A Survey: World Wide Web and the Search Engines

Bharat Bhushan Agarwal
Research Scholar
Department of Computer Science & Engineering
Teerthanker Mahaveer University (TMU)
Moradabad, India
TEN 10 Ph.D. /14.

Sonia Gupta
Department of Computer Science & Engineering
IFTM University Moradabad (INDIA)

ABSTRACT
The World Wide Web is one of the most popular and quickly growing aspects of the Internet. Ways in which computer scientists attempt to estimate its size vary from making educated guesses, to performing extensive analyses on search engine databases. We present a new way of measuring the size of the World Wide Web using "Quadrat Counts", a technique used by biologists for population sampling. There has been an exponential growth in hypermedia and web modeling languages in the market. This growth has highlighted new problems and new areas of research. This paper categorizes and reviews the main hypermedia and web modeling languages showing their origin and their primary focus. It then concludes with recommendations for further research in this field. When automatically extracting information from the world wide web, most established methods focus on spotting single HTML documents. However, the problem of spotting complete web sites is not handled adequately yet, in spite of its importance for various applications. Therefore, this paper discusses the classification of complete web sites. First, we point out the main differences to page classification by discussing a very intuitive approach and its weaknesses. This approach treats a web site as one large HTML-document and applies the well-known methods for page classification. Next, we show how accuracy can be improved by employing a preprocessing step which assigns an occurring web page to its most likely topic. The determined topics now represent the information the web site contains and can be used to classify it more accurately. We accomplish this by following two directions. First, we apply well established classification algorithms to a feature space of occurring topics. The second direction treats a site as a tree of occurring topics and uses a Markov tree model for further classification. To improve the efficiency of this approach, we additionally introduce a powerful pruning method reducing the number of considered web pages. Our experiments show the superiority of the Markov tree approach regarding classification accuracy. In particular, we demonstrate that the use of our pruning method not only reduces the processing time, but also improves the classification accuracy.

Keywords WWW size estimation using biological techniques, Modelling, Methodologies, Web application, Web site, Hypermedia, Hypertext.

1. INTRODUCTION
How much of the Internet does the World Wide Web (WWW or simply Web) actually populate? Are web servers abundant throughout the 'Net, or is the vastness of cyber space still a relatively empty void?

Unfortunately, due to the size and dynamic nature of the WWW, it is infeasible to conduct an exhaustive census. Instead, recent studies have attempted to estimate its size using a variety of techniques. We take an interdisciplinary approach to the problem of estimating the size of the WWW by adopting a technique utilized by biologists who conduct population sampling using quadrat counts – a new, and intuitive way of looking at the problem.

This is confirmed by research into the challenges of hypermedia and web development and complemented by a survey of Irish companies at the National University of Ireland (Barry and Lang, 2001:30, Lang 2001b). The survey suggests that the hypermedia development markets are relying on ad hoc or in-house build methodologies. Standard index-based search engines on the web have astonishing properties: they have over a billion of web documents in their index, they can handle millions of requests per day, they give voluminous answers almost in real time, they require huge human and computer resources. But even if those search engines have many positive advantages, they also have drawbacks, like for instance rather simple queries, a presentation of the results which is poor of information, or the fact that the user must very often explore a lot of results by himself before finding interesting pages. In this paper, we make the assumption that the user can wait for his results during one or two hours for instance but provided that he spends only a very short time in manual analysis of the results. Our final aim is to define a search engine for strategic watch that establishes at given time interval (one day for instance) a complete report on how a given subject is presented on the Web. Our approach is complementary to standard index-based search engines.

2. PREVIOUS STUDIES
The past decade has witnessed the birth and explosive growth of the World Wide Web, both in terms of content and user population. Figure 1 shows the exponetial growth in the number of Web servers. The number of users online has been growing exponentially as well. Whereas in 1996 there were 61 million users, at the close of 1998 over 147 million people had internet access worldwide. In the year 2000, the number of internet users more than doubled again to 400 million. With its remarkable growth, the Web has popularized electronic commerce, and as a result an increasing segment of the world's population conducts commercial transactions online.

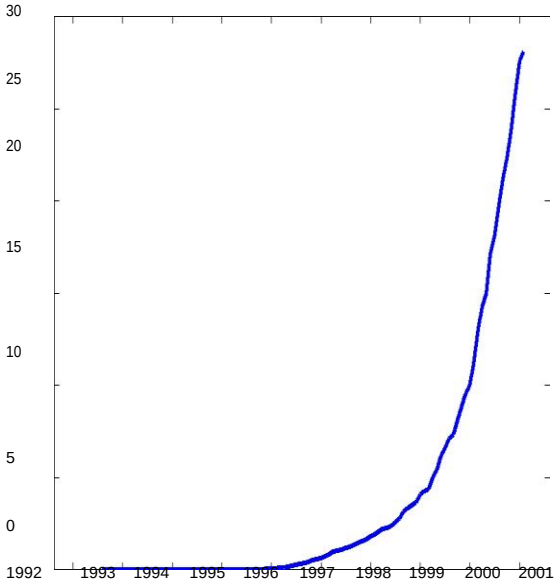
From its very onset, the Web has demonstrated a tremendous variety in the size of its features. Surprisingly, we found out that there is order to the apparent arbitrariness of its growth. One observed pattern is that there are many small elements contained within the web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others. This diversity can expressed in mathematical fashion as a distribution of a particular form, called a power law, meaning

that the probability of attaining a certain size x is proportional to 1/x to a power τ, where τ is greater than or equal to1.

When a distribution of some property has a power law form, the system looks the same at all length scales. What this means is that if one were to look at the distribution of site sizes for one arbitrary range, say just sites which have between 10,000 and 20,000 pages, it would look the same as for a different range, say 10 to 100 pages. In other words, zooming in or out in the distribution, one keeps obtaining the same result. It also means that if one can determine the distribution of pages per site for a range of pages, one can then predict what the distribution will be for another range.

Power laws also imply that the average behavior of the system is not typical. A typical size is one that is encountered most frequently, while the average is the sum of all the sizes, divided by the number of sites. If one were select a group of sites at random and count the number of pages in each one, the majority of the sites would be smaller than average. This discrepancy between average and typical behavior is due to the skew of the distribution. Equally interesting, power law distributions have very long tails, which means that there is a finite probability of finding sites extremely large compared to the average.

Figure 1: Growth in the number of web servers 1992-2001



3. WEB MODELLING LANGUAGES:

We can identify three categories of web-modelling language: structured modelling, object-oriented modelling, and other independent modelling languages that are based on hypermedia and human computer interface (HCI) principles.

3.1 STRUCTRED MODELLING

These modelling languages are based on traditional tools for software development such as the Entity Relationship Diagram (E-R Diagram), Data Flow Diagram (DFD), and other diagrams such as Flow charts. The E-R diagram is the focus of much of the work done here. Entity relationship modelling was one of the first data modelling techniques to be developed and has become very popular, with numerous texts introducing it for the use of systems analysis and design. E-R diagrams are conceptual models, distant from any given implementation. It is central to most structured system development methodologies such as Structured System Analysis and Design Method (SSADM). Thus, by using this notation, web developers would be buying into a model that is already popular among designers and understood by customers, instead of hoping that they would both find time to learn a new formalism.

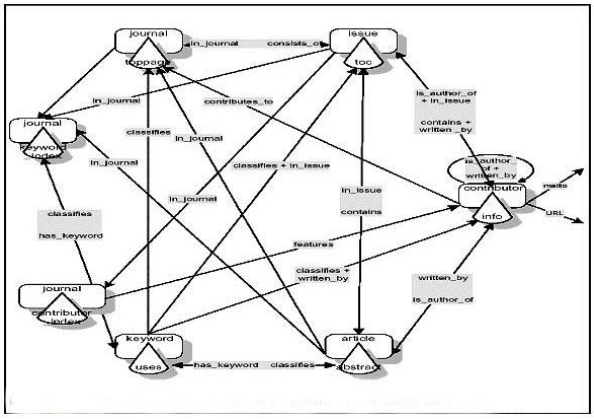


Figure 2. Complete RMM application diagram

We can see the temptation by web model developers in modelling web applications using E-R diagrams. In principle any serious web application would need a pool of data, a database, which would support its activities and transactions. This is especially relevant when almost all databases developed are structural relational databases. However, one of the questions that arise is how an E-R diagram or even an extended version of its model represents the complex navigational and multimedia object links that a web application requires.

3.2 Object-Oriented based web modelling languages

Object oriented principles are based on modelling our environment as a set of objects. These objects have three sections: A name, a set of attributes that belong to the object, and a set of operations (behaviour /methods) that are performed by this object. Object oriented principles are a relatively new and radical way of viewing programming compared to structured modelling. The principal decomposition of components and sub-components is shifted from the traditional procedures and functions to the object-oriented encapsulation of objects that have attributes and operations.

Object-Oriented Hypermedia Design Method (OOHDM) uses abstraction and composition mechanisms in an object oriented framework to allow a concise description of complex information items, and on the other hand it allows the specification of navigation patterns and interface transformations.

3.3 Hypermedia and other web modelling languages

One of the more surprising areas to contribute to web application modelling is that of hypermedia modelling. Much

of the characteristics of hypermedia application development are shared by web applications development. Therefore, several hypermedia modelling tools have been imported to be applied for Web modelling. We briefly introduce more strategy used in hypermedia modelling design: the Lite-HDM.

Lite-Hypermedia Design Modelling (Lite-HDM), is a design notation that claims to support the specifications of the structural, navigational, and presentational semantics of the application (in our case, web applications). Lite-HDM conceptual model is an evolution of Hypermedia Design Model (HDM) and is the union of three perspectives: a hyper-schema, an access schema and a presentation schema (Auto web, 2002).The Lite-HDM is used by Auto Web application in a CASE tool that would allow the creation of a web application that is model driven. The developer does, however, acknowledge some drawbacks in the process, mainly that the final product requires significant customization (Auto web, 2002). Our focus will be on the modelling language used.

Web Modelling Language (WebML) is a notation for specifying web applications at the conceptual level. WebML claims to enable the high-level description of a web application under distinct orthogonal dimensions: its data content (structure model), the pages that compose it (composition model), the topology of links between pages (navigation model), the layout and graphic requirements for page rendering (presentation model), and the customization features for one-to-one content delivery (personalization model) (Ceri, 2000). Figure 2 shows another example on how a modelling language models the navigational aspects of a Web application. In the case of figure 3, (Ceri, 2000) demonstrates using the example of an online CD shop.

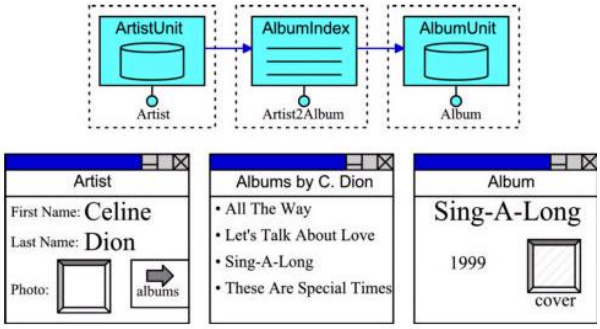


Figure 3. Modelling notation for Contextual Navigation in WebML (Ceri 2000)

4. SEARCHING THE WEB AS AN OPTIMIZATION PROBLEM

Table 1. Modeling the problem of information search on the Web as an optimization problem

Optimization	Web	Notation
Search space	S	Set of pages/documents
Fitness function	f: S R+	Relevance of the page w.r.t. user request
Optimal solution	s* = arg maxs Î Sf(s)	Page that maximizes relevance
Neighborhood	V: S Sk	Links going out of a page
Local search operator	O: S Sk	Exploration of a page's link

The search space S of our optimization problem is the set of Web pages and is structured with neighborhood relationship V: S ® Sk between the points of S thanks to the links between pages. We associate to this search space an evaluation or fitness function f: S ® R+ which can numerically evaluate web pages. A search engine tries to output pages which maximize this function, and thus tries to solve this optimization problem.

5. A GENETIC BASED SEARCH ENGINE

5.1 Main algorithm

1. Get the user request and Define the evaluation function f,

2. Pop ← Ø (initially empty, the population will progressively grow until it reaches a size of PopMax),

3. Generate an offspring page O:

4. With probability (1- Pmut) (or if |Pop| < 2) Then O ←heuristic creation (page from standard search engines)

5. With probability Pmut Then Select one parent page P from the best pages in Pop and let O←Mutation(P) (P's links exploration)

6. Evaluate f(O)

7. Insert O in Pop if (|Pop| < PopMax) or if f(O) is greater than the fitness of the worst page in Pop which is deleted,

8. Go to 3 or Stop (Pop is the output given to the user).

5.2 Genetic operators and other search mechanisms

We use a heuristic creation operator which outputs an address of a web page from the results given by five standard search engines (Altavista, Google, Lycos, Voila, Yahoo). It consists in querying each search engine with the keywords (K1, K2, ...) and in extracting the results. The links found are stored in a list sorted in the same order given by each search engine (1st link of the 1st engine, 1st link of the 2nd engine, …, 2nd link of the 1st engine, …), and each time the creation operator is called then it outputs the next link on this list. When none of these engines can provide further links, then the creation operator is not used anymore and is replaced by the mutation operator. This creation operator allows the genetic search to start with points of good quality. As will be seen in the results section, those heuristically generated individuals can be greatly improved with the mutation operator.

From a selected parent page P, the mutation operator generates an offspring O by exploring the local neighborhood of P. For this purpose, the links found in P are ordered in a list in decreasing order according to the values of the link evaluation function Linkeval . sIn this way, the most promising links are explored first. Each time the mutation operator is called, the next link on the list is given as an output. When the list is empty, then the creation operator is used. In order to speedup the pages evaluation and to avoid downloading twice the same page, we maintain a "black list"

of pages which have already been explored. If one of the two previous operators outputs a page of this list, then this operator is ran again. One should notice that the graph structure of the Web does not allow us to define a crossover operator in a straightforward way. It could be possible to define such an operator by combining links present in two parent pages P1 and P2: if those two pages have a link in common, or links pointing to the same web site, then it might desirable to combine these information and to focus the search on this common links or web sites.

CONCLUSION

Although only 100 out of a possible 16,777,216 quadrats were scanned, the quadrat counting method of estimating the size of the World Wide Web appears reasonable. An estimate of 18.5 million web servers seems neither too high, nor too low and is probably as valid, if not better, an estimate than any previously devised method could provide. By repeating this study after given time intervals, we would be able to make further estimates on the growth of the Web relative to time, and by increasing the level of information gathered by our software, we would be able to speculate on the distribution of server types across the Web.Testing to determine how well the Poisson distribution fit our data, and to establish confidence limits (to determine how many quadrats will provide a good estimate) would also increase the validity of the study.

If future research were to be conducted in this area it may be beneficial to change the nature of the quadrats yet again. Originally, our problem was that the quadrats we had chosen were of the form *a*.*b*.*c*.(0 - 255). This allowed a high level of homogeneity within the quadrats, and it was possible to stumble upon a server farm subnet as a quadrat (giving us an extremely abundant quadrat), while another quadrat could be a subnet composed entirely of unused addresses (giving us an extremely sparse quadrat). Additionally, quadrats of this "shape" meant that all of our probes would be received by the same subnet and could be interpreted as an attack by a zealous System Administrator.

We have therefore identified the following areas for further research in this field:

- empirical evaluation of methodologies and modelling languages in a variety of web application contexts;
- development of a flexibly and unified modelling language for web development based on the reviewed models suggested above;
- development of CASE tools that would support such modelling languages;
- the production of academic text books on hypermedia/web methodologies and modelling languages.

REFRENCES

[1] Atzeni, P. et al, 1998. Design and Maintenance of Data-Intensive Web Sites, *Proc. EDBT 1998*, pp. 436-450.

[2] Bennett, S. et al,1999. *Object-Oriented System Analysis and Design*. McGraw-Hill Publishing Company, London.

[3] Benyon, D, 1990. *Information and Data Modelling*. Blackwell Scientific Publications, Oxford.

[4] Bichler, M. and Nusser, S, 1996. Developing structured WWW-sites with W3DT. *In: Proceedings of the WebNet – World Conference of The Web Society.* October 16-19, San Francisco, CA USA.

[5] Monmarché N., Nocent G., Slimane M. and Venturini G. (1999), Imagine: a tool for generating HTML style sheets with an interactive genetic algorithm based on genes frequencies. 1999 IEEE International Conference on Systems, Man, and Cybernetics (SMC'99), Interactive Evolutionary Computation session, October 12-15, 1999, Tokyo, Japan.

[6] Morgan J.J. and Kilgour A.C. (1996), Personalising information retrieval using evolutionary modelling, Proceedings of PolyModel 16: Applications of Artificial Intelligence, ed by A.O. Moscardini a nd P. Smith, 142-149, 1996.Moukas A. (1997), Amalthea: information discovery and filtering using a multiagent evolving ecosystem, AppliedArtificial Intelligence, 11(5):437-457, 1997