then collected, and from these the links with similar anchor text to the current are returned to the user. The current implementation uses the Google API to compute similarity functions, although a new lightweight implementation and evaluation of the method are planned.

Session 2: Web Data Mining

A paper on " A Context Ultra-Sensitive Approach to High Quality Web Recommendations based on Web Usage Mining and Neural Network Committees" was given by Olfa Nasraoui and co-authored by Mrudula Pavuluri. Single-step recommenders match user profiles based on the nearest profile to the user using a similarity measure. However, this approach only handles linearly separated user navigation sessions. This paper proposes multi-layer perceptron neural networks (and also decision trees) to handle user sessions which are not linearly separable. In order to enhance coverage and precision of the recommendations, a more sophisticated two-step recommender system is proposed. In the first step a user session is mapped onto one of the pre-discovered profiles, which are all trained offline, and in the second step the profile-specific neural network, chosen during the first step, is used to provide the final recommendation URL. A comprehensive set of experimental results were shown to validate this two-step approach.

A paper on "Mining Evolving Web Click-

streams with Explicit Retrieval Similarity Measures" was also given by Olfa Nasraoui and co-authored by Cesar Cardona and Carlos Rojas. It discusses an immune-based approach for mining evolving user profiles. Immune systems algorithms are biological models of intelligent systems that consist of a network of cells with stimulating and suppressing links between them. Each cell in the network represents a learned pattern, and a link between two cells is strengthened the more similar they are. Cells that are no longer stimulated can forget and eventually die, which is why periodical reminders in the form of re-vaccinations are needed. In the context of web usage mining, a cell represents a profile and an antigen, which has an effect on the cell's stimulation level, is modelled through input training data from web server log files. The paper proposes a single-pass algorithm for mining evolving user profiles using an immune-based clustering approach. It uses a simple similarity measure that has the advantage of explicitly coupling the coverage and precision of learning to its early stages. Simulation results show that the algorithm is both efficient and competitive in terms of its adaptability to evolving patterns.

A paper on "A Recommendation Model Based on Latent Principal Factors in Web Navigation Data" was given by Xin Jin and co-authored by Yanzan Zhou and Bamshad Mobasher. An approach based on a latent variable model to discover hidden factors in web usage data is proposed. A personalisation algorithm based on principal factor analysis is used to discover latent factors, which are in turn used to create aggregate usage profiles representing common navigation patterns. The usage profiles together with the user's active session form the basis for generating dynamic recommendations to the user. Experiments to validate the approach were carried out on two data sets showing that this approach can successfully uncover patterns that characterise a web site's functional structure, and distinguish between different types of user interests and navigational tasks.

Session 3: Events, Change, and Web Services

A paper on "RDFTL : An Event-Condition-Action Language for RDF" was given by Alex Poulovassilis and co-authored by George Papamarkos and Peter T. Wood. RDF is one of the technologies being developed to realise the vision of the Semantic Web and it is being increasingly used in distributed web-based applications. Such applications may need to be able to detect the occurrence of specific changes in the RDF descriptions, and to respond by automatically executing the appropriate application logic. Event-Condition-Action (ECA) rules are a natural candidate to fulfill this need. This paper proposes a language for defining ECA rules on RDF repositories, giving its syntax and semantics and illustrating its use via several examples. Aslo described is the architecture of a system implementing the language, both for centralised and distributed environments.

A paper on "A Dataflow Approach To Efficient Change Detection of HTML/XML Documents in WebVigiL" by Anoop Sanka, Shravan Chamakura and Sharma Chakravarthy was accepted to the workshop but could not be presented. The paper presents the WebVigiL system, which monitors specific web pages for specified changes and notifies the user in a timely fashion when such changes occur in these pages. A dataflow approach is used for detecting multiple types of changes to a web page. Also described in the paper is the change detection graph lying at the heart of the system. WebVigiL's change detection module uses ECA rules to activate

and deactive 'sentinels' on web pages and to maintain the change detection graph.

A paper on "Modeling Semantic Web Services with OPM/S A Human and Machine Interpretable Language" was presented by Benjamin Koo on behalf on behalf of its authors Dov Dori, Eran Toch and Iris Reinhartz-Berger. OWL-S has emerged from the Semantic Web community as a modelling language for enabling web service discovery, invocation and interoperation. However, OWL-S is hard to understand by humans and thus tools which generate OWL-S specifications from higher-level descriptions are being developed. This paper presents the OPM/S graphical environment for modelling web services and their interoperability. It shows how OWL-S's Service Profile, Process Model and Service Grounding ontologies for web services can all be expressed using OPM/S's Object-Process Diagrams. OPM/S has built-in mechanisms for ensuring consistency of the Object-Process Digrams describing a particular web service. Future work includes developing a translator for converting OMP/S models to OWL-S specifications, and vice versa.

Session 4: Evolution of Web Structure

A paper on "Dynamics of the Chilean Web Structure" was given by Ricardo Baeza-Yates and co-authored by Barbara Poblete. It reports on results regarding the evolution of web sites in Chile between 2000 and 2003. The major components of the web graph are: MAIN (the strongly connected component of the web graph), IN (sites that can reach MAIN but cannot be reached by MAIN), OUT (sites that can be reached from MAIN but cannot reach MAIN), and other sites not in these three components e.g. ISLANDS which are not connected to the other components. The intuition of a typical new web site is that it begins its existence in IN or ISLANDS, then migrates to MAIN as it becomes popular and other sites create links to it, and finally if the site's links are not well maintained it migrates to OUT. However, the study reported in this paper shows that the most frequent cases are for a web site to remain in MAIN or OUT or to switch between these components (50.8%), and that the third most frequent case is for a site to migrate directly from ISLANDS to OUT. Contrary to intuition, there is almost no migration from IN to MAIN, and some web sites appear for the first time in MAIN or OUT.

A paper on "WebRelievo: A System for Browsing and Analyzing the Evolution of Related Web Pages" was given by Masashi Toyoda and co-authored by Masaru Kitsuregawa. The WebRelievo system browses and analyses the evolution of the web based on link analysis. It allows visualisation of the changing structure of web sites as a time series of graphs. Visualisation of the relationship between web pages is done by link analysis based on a variation of Kleinberg's HITS algorithm. The system allows the user to view when pages appear and disappear and how their relationship changes over time. The visualisation algorithm is based on the force-directed model and layout is synchronised so that changes in the graph are easy to detect by maintaining the position of nodes and highlighting insertions and deletions. The user can interact with the graph by typing a URL, scrolling and zooming, and changing the density of the graph.

A paper on "Geographic Information Retrieval" was given by Alexander Markowetz and co-authored by Thomas Brinkh and Bernhard Seeger. It presents some preliminary results regarding mapping internet resources to geographic locations, forming the basis for geospatial search engines. In such a search engine, the user may specify a location as

part of the query, and the search engine then returns results which are not only relevant to the query but also near the desired location. Due to that fact that location metadata is sparse, the location of a web resource needs to be inferred. The URL of the resource helps to deduce the location, and often the content within the resource, such as the text in a web page, provides useful hints as to the location of the resource. Locality of a web resource is especially important, for example when locating a nearby restaurant or shop, where the relevance to the user diminishes with distance. It is argued that the balance between proximity to a location and relevance of the content is crucial for geospatial search engines and may change from query to query. It is also observed that goespatial search will be instrumental in providing users with personalised services.

Conclusions

The field of Web Dynamics is still young and many of the ideas discussed at the workshop are still maturing.

Issues relating to studying the evolution of the web will remain important, for example a model that unifies all aspects of web evolution and is consistent with the real web is yet to emerge. Predicting how the web is changing at different levels of granularity, from the components within an individual web page to the web as a whole, will allow search engine crawlers to prioritise their operation, and web sites and portals to better plan their maintenance policies.

Technologies that support personalisation and collaboration will become more pervasive. The web is not only a communication network, it is also a social network and we expect more technologies to emerge that identify and support different communities of web users. Mobile technologies have widened the reach of the web and there is a strong

need to adapt and integrate applications so that they are operable across the different platforms that the user is interacting with.

New applications and technologies such as sensor networks and ubiquitous computing are resulting in the development of new techniques for handling events and changes in web-based information. We expect that future research will need to focus on areas such as language design, performance optimisation, security, robustness, and interoperability of reactive web-based applications.

Finally, the Semantic Web initiative, with languages such as RDF/RDFS and OWL is aiming at standardised description of web-based resources, and promises to lead to better validation, sharing, and personalisation of web-based information.

References

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

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

[3] *Proceedings of the 3rd Web Dynamics Worskhop*, May 2004. See `www.dcs.bbk.ac.uk/webDyn3`.

[4] M. Levene and A. Poulovassilis, editors. *Web Dynamics: Adapting to Change in Content, Size, Topology and Use.* Springer-Verlag, Berlin, 2004.