Web Dynamics

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

June 18, 2001

Abstract

The global usage and continuing exponential growth of the World-Wide-Web poses a host of challenges to the research community. In particular, there is an urgent need to understand and manage the dynamics of the Web, in order to develop new techniques which will make the Web tractable. We provide an overview of recent statistics relating to the size of the Web graph and its growth. We then briefly review some of the key areas relating to Web dynamics with reference to the recent literature. Finally, we summarise the talks given in a recent workshop devoted to Web dynamics which was held in the beginning of January 2001 at the University of London.

Key words. Web dynamics, Web graph, information retrieval, collaborative filtering, Web navigation, Web site design, data-intensive Web applications, workflow management, e-commerce, mobile computation

1 Introduction

The World-Wide-Web is now a ubiquitous, global tool, being used in day-to-day work, for finding information, communicating ideas, carrying out distributed computation, and conducting business. The Web is continuing to grow at an exponential rate, in terms of both the amount and diversity of the information that it encompasses, and the size of its user base. This growth poses a host of challenges to the research community. In particular, there is a need to understand and manage the *dynamics* of the Web, i.e. how its information content, topology and usage change, and what kinds of models and techniques will scale up to the rate of change. There is a need for mechanisms and algorithms for organising and manipulating the information on the Web in order to make the Web tractable.

In Section 2 we briefly review recent statistics relating to the growth of the Web so that the reader can appreciate the size of the task faced. In Section 3 we summarise some of the key issues in the broad area of Web dynamics. An international workshop on Web Dynamics was held in London on 3 January 2001, in conjunction with the 8th International Conference on Database Theory (ICDT), with the aim of brining together researchers from both academia and industry who are working on novel approaches to tackling these problems. In Section 4 we give a summary of the invited talks and papers that were presented at the workshop. More information about the workshop, including online versions of the papers presented, can be found at at www.dcs.bbk.ac.uk/webDyn.

2 The Size of the Web Graph

One way of measuring the size of the Web is to estimate the number of pages that can be indexed by a major public search engine such as Altavista, Google or Fast. This measure is often referred to as the *publicly indexable Web* [55] and excludes pages that reside in searchable databases; these searchable databases are often referred to as the "invisible" or "deep" Web [13]. In February 1999, Lawrence and Giles estimated that the publicly indexable Web contains approximately 800 million pages, which is more than double their previous estimate in December 1997 of 320 million pages. A recent estimate in July 2000 by Murry of Cyveillance [61] reported that the current size of the Web is 2.1 billion unique pages and the growth rate is 7.3 million additional Web pages per day. A recent Altavista crawl of 200 million Web pages reported 1.5 billion links [21], which implies that the ratio of links to nodes is approximately 7.5. It is an open question how long the Web can and will sustain this kind of exponential growth.

Despite the size and predicted growth of the Web and the fact the largest search engines cover only between 25% to 30% of the Web, searching the Web can in many cases be carried out effectively by a combination of querying search engines, and navigation. One question which was recently tackled in [5] is what is the *diameter* of the Web or, in other words, what is the average shortest path between two randomly selected Web pages ? To answer this question Albert *et al.* [5] analysed the distribution of links in several Web sites concluding that they follow a power-law distribution (i.e an inverse polynomial distribution, see also [21]). Using this analysis they constructed random graphs following the observed power-law distribution and estimated that the diameter of the Web is 19, that is, on average any two Web pages are separated only by 19 clicks. Due to the logarithmic dependence of the diameter on the size of the Web, even if the Web size increases ten-fold the diameter will only increase from 19 to 21. Thus the Web graph is a *small-world network* [72], popularly known through the notion of "six degrees of separation", where any random two people can discover that there is only a short chain of at most six acquaintances that separates them. Consider a sparse network where nodes are connected to other nodes in their neighbourhood but otherwise the average distance between nodes is rather high. Watts and Strogatz [72] have shown that by making a small fraction of short-cuts between distant nodes in such a sparse network the average distance between any two nodes can be made small or, more precisely, logarithmic in the size of the graph, resulting in a small-world network. Recently, Adamic [4] has verified, on a Web crawl of 50 million Web pages, that the core component of the Web (see below) is a small-world network in the sense that Web sites tend to be clustered, yet only a few links separate one Web site from another.

A recent study by Broder *et al.* [21] has shown that the 19 degrees of separation for the Web does not reveal the full story, since the Web graph is not a connected graph but rather has the shape of a bow-tie (see Figure 1). The bow-tie has three major components: a knot and two bows, and a smaller fourth component. The knot, which contains about 30% of the Web, is the core component of the Web graph and is strongly connected, i.e. there is a directed path from any node in the core to any other node in the core. The left bow, which contains about 21.5% of the Web, consists of pages that can reach the core via a directed

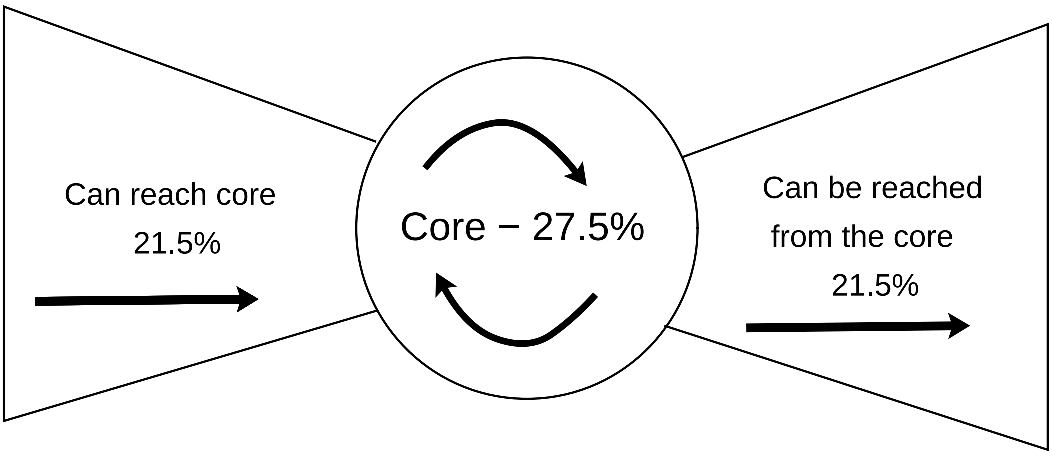

Figure 1: Bow-tie shape of the Web

path but cannot themselves be reached from the core via a directed path. The right bow, which also contains about 21.5% of the Web, consists of pages that can be reached from the core via a directed path but cannot themselves reach the core via a directed path. Finally, the last component, which also contains about 30% of the Web, is either disconnected from the bow-tie, can be reached from the left bow but is not in the core, or can reach the right bow but not from the core. The results in [21] show that 75% of the time there is *no* directed path from one random Web page to another random Web page. When such a path exists its average distance is 16 clicks for directed paths and only 7 clicks for undirected paths.

This analysis of the Web graph may provide clues on how to incorporate navigation, i.e. "surfing", into search engine technology in order to improve the "hit rate". For example, search engines such as Google are already using link analysis to improve the quality of relevance assessment of Web pages, and possible algorithmic solutions incorporating navigation are suggested in [57]. Kleinberg [49] investigated the algorithmic nature of small-world networks, when, on average, short paths exist between any two nodes. In particular, he was interested in the computational complexity of finding a short path between two nodes in such a network. Klienberg's setting is a two-dimensional grid where the likelihood of a node being connected to another node is proportional to an inverse function of its distance from that node, i.e. a function of the form $d^{-\alpha}$, where d is the distance and α is a non-negative parameter. That is, nodes that are close to each other are more likely to be connected than nodes that are far away from each other. In addition, the shortest path algorithm can only use local knowledge when deciding which link to follow next. What Kleinberg showed is that, in this model, only when $\alpha = 2$ is it computationally easy to get from a source node to a destination node, and in this case the algorithm is simply the "greedy" heuristic, i.e. at each stage choose to follow the node that brings you closest to the destination.

3 Key Issues in Web Dynamics

In this section we give a summary of some of the most important issues in the area of Web dynamics, with reference to the literature.

Web data mining [51] concerns the use of data mining techniques to discover, extract and analyse Web data. It can be categorised into three subareas: *Web content mining*, *Web structure mining* and *Web usage mining*. Web content mining is concerned with the discovery of useful information from Web documents. Much of the information to be analysed is unstructured text which is mined using statistical learning techniques [27]. In particular, in supervised learning or *classification* the learner uses training data with items labelled from a finite set of classes in order to build a classifier, and in unsupervised learning or *clustering* the learner, using some notion of document similarity, tries to group together similar documents. Web structure mining is concerned with discovering patterns in the link structure of the Web graph, for example discovering hubs and authorities [28, 45]. Another example is the *pagerank* algorithm used by the search engine Google [20], which interprets a link from page A to page B as a vote by page A for page B and thus considers a page to be "more important" if the votes it obtains are from "more important" pages. Web usage mining is concerned with the analysis of surfers Web log data in order to find patterns in their navigation behaviour — see [47, 56]. It is a fast growing area that has resulted in several workshops. A recent proposal for a novel algorithm for finding the users preferred navigation trails, based on modelling a collection of users navigation sessions as a probabilistic grammar, can be found in [16].

Related to Web structure mining is the notion of a *Web community*, which is a collection of Web pages sharing a common theme. For example, the XML Web community is such a community, which is represented by the collection of XML resource pages. In most cases such Web communities emerge spontaneously rather than being organised by a central body and thus represent the sociology of the Web. Kumar *et al.* [54] use a graph-theoretic approach for identifying Web communities based on the concept of *co-citation*, where two Web pages are co-cited if there is a link to both of them from a third page. In particular, they characterise a Web community as a dense directed bipartite subgraph which distinguishes between two types of page: *hub* pages which are pages that contain useful links to relevant pages on a particular topic and *authority* pages which are high quality pages for the given topic.

Collaborative filtering [18, 32] is a recent technology whose aim is to provide recommendations to a user based on preferences of other, similar, users. It is already being widely used to recommend music and books to consumers. For example, if you like a certain type of music and the system can locate other users with similar musical taste, this information can be used to recommend to you albums which were purchased by these users. Such systems also make use of ratings obtained from users on various items which are recommended directly to these users, in order to train the recommender algorithm to learn users preferences.

The ability to search and locate information is one of the fundamental enabling technologies necessary to realise the full potential of the Web. The field of *information retrieval* has a long history spanning the 1960's and 1970's, when the foundations of automatic text indexing and full text retrieval were established (see [69] for a comprehensive collection of historical papers). The Web poses a host of challenges for search engine developers, which has caused a revival in the information retrieval area. Some of the challenges are: coping with the dynamic and heterogeneous nature of Web information, coping with scalability and distribution issues, dealing with varying quality of information, and making use of the hypertextual nature of Web data.

The most common model for information retrieval used by search engines is the *vector space model*, where documents are viewed as vectors of weights assigned to the collection of their index terms. A query is also viewed a vector. The similarity of a document to a query is determined by the cosine of the angle between their vector representations. A review of alternative models such as probabilistic information retrieval models can be found in the recent book [9]; see also [44, 8] for overviews on how information retrieval is utilised by search engines.

Utilising the hyperlink information which is embedded in Web pages can substantially improve the quality of Web information retrieval. In particular, two recent influential algorithms are the *pagerank* algorithm [20] mentioned above, and the *HITS* (Hyperlink Induced Topic Search) algorithm [50] which computes hubs and authorities for a given search query; see [35] for a review on using the Web graph structure to aid search.

Web *crawlers* or *spiders* are mobile agents that traverse the Web in search of new and up-to-date information in order to refresh the databases of Web pages maintained by search engines for answering user queries [46]. To give an idea of the scale of these databases, Google and Fast have databases covering on the order of 500 million Web pages while AltaVista's database covers on the order of 300 million Web pages (see www.searchengineshowdown.com/stats/ for a detailed summary of search engine statistics).

One of the still unresolved problems in Web interaction, whereby users lose track of the context when following a sequence of links and are unsure how to proceed in terms of satisfying their original goal, is called the *navigation problem*. This phenomenon is also known as getting "lost in hyperspace". In order to resolve this problem navigation systems need to help users address the following three questions: where am I now? where have I been previously? and where can I go from here? A review on the navigation problem which includes the traditional methods of tackling it using orientation tools to provide contextual and spatial cues can be found in [57]. More recent techniques involve the use of machine learning and data mining techniques which perform a navigation task and adapt to the situation at hand [7, 57, 59, 67].

A recent emerging area which utilises both human-computer interaction and Web data mining techniques, is that of *adaptive hypertext* [22], whose aim is to build a model of an individual user of a hypertext system and apply this model for the purpose of adapting the system to that user. A distinction is made between *adaptive presentation*, which deals with adapting the contents of a Web page according to the user's knowledge and goal, and *adaptive navigational support*, whose aim is to help the user find the most relevant trails to follow by adapting the choice of links that the user can traverse.

Web site design is a growing area which needs to be investigated from several angles. Web site usability is one of the most important consideration in Web site design, since gaining users' attention is the most valuable commodity due to the fact that users are not willing to invest time in learning how to interact with the Web site's content. Nielson [62] advocates simplicity of use as the guiding principle in Web site design, with the aim of enabling users to carry out their main tasks in as simple a manner as possible. An approach which utilises both adaptive hypertext and Web data mining is that of adaptive Web sites advocated by [65]. The idea is that Web sites automatically improve their structure and presentation through the use of machine learning techniques. It is based on an algorithmic approach which clusters pages that co-occur in the Web site log data and then finds the concepts that best describe each cluster. One objective in Web site design is to minimise the cost to the user in locating a Web page, where the expected cost could be measured in terms of the frequency of visit to that Web page multiplied by the number of clicks necessary to reach the page starting from the home page or some other landmark page in the Web site; a step in this direction was taken by Aldous [6].

Due to the dynamic nature of the Web and its growth rate, it is important to build mathematical models which simulate the evolution of the Web. Several authors, notably [12, 17, 34], have proposed a stochastic model of the evolution of the Web leading to a power-law distribution. In these models new Web pages are continuously added to the Web graph at a certain rate. New links are added to the Web graph by means of preferential attachment, whereby Web pages that have more incoming links are more likely to obtain further incoming links i.e. success breeds success. These theoretical models have proven to be consistent with empirical evidence [21, 53]. It is interesting to note that the foundations for the stochastic models of the Web have their roots in the works of Simon 1955 [68] and de Solla Price 1976 [66]. Simon investigated a class of long-tailed distribution functions, which could, for example, model the frequency of words in a given text, and de Solla Price investigated empirical laws of bibliometrics such as those generated by citation data.

Being able to predict the rate of change of Web pages is important for the formulation of scheduling strategies in Web crawlers. Brewington and Cybenko [19] define a Web page indexed by a search engine to be β-*current* if the page did not change between the last time the crawler downloaded it and β time units ago i.e. if the page is at most β units of time out of date. A search engine is defined to be (α, β)-*current* if the probability that a randomly chosen Web page being β-current is at least α. As an example of the scale of the problem, AltaVista have reported that their main crawler, *Scooter*, visits on the order of 10 million Web pages per day, so it would take several months to cover a large portion of of the Web. Recent research [31] suggests that in order to keep the Web database of a search engine fresh, a crawler should be designed that continuously and incrementally updates its Web database, taking into account the estimated rate of change of pages.

So far in this section, we have discussed issues related to the management of *information* on the Web. Equally important is the management of *computation* over the Web, again in the face of change and growth of the Web and of its usage.

The Web is an ideal platform for supporting distributed applications over the Internet, due to its open architecture which facilitates the integration of heterogeneous resources [43]. *Data-intensive Web applications* pose specific challenges due to the volume, heterogeneity and dynamicity of their data, and the numbers and diversity of their users. In order to make such applications robust, scalable, and extensible, it is important to separate the data management issues, the specification of the structure and content of the Web site, and the visual presentation of the Web pages. Fernández *et al.* describe a data-intensive Web site implementation system called STRUDEL [42] whose underlying data model is a semistructured one [2]. In STRUDEL the data graph of the Web site is specified as a query in a semistructured language which can also be used for restructuring the data. The HTML rendering of the Web site is generated separately using a template language. Fraternali [43] gives a comprehensive survey of tools and approaches to developing data-intensive Web applications.

Web-enabled and *Web-based workflow management systems* coordinate and schedule the heterogeneous and distributed activities related to business processes over the Web [15, 38]. In the dynamic environment of the Web, support for workflow *evolution* is a major issue. The Vortex language [39] is a rule-based workflow specification language targeted at end-users. It is provides a predefined set of decision modules for combining workflow decision logic. It's modularity and rule-based approach provide a powerful basis for constructing and modifying workflow schemas. Kradolfer and Gepper [52] discuss a multi-version framework for supporting the run-time evolution of workflow schemas. Currently running workflow instances are migrated to the new workflow schema if they are compatible with it, otherwise an older version of the schema is maintained for as long as needed.

E-Commerce is a major Web-based application application which promises to improve economic efficiency, reduce prices, speed-up complex business deals, and reach global markets [40, 41]. E-commerce is expected to exceed one trillion dollars by 2002. In order to realise a satisfactory service there are serious technical challenges to overcome in areas such as privacy and security, content management, business transaction management, distributed data management, online negotiation protocols, and establishing trust. In the area of online negotiation, the *auction* paradigm has gained widespread usage as a mechanism for price discovery and differentiation [71]. For example, consumer auctions are projected to grow to 15.5 billion in revenue in 2001, and business-to-business and business-to-consumer auctions are projected to become much larger markets.

Event-condition-action (ECA) rules are a significant computational paradigm deployed in a variety of ways in Web applications: specifying and implementing Web-based business processes [1, 3, 26, 70], encoding user preferences in personalisation and push technology [64, 26], encoding negotiation strategies in Web-based negotiation [37], and change notification [58]. ECA rules have their origin in production systems such as OPS5 and have been much studied in the context of modern-day database systems where they are also known as *triggers* and are supported in some form by the major commercial DBMSs. Triggers make a database able to automatically respond to the occurrence of events by the execution of pre-programmed actions provided specified conditions hold. These actions may in turn cause further events to occur, which in turn cause more triggers to execute. Thus, triggers turn a "passive" database into an "active" database. Comprehensive collections of papers on ECA rules in the context of databases can be found in [73, 63].

The use of triggers in Web applications poses two major challenges that are not met by today's active database technologies: firstly, there is a need for more powerful techniques and tools for analysing and verifying the behaviour of evolving ECA rule sets before they are deployed; secondly, there is a need for optimisation techniques that will enable trigger processing to scale up from the hundreds of triggers that can be handled effectively by today's databases to the much larger (e.g. 3-4 orders of magnitude) numbers of triggers that can be generated by Web applications. For the first of these problems, we believe that our recent work on using abstract interpretation for static analysis of triggers is a promising foundation for more powerful trigger analysis tools — see [10, 11] for a description of our work and references to other static analysis techniques for triggers. For the second of the problems, work has recently been done into more compact encodings of the event and condition parts of triggers and more efficient matching of event occurrences against them [64, 36, 30, 58] and these approaches promise to achieve the necessary scalability in trigger management and processing.

Mobile agents [60] are software fragments that can migrate across different parts of a network, and are thus important in the general context of distributed systems and in the specific context of the Web. Examples of mobile agent applications include: data-intensive applications where the data is remotely stored and an agent needs to perform the task on a remote server, applications where agents launched from an appliance, such as a PDA, to a remote server to carry out a task for the user, and information retrieval situations where an agent is sent to collect some piece of information from remote sites. Huhns and Singh [48] contains a comprehensive collection of papers on agent technology.

With this kind of *mobile computation*, agents and devices can dynamically connect to the network, change location, or become unavailable, so no fixed topology or set of resources can be assumed for the network. Cardelli and Gordon have proposed the *Ambient Calculus* [24] as a general model of mobile computation. This is a process calculus where processes reside at the nodes of a dynamically evolving hierarchy of locations. The *Ambient Logic* [25] is a modal logic of space and time that has the Ambient Calculus as a model. This logic includes spatial connectives as well as the standard and temporal connectives and can be used to formally specify and verify properties holding at particular locations of the hierarchy, and the dynamic evolution of the hierarchy. Since it reasons about labelled trees, it turns out that the Ambient logic is also applicable to querying semistructured *data*, and thus has the potential to serve a unifying formalism for both semistructured data querying and semistructured computation on the Web [23].

4 Summary of the Web Dynamics Workshop

There were three invited talks at the Web Dynamics workshop. The first was given by Soumen Chakrabarti from the Indian Institute of Technology, Bombay, on the topic of hyperlink analysis for the purpose of enhanced ranking of Web pages in response to a query. In the first part of his talk Chakrabarti introduced a learning system called a *focused crawler* [29], which is a specialised crawler with the objective of seeking out Web pages on a pre-defined set of topics (see also [33]). This is in contrast to general-purpose crawlers utilised by search engines to create their offline databases of Web pages. The focused crawler starts from a small seed-set of topic-related pages and during the crawl the classifier used by the system decides at each stage which outlinks to follow from each fetched pages according to their potential relevance to the query. In the second part of the talk, a fine-grained model for extracting *micro-hubs* was introduced, which is motivated by the fact that hubs computed by the HITS algorithm, mentioned above, are often *mixed* in the sense that only specific regions within these pages are actually relevant to the query. The algorithm for identifying micro-hubs within mixed-hubs is based on the minimum description length principle.

The second invited talk was given by Knut Magne Risvik, Director of Search Technology at Fast Search & Transfer, on the topic of search engine challenges in the face of the rapidly growing and dynamic Web. Fast's search engine (www.alltheweb.com) is one of the largest, covering on the order of half a billion Web pages. In order to maintain its relative coverage and retrieval quality, Fast must be able to deal with a wide variety of file formats including text and multimedia, the problem of scalability of its architecture, and the problem of efficient crawling and duplicate detection in order to maintain the quality of its offline database. Risvik noted that, weighted by site popularity, the average age of an interesting Web page is less than one week and thus search engine crawlers must be adaptive and must index dynamic Web sites more frequently. This implies that the age of a Web page is an important factor in measuring its relevance. In fact, it may be the case that dynamic Web sites render current crawling practices redundant and some searching will have to be carried out in real-time. Finally, to deal with the issue of scalability a search engine must decide what content is important and for this purpose usage statistics may be helpful.

The third invited talk was given by Luca Cardelli from Microsoft Research Labs, Cambridge, on the topic of formal foundations for specifying and verifying the properties of mobile computation. In his talk Cardelli described a modal logic for the Ambient Calculus which is an extension of the Ambient Logic to express issues of privacy and security. Cardelli discussed how this logic can be applied to areas such as the specification of mobility protocols and mobility policies, and for model checking of security and privacy properties of mobile computations.

An invited talk that related to the theme of the workshop was given at the ICDT conference following the workshop (4-6 January 2001) by Andrei Broder, Vice President for Research and Chief Scientist at AltaVista (www.altavista.com). Broder talked about the explosive growth of unstructured search and the move to third-generation search engines. First-generation search engines were based on classical information retrieval models, such as the vector space model, and support HTML parsing and weighting. Second-generation search engines employ link analysis, such as the Google pagerank algorithm, and utilise anchor text in order to provide some navigational support. Third-generation search engines aim to reflect users' needs by detecting the context of a query. This context could be spatial, textual, the user profile, or previous user queries based on data mining analysis. Third-generation search engines also have to deal with dynamic Web-page content, an issue that was also raised by Risvik in his talk. Finally, Broder stated that Intranet users querying a local search engine are now expecting a similar interaction style as with global search engines, rather than the more traditional structured query language style.

Four papers and one demonstration were presented at the workshop. Géry and Chevallet from the MRIM Research Laboratory, University of Grenoble, described experiments they have carried out for extracting structure from HTML Web pages in order to develop an improved information retrieval model on which search engine technology may be based. Their work is aimed at the validation of two hypotheses. The first is that information units on the Web come in different granularities such as a section in a document, a whole document, or a cluster of documents within a Web site. The second hypothesis is that there are a variety of relationships between documents such as hierarchical, sequence and reference. From an experimental analysis they reach the conclusion that the first hypothesis is correct. Regarding the second hypothesis they conclude that most links within a Web site are internally linked to other pages within the Web site and that there is strong evidence that links follow hierarchical paths emanating from the home page.

Omelayenko from the Division of Mathematics and Computer Science, Vrije University Amsterdam, provided a survey of the state-of-the-art in ontology learning from the Web. He distinguishes between three types of ontologies: natural language ontologies, domain ontologies, and ontology instances. Natural language ontologies capture background knowledge and are used to expand user queries. Domain ontologies capture specific domain knowledge; they are usually constructed manually but machine learning techniques can aid their construction. Ontology instances are actual Web pages many of which will be written in XML. In order to realise Berners-Lee's vision of a *semantic Web* [14], machine learning techniques are being developed to automate ontology learning tasks, with the aim of producing human readable and understandable representation of knowledge. Omleayenko surveys various approaches to learning ontologies including clustering, association rule detection, decision tree learning and naive Bayes.

Ishikawa and Ohta from the Department of Electronics and Information Engineering, Tokyo Metropolitan University, discussed a new modelling language called XBML for specifying the dynamic aspects of business processes in e-commerce. This language integrates active database functionality in the form of event-condition-action rules with a query language for XML-based databases. The event-condition-action rules are used to specify the flow of control in the business processes. The paper discusses how a variety of e-commerce business processes can be implemented using this language, including the auction model. The paper also describes the architecture and implementation of the XBML distributed system.

Cannataro, Cuzzocrea, and Pugliese from the Institute of Systems Analysis and Information Technology in Rende presented a probabilistic approach to adaptive hypertext. Hypertext is modelled as a weighted directed multigraph of elementary abstract concepts, each concept being represented by a weighted directed multigraph of presentation descriptions which map onto XML documents. User models, or *profiles*, describe users' characteristics and preferences. Information can be collected about users' behaviour to build a discrete distribution that measures the extent to which a user belongs to each possible profile. The authors also propose an architecture for a prototype implementation of their approach.

Bouras and Konidaris from the Computer Science Institute and the Computer Engineering and Informatics Department, University of Patras, discussed run-time management policies for data-intensive Web sites, with the aim of reducing Web latency i.e. the delay between the time a user makes a request and the time the page reaches the user's computer. They propose a hybrid model between on-the-fly creation of dynamic Web pages at the time they are requested and pre-computation and caching of all dynamic pages prior to them being requested. The tradeoff between consistency of the cache and query efficiency is described by a *compromise factor* which states the acceptable number of changes to a dynamic Web page before a new copy of the page is cached. The authors describe experiments they carried out with their hybrid approach with different compromise factors showing when performance can be improved.

Levene, Wheeldon and Bitmead from the Department of Computer Science at University College London (now at Birkbeck College, University of London) and NavigationZone Ltd., demonstrated a navigation tool for "surfing" within Web sites. This tool semi-automates user navigation by proposing to users a set of preferred trails that are relevant to their query. This is in contrast to a site search engine which merely outputs a list of pages which are relevant to the user query without addressing the problem of which trail the user should follow. The navigation tool addresses the *navigation problem* of users "getting lost in hyperspace", when they lose the context in which they are browsing and are unsure how to proceed in terms of satisfying their original goal. The alpha release of the system operates over the University College London (UCL) Web site. A screen shot of the output of the navigation tool for the query "computers" can be seen in in Figure 2, The results from the navigation engine help the user to find the context within the Web site of the information pertaining to "computers" by displaying the trails which are relevant to this query. For example, the first trail indicates that the Information Systems Division (ISD) at UCL provides computing services for students, while the second trail implies that ISD provides a help desk facility including Macintosh support. Search engine's results do not provide this contextual information. The best it can do for this type of query is provide a good starting point for the user to initiate a navigation session.

An additional paper by Maurer, Huberman and Adar from Xerox Palo Alto Research Centre was accepted for the workshop but the authors could not attend the workshop to deliver the paper. In this paper Maurer *et al.* provide an analysis of the optimal size of a *Web ring*, which is a Web structure where sites are linked together to form a ring. The analysis utilises their *law of "surfing"* [47] based on a random walk model where the user has

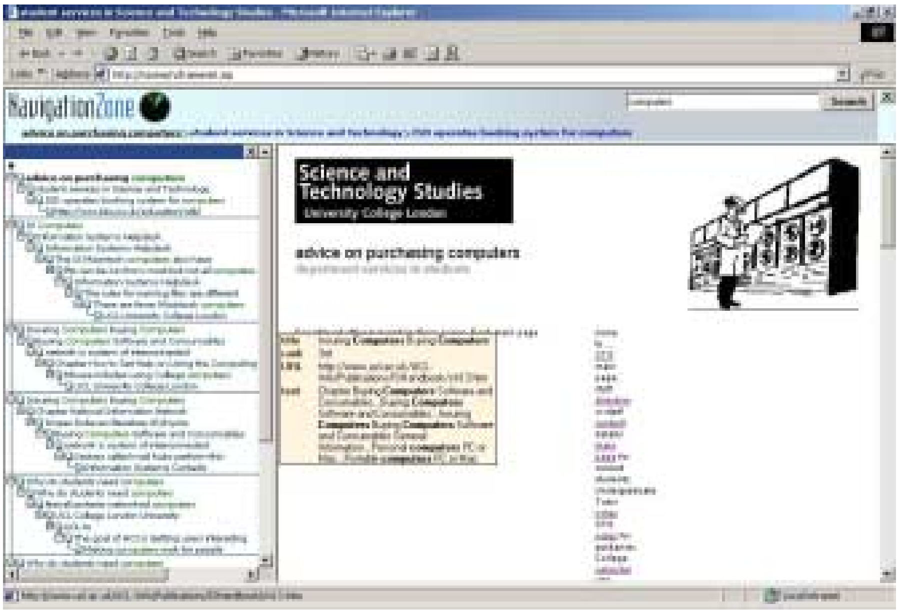
NavigationZone
Search
Science and Technology Studies
University College London
advice on purchasing computers

Figure 2: Navigation engine results for query "computers"

to decide whether it is beneficial to continue or stop "surfing". Under the assumptions of their model the authors show that it is beneficial to split a large ring into two equally sized smaller subrings if, when split into these subrings, the probability of choosing the correct subring is greater than one half. In conclusion they note that combing their analysis with machine learning algorithms may lead to automation of optimal organisation of Web sites.

5 Conclusions

There is a growing body of work in the area of Web dynamics both theoretical and practical. We are beginning to understand how the Web is evolving and what properties it has as a massive graph. As we have shown, there are many relevant subareas dealing with different aspects of Web dynamics, which are contributing to advances in Web technologies. Still it is hard to predict what the Web will look like one year from now, since, apart from technological factors, it is also influenced by economic, social and political factors. There are many unresolved problems to do with the scale of information and computation management in such a heterogenous, distributed and dynamic environment. With the advances in Web technologies it is important not forget the user and the information overload the Web creates. Thus an important challenge relating to many of the areas of Web dynamics is that of understanding and acting upon user *context*. We need to further develop algorithmics that make use of knowledge about what users are doing now, what they did previously, how users can be associated with similar groups of users, and the kinds of Web content and Web-based activities that users are interacting with.

References

[1] S. Abiteboul, B. Amann, S. Cluet, A. Eyal, L. Mignet, and T. Milo. Active views for electronic commerce. In *Proc. 25th Int. Conf. on Very Large Databases*, pages 138–149, 1999.

[2] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistrucutred Data and XML*. Morgan-Kaufmann, San Francisco, Ca., 2000.

[3] S. Abiteboul, V. Vianu, B.S. Fordham, and Y. Yesha. Relational transducers for electronic commerce. *JCSS*, 61(2):236–269, 2000.

[4] L.A. Adamic. The small world web. In *Proceedings of European Conference on Research and Advanced Technology for Digital Libraries*, pages 443–452, Paris, 1999.

[5] R. Albert, H. Jeong, and A.-L. Barbási. Diameter of the world-wide web. *Nature*, 401:130, 1999.

[6] D.J. Aldous. Reorganizing large web sites. *American Mathematical Monthly*, 108:16–27, 2001.

[7] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. WebWatcher: A tour guide for the World Wide Web. In *Proceedings of AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Sources*, Stanford, Ca., 1995.

[8] R. Baeza-Yates. Searching the web: Challenges and partial solutions. In *IEEE Workshop on String Processing and Information Retrieval*, Santa Cruz, Bolivia, 1998.

[9] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press and Addison-Wesley, Reading, Ma., 1999.

[10] J. Bailey and A. Poulovassilis. Abstract interpretation for termination analysis in functional active databases. *JIIS*, 12 (2/3):243–273, 1999.

[11] J. Bailey and A. Poulovassilis. An abstract interpretation framework for termination analysis of active rules. In *Research Issues in Structured and Semistructured Database Programming (Proc. DBPL'99)*, LNCS 1949, pages 249–266. Springer-Verlag, 1999.

[12] A.-L. Barbási, R. Albert, and H. Jeong. Mean-field theory for scale free random networks. *Physica A*, 272:173–189, 1999.

[13] M.K. Bergman. The deep web: Surfacing hidden value. White paper, Bright Planet, July 2000.

[14] T. Berners-Lee. *Weaving the Web*. Orion Books, London, 1999.

[15] G.A. Bolcer and G. Kaiser. SWAP: Leveraging the web to manage workflow. *IEEE Internet Computing*, 3:85–88, 1999.

[16] J. Borges and M. Levene. A fine grained heuristic to capture web navigation patterns. *SIGKDD Explorations*, 2:40–50, 2000.

[17] S. Bornholdt and H. Ebel. World-Wide Web scaling exponent from Simon's 1955 model. *Condensed Matter Archive*, cond-mat/0008465, 2000.

[18] J.S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predicitve algorithms for collaborative filtering. In *Proceedings of Conference on Uncertainty in Artificial Intelligence*, Madison, Wisc., 1998.

[19] B.E. Brewington and G. Cybenko. Keeping up with the chaanging web. *IEEE Computer*, 33:52–58, 2000.

[20] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of International World Wide Web Conference*, pages 107–117, Brisbane, 1998.

[21] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, A. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph strucutre in the web. *Computer Networks and ISDN Systems*, 30:309–320, 2000.

[22] P. Brusilovsky, A. Kobsa, and J. Vassileva, editors. *Adaptive Hypertext and Hypermedia.* Kluwer, Dordrecht, 1998.

[23] L. Cardelli. Semistructured computation. In *Research Issues in Structured and Semistructured Database Programming (Proc. DBPL'99)*, LNCS 1949. Springer-Verlag, 1999.

[24] L. Cardelli and A.D. Gordon. Mobile ambients. In *Proc. FoSSaCS'98*, LNCS 1378, pages 140–155. Springer-Verlag, 1998.

[25] L. Cardelli and A.D. Gordon. Anytime, anywhere. modal logics for mobile ambients. In *Proc. 27th ACM Symp. on Principles of Programming Languages (POPL'00)*, pages 365–377, 2000.

[26] S. Ceri, P. Fraternali, and S. Paraboschi. Data-driven one-to-one web site generation for data-intensive applications. In *Proc. 25th Int. Conf. on Very Large Databases*, pages 615–626, 1999.

[27] S. Chakrabarti. Data mining for hypertext: A tutorial survey. *SIGKDD Explorations*, 1:1–11, 2000.

[28] S. Chakrabarti, B. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J.M. Kleinberg. Mining the web's link structure. *IEEE Computer*, 32:60–67, 1999.

[29] S. Chakrabarti, M. Van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of International World Wide Web Conference*, pages 1623–1640, Montreal, 1999.

[30] J. Chen, D. DeWitt, F. Tian, and Y. Wang. Niagaracq: a scalable continuous query system for internet databases. In *Proc. 2000 ACM SIGMOD Int. Conf.on Management of Data*, pages 379–390, 2000.

[31] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of International Conference on Very Large Data Bases*, pages 200–209, Cairo, Egypt, 2000.

[32] W.W. Cohen and W. Fan. Web-collaborative filtering: Recommending music by crawling the web. In *Proceedings of International World Wide Web Conference*, Amsterdam, 2000.

[33] M. Diligenti, F. Coetzee, S. Lawrence, C.L. Giles, and M. Gori. Focused crawling using context graphs. In *Proceedings of International Conference on Very Large Data Bases*, pages 527–534, Cairo, Egypt, 2000.

[34] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. WWW and internet models from 1955 till our day and the "popularity is attractive" principle. *Condensed Matter Archive*, cond-mat/0009090, 2000.

[35] K. Efe, V.V. Raghavan, C.H. Chu, A.L. Broadwater, L. Bolelli, and S. Ertekin. The shape of the web and its implications for searching the web. In *Proceedings of International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, Rome, 2000.

[36] E. Hanson et al. Scalable trigger processing. In *Proc 15th Int. Conf. on Data Engineering (ICDE'99)*, pages 266–275, 1999.

[37] J. Hammer et al. The ideal approach to internet-based negotiation for e-business. In *Proc 16th Int. Conf. on Data Engineering (ICDE'2000)*, pages 666–667, 2000.

[38] J.A. Miller et al. Webwork: Meteor$_2$'s web-based workflow management system. *JIIS*, 10(2):185–215, 1998.

[39] R. Hull et al. Declarative workflows that support easy modification and dynamic browsing. In *Proc ACM Int. Joint Conf. on Work Activities Coordination and Collaboration (WACC'99)*, pages 69–78, 1999.

[40] S. Feldman. The changing face of e-commerce: Extending the boudaries of the possible. *IEEE Internet Computing*, 4:82–83, 2000.

[41] S. Feldman. Electronic marketplaces. *IEEE Internet Computing*, 4:93–95, 2000.

[42] M. Fernáindez, D. Florescu, A. Levy, and D. Sucui. Declarative specification of web sites with STRUDEL. *The VLDB Journal*, 9:38–55, 2000.

[43] P. Fraternali. Tools and approaches for developing data-intensive web applications - a survey. *ACM Computing Surveys*, 31(3):227–263, 2000.

[44] V.N. Gudivada, V.V. Raghavan, W.I. Grosky, and R. Kasanagottu. Information retrieval on the world wide web. *IEEE Internet Computing*, 1:58–68, 1997.

[45] M.R. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 1:45–50, 2001.

[46] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2:219–229, 1999.

[47] B.A. Huberman, P.L.T. Pirolli, J.E. Pitkow, and R.M. Lukose. Strong regularities in world wide web surfing. *Science*, 280:95–97, 1998.

[48] M.N. Huhns and M.P. Singh, editors. *Readings in Agents.* Morgan-Kaufmann, San Francisco, Ca., 1998.

[49] J. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[50] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, 1999.

[51] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2:1–15, 2000.

[52] M. Kradolfer and A. Geppert. Dynamic workflow schema evolution based on workflow type versioning and workflow migration. In *Proc 6th Int. Conf. on Cooperative Information Systems (CoopIS'1999)*, pages 104–114, 1999.

[53] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. The web as a graph. In *Proceedings of ACM Symposium on Principles of Database Systems*, pages 1–10, Dallas, Tx., 2000.

[54] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of International World Wide Web Conference*, pages 1481–1493, Montreal, 1999.

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

[56] M. Levene, J. Borges, and G. Loizou. Zipf's law for web surfers. *Knowledge and Information Systems*, 3:120–129, 2001.

[57] M. Levene and G. Loizou. Web interaction and the navigation problem in hypertext. In A. Kent, J.G. Williams, and C.M. Hall, editors, *Encyclopedia of Microcomputers.* Marcel Dekker, New York, NY, 2001. To appear.

[58] L. Liu, C. Pu, and W. Tang. Supporting internet applications beyond browsing: Trigger processing and change notification. In *Proc 5th Int. Computer Science Conf (ICSC'99), Hong Kong*, pages 294–304, 1999.

[59] F. Menczer and R.K. Belew. Adaptive information agents: Internalizing local context and scaling up to the web. *Machine Learning*, 39:203–242, 2000.

[60] D. Milojicic. Trend wars - Mobile agent applications. *IEEE Concurreny*, 7:80–90, 1999.

[61] B.H. Murray and A. Moore. Sizing the internet. White paper, Cyveillance, July 2000.

[62] J. Nielsen. *Designing Web Usability: The Practice of Simplicity.* New Riders Publishing, Indianapolis, Indiana, 2000.

[63] N. Paton. *Active Rules in Database Systems.* Springer-Verlag, 1999.

[64] J. Pereira, F. Fabret, F. Llirbat, and D. Shasha. Efficient matching for web-based publish/subscribe systems. In *Proc 7th Int. Conf. on Cooperative Information Systems (CoopIS'2000)*, pages 162–173, 2000.

[65] M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence*, 118:245–275, 2000.

[66] D. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society of Information Science*, 27:292–306, 1976.

[67] J. Rennie and A.K. McCallum. Using reinforcement learning to spider the web efficiently. In *Proceeding of the International Conference on Machine Learning (ICML)*, Bled, Slovenia, 1999.

[68] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[69] K. Sparck Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan-Kaufmann, San Francisco, Ca., 1997.

[70] M. Spielmann. Verification of relational transducers for electronic commerce. In *Proc. 19th ACM Symp. on Principles of Database Systems (PODS'2000)*, pages 92–103, 2000.

[71] M. Ströbel. On auction as the negotiation paradigm of electronic markets. Research Report RZ 3166 (93212), IBM Research Laboratory, Zurich, 1999.

[72] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[73] J. Widom and S. Ceri. *Active Database Systems*. Morgan Kaufmann, 1996.