Scandinavian Journal of Information Systems

Volume 11 | Issue 1 Article 2

1999

Surveying the World Wide Web

Henrik Fagrella
The Viktoria Research Institute, fagrell@informatik.gu.se

Carsten Sorensen
London School of Economics, c.sorensen@lse.ac.uk

Surveying the World Wide Web

Henrik Fagrell[a] & Carsten Sørensen[b1]
fagrell@informatik.gu.se[a] and c.sorensen@lse.ac.uk[b]
[a]Department for Informatics, Göteborg University, and
The Viktoria Research Institute, Göteborg, Sweden
[b]Department for Information Systems
The London School of Economics and Political Science, United Kingdom, and
Laboratorium for Interaction Technology, Trollhättan Uddevalla, University, Sweden

Abstract

The World Wide Web (the Web) is the main driving force behind the rapid diffusion of Internet technology. As a result, we are beginning to live a significant part of our lives in Cyberspace. Measuring and monitoring our surroundings is an essential human activity that helps us both to understand and shape the world we live in. Substantial efforts have in the past years been invested into further understanding the Internet in general and the Web in particular through, for example, surveys of user attitude and behaviour, maps of Internet traffic, and indexing of content. Very little research has, however, investigated how to measure and monitor the contents of Web sites based on a combination of linguistics and data visualisation measures. Many efforts have demonstrated the use of techniques from within a particular discipline such as information retrieval, data mining, or autonomous agents. This paper, however, explores issues related to the monitoring of contents and changes to the Web based on a range of measures. The paper aims to demonstrate the principles behind the application of semi-automatic measurement instruments to forward our understanding of the Web as a body of textual traces of human activity. The paper suggests five basic types of measures for studying the Web: volume, density, vocabulary, structure, and relative measures. A survey of 82 Swedish Web sites was conducted using semi-autonomous Web robots for information retrieval and filtering based on techniques from linguistics and information visualisation. Examples demonstrate how such data can be applied to summarise site contents, identify site topic, map site structure, and compare Web sites. The results are discussed and related to emergent issues, such as Web navigation, electronic commerce and the management of knowledge.

Keywords: Data visualisation, linguistics, textual trace, indexing of content, maps of Internet traffic.

1. Introduction

The Internet has experienced an explosive rate of growth since it's inception in the late sixties, initially designed as an experimental network for scientists and US Government contractors. During the 1980 it gained widespread popularity among faculty, staff and students at universities and research centres as an information infrastructure. Restrictions for commercial use of the Internet were lifted in 1991, and that year the World Wide Web (the Web) (Berners-Lee et al., 1994) was introduced to the public. These two events proved to be a very strong driving force for diffusion of the technology to most aspects of public and private domain around the world at a previously unseen rate (Leiner et al., 1997; Guice, 1998; Hannemyr, 1998; Zakon, 1998). As a result, many spend an increasing proportion of both

1.Authors are listed in alphabetical order only. This paper was submitted for review October, 1997, and accepted for publication October 1999. It represents a survey conducted in 1996.

their professional and private lives in this new 'world' of bits (Mitchell, 1995). We search for information, purchase products and services, and interact with people using this particular information infrastructure. We also experience new problems related to living in and with this new world, such as finding our way, finding specific information, remembering where we have been, and in general getting an overview (Li, 1998; Sørensen, 1998; Nielsen, 1999). In the physical world we have been monitoring an abundance of both cultural and natural attributes over hundreds of years. Monitoring and measuring are fundamental activities for understanding the world we inhabit and shape and substantial efforts have been put into understanding and describing both the Internet and the Web (Dodge, 1999). The behaviour and attitudes of people using both the Internet and the Web have been surveyed. Others have studied how to represent, map, visualise and analyse the Internet and the Web using quantitative measures. Most approaches that have been suggested for characterising the contents of the Web, however, tend to focus on the application of a particular technique, and a large proportion of the efforts are directly related to support search engines indexing the Web or to visualising the hyperlink structure of the Web. There has been relatively little research that has explored the contents of the Web as a body of text undergoing constant change, in an attempt to provide an overview and to compare the contents and structure of Web sites.

This paper asks the question; How can we apply semi-automatic measurement instruments to better understand and present the contents of and changes to sites on the World Wide Web? This article presents results based on a comprehensive survey where 82 Swedish Web sites within six sectors were retrieved a number of times throug seven weeks in 1996. The process was initiated by the specification of the Web measures to be calculated. A series of Web robots were developed and tested. After selecting the most appropriate type of robot for the task at hand, the 82 Web sites were selected and the robots were employed as a means of semi-automatic Web retrieval (Leonard, 1997; Schneiderman and Maes, 1997; Schubert et al., 1998; Wooldridge and Jennings, 1999). Techniques from computational linguistics and information visualisation were used in order to filter and represent the Web site contents (Tufte, 1983; Tesitelová, 1992; Young, 1996). Applying techniques from computational linguistics enabled us to provide overviews of not only superficial syntactical changes to the Web sites, but also to gain some insight in more substantial changes to the type of contents of the Web sites as well as the degree of sophistication of the language employed. The task was basically one of collecting, detecting and displaying information based on a stable need, from a dynamic and unstructured source (Oard, 1997). The process was semi-automatic, in that the robots were manually configured with a list of sites to be visited. The data collection was automatic, and data aggregation was partly automatic and partly manual. It was based on measures selected from each of the five basic categories; volume, density, vocabulary, structure and relative measures, we demonstrate how to summarise site contents, identify site topic, map site structure, and compare Web sites. The examples are all drawn from the newspaper sector. Given the focus on measuring and understanding changes, this strata appeared to be the most interesting.

Obviously, a concern could be raised regarding the age of the data presented, given the survey was conducted throughout 1996. Given the rapid changes to the Web, the actual data collected may now be atypical. Given the focus on newspaper Web sites, Eriksen et al. (2000) illustrates very well in their analysis of how a Swedish, a Norwegian and a Danish newspaper Web site all have changed from 1996 to 1999. However, two issues need to be considered here. Firstly, the delay in publishing the results are caused by factors outside the authors' control. Secondly, the important aspect of this paper is not to provide an actual up-to-date overview of the Swedish Web. The purpose is to demonstrate how the approach selected can add to the way we understand and present the Web. Given our choice of perspective where we look at the Web as individual sites containing hyper-linked text with embedded graphics, relatively little has changed. Although there is an increase in more sophisticated Web sites where a significant proportion is generated dynamically, the end product is still a Web page. Although there is a drive towards digital conversion of various

media, involving streamed video and audio as an alternative to textual contents, the emergence of mobile computing with small networked information devices and standards for that platform has actually reinforced the importance and 'centre-stage' of textual information.

In the following section we survey related research and outline the problem setting. Section 3 suggests a set of five basic measures for surveying the Web. Section 4 presents the instruments and procedures for collecting and analysing data in the survey. Section 5 presents examples of survey results applying Web measures from each of the five basic types suggested in Section 3. Section 6 discusses the approach and relates it to emergent issues within Web development and use such as Web navigation, electronic commerce, and the management of knowledge. It is argued that there is a need for mechanisms providing an overview of both the structure and the contents of complex Web sites. We conclude that this approach can inform future design of a number of specialised Web services such as navigational support, search engines and advertisement.

2. Related Research

We are increasingly relying on information and communication technology when we interact with others or document our activities. This implies that the traces we leave behind increasingly are electronic as opposed to being paper based. Little research has, however, investigated how we can survey and analyse changes to the Web, when viewed as textual traces of human activity. Other types of traces have been studied previously. In 1992, for example, Hill et al. (1992) discussed electronic traces as implications for design of user interfaces. Whittaker et al. (1998) have investigated USENET news messages for aspects of people's conversational strategies, demographic belonging and interaction frequency. Greenhalgh (1997) has investigated the traces of movement of avatars in a virtual collaborative environment to inform requirements for look ahead, bandwidth and special treatment of some participants.

A number of approaches for surveying the Web consisting of a dense weave of texts, pictures, interactive components, CGI scripts, have been suggested and applied since its conception in the early 1990. This section relates the approach applied in the survey of the Swedish Web to other research efforts. It is beyond the scope of this paper to compile and discuss all related research. Dodge (1999) however, presents the most comprehensive list of references we have found, and his index on has proved a valuable resource (see `www.cybergeography.org`).

In this paper we investigate the lessons that can be learnt about changes to the Web from semi-automatic data collection and analysis. The unit of analysis is the individual Web site defined by the registered domain name, and viewed as hyperlinked texts. The project surveyed the publicly accessible part of the Swedish Web. Within that, we chose 82 Web sites within six sectors. A Web robot surveyed the hypertext contents of the chosen Web sites a number of times during a six week period. During each measurement the robot collected the entire public Web site by traversing the hyperlink structure for the site. The data was downloaded and subjected to further analysis. The survey method can be characterised as the application of semi-autonomous robots (Leonard, 1997; Schneiderman and Maes, 1997; Schubert et al., 1998; Wooldridge and Jennings, 1999) to the World Wide Web combined with information filtering (Oard, 1997) based on techniques from linguistics (Tesitelová, 1992) and the visualisation of information (Tufte, 1983; Young, 1996). Because of the magnitude of the task of surveying the Web, it was important to stratify the sampling, i.e. select a target population. The survey, therefore, focused on collecting and analysing data from Swedish Web sites within relatively few sectors, e.g., newspapers, companies registered on the stock exchange and government agencies. This is not significantly different from surveying the physical world. Geographers, sociologists, economists and statisticians are also forced to stratify their areas of inquiry. The annual Swedish statistics report (SCB,

1996) for example, only contains a tiny fraction of attributes measured, which in turn only represents an infinite fraction of the attributes measurable. This paper only attempts to investigate the publicly accessible part of the Web. It was not the intent to study aspects of Web sites which were not publicly available, such as, server activity logs and restricted access-areas. In the following we present and discuss research related to our apporach. Section 3 presents the measures applied in the survey and Section 4 presents and discusses the survey method in further detail.

A major area of research with regard to both the Internet and the Web concerns relationships between users and the technology. A number of surveys have investigated demographics, behaviour and attitude of Web users. Pitkow and Kehoe (1997) present an ongoing series of comprehensive demographic surveys of Web-use patterns, conducted by researchers at Georgia Tech. (`www.gvu.gatech.edu/user_surveys/`). Hoffman et al. (1996) also present a study of the use of the Web. Electronic Commerce on the Web is continuously surveyed by the CommerceNet/Nielsen Internet Demographic Survey (CommerceNet, 1996) from questionnaire data. Some researchers have applied methods measuring user behaviour when accessing the Web. Garofalakis et al (1999) show how to optimise the structure of Web sites based on surveys of user behaviour. Tauscher and Greenberg (1997) analyse how people revisit web pages with the intention of informing the design of better history mechanisms for browsers. This approach relates to research within user modelling where models of user preferences and behaviour can inform the design of new functionality (Allen, 1990; Ambrosini et al., 1997; Maglio and Barrett, 1997). The approach adopted in this paper does not investigate the relationships between users and the Web. Instead it analyses the Web as such, and attempts to derive lessons from looking at the hypertexts as language and structure.

Viewing the Web as a 'world' of bits naturally raises the issue of space. In geometry, space is defined by two concepts: topology and metric. If we use the geometrical definition of space as a metaphor, the Web's topology can loosely be describes as a graph with nodes and uni-directional links. The nodes represent retrievable documents, i.e., files containing texts, images, links, and several other types of information. However, a plain distance metric does not capture the phenomenon accurately. Increased physical distance between the computers connected in the network does not necessarily lead to higher transaction costs. The metrical aspects can, however, be based on other variables than distance. Other researchers using the geometric metaphor considers the Web's metric to be calculations on how to traverse the graph formed by the link structure (Drew and Hendley, 1995; Mukherjea and Foley, 1995; Girardin, 1996). Chakrabarti et al. (1999) apply link clustering algorithms to determine authorities on topics based on the link topology.

Researchers have studied the Web in order to describe emergent properties. For example, Palmer and Eriksen (1999) and Eriksen et al. (2000) study newspapers on the Web as new forms of news products. Smithson (1999) has produced a ranking of 100 web sites representing commercial organisations from 8 sectors with respect to the support for electronic commerce. Characteristic for this type of investigation of the Web is a qualitative approach based on manual navigation of the Web sites. This paper, however, explicitly subscribes to a quantitative approach based on semi-automatic techniques. Given the aim of measuring and monitoring changes to a relatively large number of web sites, the software agent approach was deemed more suitable for our purpose compared to the direct manipulation approach offered by the conventional Web browser (Schneiderman and Maes, 1997). As argued by Nielsen (1999), the dramatic increase in the size and complexity of the World Wide Web will lead to new challenges for user access methods, and the application of semi-automatic filtering and retrieval techniques thus seems to be a viable approach.

A number of research efforts are concerned with the visualisation and mapping of both the Internet and the World Wide Web in order to provide support for navigation (Dodge, 1999). These address issues such as: maps of the Internet structure, Internet repositories and indices, statistics of Internet traffic and size, and visualisation of Web spaces. For example, Girardin (1996) and Drew and Hendley (1995) present visualisation of hyper-link

structures. Barry and Batty (1994) analyse the diffusion of the Internet in order to predict future growth. Dodge (1996) applies a spatial metaphor to analyse the Web using Geographical Information System (GIS) technology. Young (1996) presents an extensive survey of 3D information visualisation research. The approach adopted in this paper uses information visualisation techniques to represent the results of the semi-automatic survey. Visualisation is a means rather than a goal in itself however.

Guan and Won (1999) promote keyword based datamining of the Web using predefined patterns. This approach is suitable for identifying and extracting pre-defined patterns, and not appropriate for characterising the contents of a Web site without prior assumptions about the nature of the contents. Similarly, Kumar et al. (1999) applies semi-automatic techniques for identifying communities of common interest based on Web contents. Stenmark (1999) applies a similar approach combined with mutual recommendation functionality (Oard, 1997; Resnick and Varian, 1997) within a company intranet.

Bray suggests collecting data and performing statistical analyses on volume and density measures of the Web (Bray, 1996). The project, furthermore, looks at the relative link topology between Web sites. Bray applies software robots for data collection. This approach has a number of similarities to the approach we have applied. There are, however, also major differences. Bray's survey of the Web is based on the Open Text Index, November 1995, covering 1,5 million pages. The parameters analysed are, however, few and they are mainly volume and density measures, e.g., distribution of page sizes, number of embedded images, and types of file extensions. These are combined with structural measures such as a ranking of sites most often referred to, and other inter-site linking measures. The interlinking measures are applied to illustrate proximity of sites through a spatial mapping.

Bray's approach and the one adopted in this paper both apply the Web site and page as the two basic sample units. It could be argued that defining the granularity of the survey based on site names is biased. Surveying sites, results in focusing on institutionalised entities on the Web. One way of taking this into consideration is to calculate "links-to-site" sets. Bray, for example, calculates the rankings of most popular site referenced to in the pages. While Bray focuses on few and relatively simple measures for a large sample, we have chosen to measure more parameters, and to apply a deeper analysis of the contents of the pages.

3. Web Measures

The Web can be viewed as a body of text containing two fundamentally different types of data: the contents and the tags. A tag is in HTML (Hyper-Text Markup Language) metadata describing the layout and linking structure between the text, graphics, audio and interactive components. In more advanced markup languages such as XML or SGML, the markup types can be defined by the user (Khare and Rifkin, 1997; Lassila, 1998; Rein, 1998). Analyses of HTML hypertexts, therefore, concern both aforementioned types of data. This paper suggests the application of the five basic different types of measures presented below. These are drawn from the research reviewed above and from computational linguistics (Tesitelová, 1992). Table 1 summarises the measures and provides examples of actual measures within each type. Most of the measures are based on compilations of frequency lists of words retrieved from the Web sites. A frequency list for a Web site can contain the tokens found, which are all the separated words, or it can be filtered further to only contain the types found, i.e., a list of unique tokens.

Volume measures count total numbers of constituents in the hypertext, such as, bytes, pages, link errors, tokens, types, headings, interactivity, internal and external links. The number of bytes and pages provide measures of the size of a site. The number of link errors reflects how well it is maintained. The total number of tokens and types provide contents-based volume metrics for a site. Interactivity is measured by counting forms, CGI-script and Java-applets. Measuring headings, external links (to other sites) and internal links (within the site) provide quantitative measures for 'page-layouts'.

Density measures relate volume measures to each other, making it possible to express more general site properties. Examples of density measures are: Bytes pr. page, average number of tokens pr. link error, and number of external links per page.

Vocabulary measures analyse site text vocabulary applying the computational linguistic measures: Stemming, Guiraud and theoretical vocabulary (Tesitelová, 1992). Stemming is a technique, which classify common words according to common meaning, e.g., reading, read, reads. Guiraud is a measure reflecting vocabulary richness. It is calculated by dividing the number of types by the square rooted number of tokens. This measure does not incorporate the size of the text, and subsequently fails on both extremely small and large texts. Because of the large variations in the size of Web sites we have used theoretical vocabulary as a complement to Guiraud. Theoretical vocabulary is not sensitive to the text size, but because it is computed based on a frequency list of types, it is computationally more complex than Guiraud. Theoretical vocabulary reflects the expected number of types if the tokens are reduced. The measure is calculated as follows: Assuming that a text containing *N* number of tokens is reduced to *M* number of tokens. Let *V* be the number of word types. The possibility that all occurrences of a particular word type is eliminated in a reduction is $(M/N)^i$. If TN is the original number of types, the theoretical vocabulary will be (*TM*) (see Figure 1). We reduced the number of types (*M*) to ten thousand. Both Guiraud and theoretical vocabulary values increases when the vocabulary gets richer.

Figure 1: The theoretical vocabulary formula.

$$T_M = T_N - \sum_V V_i \left(\frac{M}{N} \right)^i$$

Structural measures provide quantitative measures representing the spatial property distance. Two structural measures are applied in this paper, directory structure for the Web site and mean distance of the hypertext. The former is relatively straightforward, and the latter reflects whether the site link-structure is deep or flat by giving the average number of the smallest amount of clicks on links needed to get from the start page to any other page.

Table 1: The five types of measures from which the actual ones applied in the paper is drawn. For each type, one or more examples are given.

Measure	**Description**	Examples
Volume	Count absolute numbers of hypertext atoms (e.g. the tags and the text). This constitutes all raw data collected from which the remaining measures are calculated.	Number of separated words (tokens) or different words (types) within a site
Density	Density calculations based on the volume measures.	Number of errors pr. page. The standard deviation for tokens pr. page.
Vocabulary	Identifies the richness of the used vocabulary.	Guiraud or theoretical vocabulary
Structure	Attempts to measure the site hierarchy, depth and width of the link tree.	The average number of clicks on internal links needed to get from the start page to any other page.
Relative	Compare different data sets.	Lexical equality measure identifies whether two texts deal with the same topic, or have a similar content.

Relative measures compile various differences between sites. We have used lexical equality of frequency lists as a relative measure to detect if two sites used the same type of language. This approach does not take into consideration where in the text a particular word

appeared. Lexical equality can be calculated from frequency lists of tokens or of types. The types-based method does, compared to the token-based approach, not consider the frequency of a particular word. The token-based method can potentially result in a bias towards non-context carrying words. Context-carrying words often have low frequencies. They are often nouns or verbs and explains more about the texts than highly frequent words such as *and*, *or* and *I* can do. Lexical equality is expressed as a percentage of equality between two frequency lists. Calculating the lexical equality of every combination of a group of sites results in a a matrix representing the lexical distance map. The values in the matrix can be visualised using a clustering algorithm (Jain and Dubes, 1988). This yields a two-dimensionally representation of the relative distance between the frequency lists according to the clustering metric. Hagman and Ljungberg (1995) have demonstrated how clustering of lexical equality can be applied to compare conventional newspaper articles. Clustering provides a semi-automatic means for identifying patterns across Web sites. Olsen et al. (1993) have used clustering of keywords characterising individual documents to obtain an overview of a hypertext document collection.

4. Survey Setting

In the survey we applied semi-autonomous Web robots for retrieval of Web site contents (Leonard, 1997; Schneiderman and Maes, 1997; Schubert et al., 1998; Wooldridge and Jennings, 1999). The process of bringing about the Web robot for the survey can be characterised in terms of the following nine activities: (1) specification of Web measures to be calculated; (2) design and construction of software robots; (3) small-scale tests of robots; (4) selection of robot for survey; (5) Web site selection; (6) data collection; (7) data aggregation; (8) data analysis; and (9) documentation of results.

A (Web) robot is a program that automatically traverses the Web's hypertext structure by retrieving a document, and recursively retrieving all documents referenced. The term 'recursively' does not limit the definition to any specific traversal algorithm. The robot can apply various heuristic algorithms to the selection and ordering of documents to be retrieved. A Web browser is normally not in itself considered a Web robot since it is operated by a human user and does not automatically retrieve referenced documents (Sørensen, 1998). If robots do not contain rules stipulating when to terminate the retrieval of documents, they might attempt to retrieve all the public pages on the Web. The termination criterion can be defined relative to a certain link structure depth, or be based on a predefined number of retrieved documents. The criteria applied in our experiments are defined relative to the public pages within a given site or domain. Web robots or Web agents are also frequently referred to as 'Web Wanderers', 'Web Crawlers', or 'Web Spiders'. These names are however, misleading as they give the impression that the software itself moves between Web servers. Since the research reported here did not use the relatively rare mobile agent technology (Huhns and Singh, 1997; Schubert et al., 1998) the Web robot simply visited sites by requesting documents from them. This technique is similar to the one used by seach engines such as Altavista (`www.altavista.com`) and Hotbot (`www.hotbot.com`) for indexing the Web.

The survey was conducted on 82 Swedish Web sites. All of the sites surveyed were found in the Swedish University Network (SUNET) link collection. Although all sites were Swedish with server address within the ".se"-domain, this was no guarantee for the Web server physically being located in Sweden. For example, the server `www.ericsson.se` was physically located in the Netherlands. We did not have resources to investigate the entire Swedish Web, and we intended to use our contextual knowledge about the selected

sub-strata during the analysis. In order to survey a cross section of Web sites representing both public and private organisations, we chose sites from the six sectors listed in Table 2:

Table 2: Frequencies and percentages of the 82 sites analysed.

Sector	Count	%
A-List, i.e., companies on Swedish stock-exchange	24	29,3
Municipalities	11	13,4
Newspapers	8	9,76
Political parties and interest groups	13	5,9
Government agencies	19	23,2
TV- and radio stations	7	8,54

Three different robot prototypes were constructed. The first one was an extension of the maintenance robot MOMSpider (Fielding, 1994), implemented in perl and used for validating links and generating statistics. Due to performance problems with perl, a second robot was developed in C++. During the development of the second robot we came across ht://Dig (available at URL http://htdig.sdsu.edu/) implemented in C++. It is constructed to index local networks, such as Intranets, but with some adjustments it served our purpose perfectly. All the data documented in this paper is collected by this robot. Using an existing agent architecture is highly beneficial since developing a reliable architecture is a substantial effort (Wooldridge and Jennings, 1999). The robot conformed to the de facto ethical standard for Web robots (Eichmann, 1994; Koster, 1997). This entails that the robot did not squire resources from human users by retrieving pages at high speed. It also ensured that the robot identified itself to the Web server allowing the Webmaster to contact the owner of the robot if problems should occur.

Table 3 shows key sampling data on the total amount of hypertext retrieved. It also provides information on size, download time, number of tokens and types, and the calculation time for the frequency lists for both the largest and the smallest Web site.

Table 3: Key sampling data on the total amount of hypertext sampled: the largest site and frequency list, as well as the smallest site and frequency list.

All sites	310 mega-bytes uncompressed hypertext. 21 mega-byte URL-lists
Largest site and frequency list	Ericsson, Oct 14. 17,500 kilobytes hypertext downloaded in 10 hours. 2000 kilobytes frequency list. 2,489,999 tokens and 163,636 types calculated in 3500 seconds (with an optimised C-program)
Smallest site and frequency list	Dagens Industri, Oct 14. 178 kilobytes hypertext downloaded in 12 minutes. 17 kilobytes frequency list. 11,014 tokens and 1,948 types calculated in 7 seconds (with an optimised C-program)

In order to reach a sufficient depth of analysis, we chose to focus on one of the sectors surveyed, the newspaper Web sites. In general, they changed more frequently compared to sites in other sectors. As an example, there were no changes during the sample period to any of the A-list companies' Web sites. The newspapers sites were collected on five different occasions in 1996, namely, September 23rd, September 30th, October 14th, October 29th and November 4th,. The newspapers are listed in Table 4. The data aggregation was conducted with a variety of small programs implemented in several different languages, such as, C, perl and awk. Part of the data aggregation process was automated by perl scripts 'glueing' the various programs together. Standard statistical packages (DataDesk and

Microsoft Excel) were used for calculations and hypotheses testing. We have also used data clustering (Jain and Dubes, 1988) to visualise relative results in order to establish patterns in the data material.

Table 4: The Swedish newspaper Web sites analysed, with a indication of the month when the Web service had been launched.

Newspapers	Description	Launch Date
Aftonbladet	National evening paper	August 1994
Arbetet Nyheterna	Regional morning paper	March 1996
Dagens Industri	National business daily	June 1995
Göteborgs Posten	Regional morning paper	August 1995
Hallandsposten	Regional morning paper	September 1995
Nerikes Allehanda	Regional morning paper	May 1995
Sydsvenska Dagbladet	Regional morning paper	August 1995
Svenska Dagbladet	National morning paper	June 1995

5. Results

The extent of the survey was such that only parts of the data analysis can be presented. We have chosen to focus on the following five aspects: (1) Summarising the contents of a site; (2) Identifying topics presented at a site; (3) Mapping the changes to a site structure; (4) Comparing the contents of several sites; and (5) Surveying a messy world. To illustrate aspects (1), (2), and (3), we show the results from analysing the two newspaper sites Göteborgs Posten and Sydsvenska Dagbladet. To illustrate the type of analysis performed in (1) and (4), results from comparing the eight newspaper Web sites are shown.

5.1Summarising Site Contents

How can the results of a survey of a site be summarised? We have chosen to present key-data in the form of tables. Table 5 shows the types of information presented in Table 6 for Göteborgs Posten and Table 7 for Sydsvenska Dagbladet. Göteborgs Posten (GP) is geographically located in Göteborg (Gothenburg) and operates in the western part of Sweden. GP was the second largest morning newspaper in Scandinavia with an average circulation of 273,000 on weekdays and 306,000 on Sundays. These figures has not changed substantially since 1996. Sydsvenska Dagbladet (SD) is also a regional morning paper. Both GP and SD initiated a Web site in August 1995. Since there are only five observations from the sites, we can only perform a tentative qualitative analysis of the data. Furthermore, some of the variables did not change much during the sample-period. Those variables are bytes, pages, types, links, bytes pr. page, types pr. page, links pr. page and theoretical vocabulary.

Table 5: The types of key data from samples of Swedish Web site.

Measures	Explanation
Bytes	Number of bytes at the site
Pages	Number of pages at the site
Tokens	Number to tokens, i.e., separated words
Types	Number of types, i.e., number of unique words
Error	Number of link errors
ErrorLinks	Number of links
Headings	Number of html document headings

Table 5: The types of key data from samples of Swedish Web site.

Measures	Explanation
Gif imgs	Number of gif images on the site
Jpeg imgs	Number of jpeg images at the site
Bytes/pg	Average number of bytes pr. page
Tokens/pg	Average number of tokens pr. page
Types/pg	Average number of types pr. page
Lnk Err/pg	Average number of link errors pr. page
Links/pg	Average number of links pr. page
Headin/pg	Average number of headings pr. page
Largest pg	The largest page encountered
Guiraud	Guiraud calculated for the site
Theo.Voc.	Theoretical Vocabular calculated for site
Mean Dist	The average distance from the root to any other page

We can appreciate the rate and extent of change of Sydsvenska Dagbladet's site when comparing the standard deviation between the size of the samples which turns out to be more than the total size of many of the other newspaper sites. Hallandsposten only had a maximum of 601k bytes, Dagens Industri had 176k bytes, and Arbetet 621k bytes.

The theoretical vocabulary gives a quantitative measure of the diversity of a text. In general we would expect a more diverse text to represent a more complex and diverse choice of language. The average for all sites is 1819 and the top score is 2617 (The Royal Library). This makes the 2437 average for Göteborgs Posten, and 2405 for Sydsvenska Dagbladet quite high. The theoretical vocabulary increased slightly over time on the SD Web site and minor fluctuations could be detected on the GP site. Apart from changes to the vocabulary, no radical changes could be measured on the Sydsvenska Dagbladet's site. This is, of course not a reflection of lack of update of the information on the site. It merely informs us that the contents may have changed, but the type of contents have not. Furthermore, the relatively small changes may not be a significant predictor for substantial changes to the type of texts on the site.

The sample clearly showes that something happend to the GP site between sample II and IV (See table 6). Firstly, the number of link errors were5,4, and 3 in the previous samples, and suddenly increased to 85 and 91 in sample IV and V. The mean distance also changed from 2.3 to 6.6. This indicated a major restructuring of the site transforming it from having a flat links-structure to a more deeply one. The site also had obtained nine interactive forms in sample IV from having no forms in sample III. There had also been a complete change to the sites set of outgoing links. The two most popular external links in sample IV were `www.realaudio.com` and `www.netscape.com`, which occurred twenty times each. These links were `www.westnet.com` and `www.sunet.se` and they were used about 25 times each. This clearly indicated an overall change of the site´s page and linkstructure layout.

Table 6: Key data from the five samples of the Göteborgs Posten Web site.

Measure	I	II	III	IV	V	Average	Std.dev
Bytes	5113k	5152k	5209k	5629k	5980k	5417k	376665
Pages	912	914	929	902	943	920	16
Tokens	502k	509k	512k	569k	606k	641k	45874
Types	58210	58661	59171	59102	61812	59917	1407
Errors	5	4	3	85	91	38	46

Table 6: Key data from the five samples of the Göteborgs Posten Web site.

Measure	I	II	III	IV	V	Average	Std.dev
Links	9943	10083	10229	9719	10351	10065	247
Headings	1574	1541	1543	1494	1605	1551	41
GIF imga	144	117	136	177	179	150	27
JPEG imga	11	14	13	29	31	20	10
Bytes/pg	5606	5637	5608	6241	6342	5887	371.16
Tokens/pg	1108	557	552	631	643	698	282.83
Types/pg	66.71	64.18	63.69	65.52	65.55	65.13	1.20
Lnk Err/pg	0.0055	0.0044	0.0032	0.0942	0.0965	0.0408	0.0498
Links/pg	10.90	11.03	11.01	10.77	10.98	10.94	0.011
Headin/pg	1.73	1.69	1.66	1.66	1.70	1.69	0.03
Largest pg	35kb	43kb	40kb	44kb	51kb	43kb	5918
Guiraud	58.0	58.1	58.5	55.4	56.1	54.2	1.4
Theo.Voc.	2455	2454	2458	2409	2410	2437	25
Mean Dist	2.3	2.3	2.3	6.6	6.7	4.0	2.4

Although much weaker, SD also showed a change between sample III and IV, with increases in both the size of the site and in number of pages. Here the mean distance, however, remained virtually unchanged. In both GP and SD the tokens, types, Guiraud, and theoretical vocabulary showed that the types of texts did not change substantially. As an example, SD had an increase in tokens of around 9.5% over the period. This might not seem much, but the sample period was only 7 weeks, which roughly translates to 70% increase per year. It is not unrealistic to assume a steady growth, since the Web site was started in August 1995. Göteborgs Postens Web site has experienced a growth from an estimated access of 3000 people/day in 1996 to around 27,000 people/day in 1999 and has expanded dramatically in both size and type of material published (Eriksen et al., 2000).

Overview information as presented in this section, can, amongst others, provide indications of major redesign of a Web site, and perhaps also help pointing at what types of changes have been implemented. The overview information can also support cross-site analyses of relative technological advancement, in terms of, for example, comparing the use of interactive components, automatically generated pages, and linking structure. If users accessing Web sites are interested in notification of major changes, and if a Web site is indexed regularly, triggers could be configured, notifying about changes exceeding the defined threshold (Senanayake, 1998). In a world of constant change, it could be more valuable to only be informed about significant changes and not only of any change occurring

Table 7: All key data from the five samples of the Sydsvenska Dagbladet Web site.

Measure	I	II	III	IV	V	Average	Std.dev
Bytes	12887k	13130k	13854k	14316k	14437k	13725k	694378
Pages	1327	1353	1438	1520	1543	1436	96.5
Tokens	1550k	1574k	1642k	1700k	1713k	1636k	73066
Types	124015	125301	128940	131420	132127	128361	3710
Errors	21	18	17	22	21	20	2.17
Links	11052	11271	13095	13852	14042	12662	1417
Headings	1855	1876	1820	1930	1950	1886	53.5
GIF imga	321	305	288	300	307	304	12
JPEG imga	162	159	178	191	198	178	17

Table 7: All key data from the five samples of the Sydsvenska Dagbladet Web site.

Measure	I	II	III	IV	V	Average	Std.dev
Bytes/pg	9711	9704	9634	9418	9356	9565	166
Tokens/pg	1168	1163	1142	1118	1110	1140	25.89
Types/pg	93.46	92.61	89.67	86.46	85.63	89.56	3.52
Lnk Err/pg	0.0158	0.0133	0.0118	0.0145	0.0136	0.0138	0.0015
Links/pg	8.33	8.33	9.11	9.11	9.10	8.80	0.43
Headin/pg	1.40	1.39	1.27	1.27	1.26	1.32	0.07
Largest pg	426kb	426kb	426kb	426kb	426kb	426kb	0
Guiraud	70.4	70.6	71.1	71.3	71.4	71.0	0.4
Theo.Voc.	2383	2388	2409	2421	2422	2405	18
Mean Dist	4.5	4.4	4.5	4.6	4.6	4.5	0.08

5.2Identifying Site Topics

Imagine that arriving at a Web site, you could be presented with a pre-prepared list of topics covered by the Web site. As a first step in conducting a more in-depth analysis of the contents of Web sites, we chose to apply a semi-automatic technique for identifying Web site topics. The technique is based on the frequency list of types, i.e. different tokens, for the site. The frequency list is filtered using a stop-list of non-context bearing words such as; *are, is,* and *it*. Subsequently a word-stemming algorithm is applied in order to find syntactically related words and dividem into word categories. This, for example, results in the words *Göteborgs, Göteborg, Göteborgare, Göteborgarna, and Göteborgarnas,* from Göteborgs Posten's Web site being classified as one category (See Table 8). The last step in the process is to sort the list of word categories according to occurrence. This technique is also applied manually at newspapers, amongst others the British newspaper The Guardian, to highlight the core contents of political speeches. Here, a list of most common word categories in political speeches of political opponents are compared, to bring forward a deeper understanding of differences in opinion between the political parties. Table 8 presents the most frequent word categories from Göteborgs Posten and Sydsvenska Dagbladet based on the sample September 23rd. Although both Web sites are Swedish regional newspapers, the contents of the Web sites differed significantly. The GP Web site clearly still had a substantial Web coverage of the 1996 Atlanta Olympic Games: 3: Os, 7: USA, 18: Final, and 19: Atlanta. It also devoted substantial amount of text to the topic of films (1227 occurrences of words in this category). Sydsvenska Dagbladet, on the other hand, did not have any of these two topics represented on the list. SD's Web site was mainly devoted to information for a Swedish-Danish investor club, helping people locating appropriate shares. This can, amongst others, be seen from the large proportion of Swedish and Danish company names on the list: 2: Plm, 5: Trelleborg, 7: Skanska, 8: Danisco, 9: Astra, 10: Ericsson, 13: Pharmacia, 14: Carlsberg, 16: Oticon, and 19: Althin.

Words associated with the publishing institution itself or the geographical area in which it is located, were also quite frequent: (A)*Göteborgs Posten*; 1: Article, 2: Gp, 4: Göteborg, 5: Posten, 6: Copyright. (B) *Sydsvenska Dagbladet*; 1: Malmö, 6: Lund, 11: Getinge, and 12: Sydsvenskan. An initial analysis seems to indicate that Göteborgs Posten was

using a language more site-centric than that of Sydsvenska Dagbladet, relative to the rest of the text.

Table 8: Table of the most common words at the Götebortgs Posten and Sydsvenska Dagbladet Web sites on September 23rd, after filtering and stemming of the frequency lists

	Göteborgs Posten		**Sydsvenska Dagbladet**
Freq	**Topics and Synonyms**	**Freq**	**Topics and Synonyms**
2387	Article/Artikel: Artikeln	14035	Malmö: Malmoe, Malmös
2127	Göteborgs Posten/Gp	11744	Plm (Swedish company)
1840	Olympic Games /Os	10177	Novo: Novos (Danish company)
1738	Gothenburg/Göteborg: Göteborgs, Göteborgare, Göteborgarna, Göteborgarnas	10173	Nordic/Nordisk: Nordiska, Nordiskt, Nordiske
1584	Mail/Posten: Post, Postade, Postar, Poste, Postens, Poster, Posterna, Postum	9910	Trelleborg: Trelleborgen, Trelleborgs (Swedish company)
1478	Copyright	9422	Lund: Lunden, Lunds (Swedish City)
1282	USA	8710	Skanska: Skanskas (Swedish company)
1259	Year/År	8528	Danisco: Daniscos (Danish company)
1227	Film: Filma, Filmad, Filmade , Filmar, Filmare, Filmares, Filmarna, Filmas, Filmat, Filmats, Filmen, Filmens, Filmer, Filmerna, Filmernas	7087	Astra (Swedish company)
1196	Page index/Sidindex	6949	Ericsson: Ericson, Erikson, Erickson, Ericssons, Eriksson (Swedish company)
1192	Previous/Föreg	6908	Getinge: Getingar, Getingarna, Getingen
	(The stemming did not recognise the difference between the name of a geographic place (Getinge) and the Swedish word for wasp (geting))		
1157	Copy	6520	Sydsvenskan: Sydsvenska, Sydsvenskans, Sydsvensskans (Synonymous for the newspaper)
982	Sweden/Sverige: Sveriga Sveriges	5874	Pharmacia (Swedish company)
624	Day/Dag	5791	Carlsberg: Carlsbergs (Danish company)
623	Swedish/Svenska: Svensk, Svenskar, Svenskarna, Svenskarnas, Svenskars	5532	Kr (Krona is Swedish and Danish currency)
509	Germany/Tyskland: Tysklands	5102	Oticon: Oticons (Danish company)
506	Russia/Ryssland: Rysslands	5048	Registration/reg
500	Final: Finalen, Finalens, Finaler, Finalerna, Finale	4671	Tele: Talet, Talar, Tala, Talade, Talades, Talan, Talare, Talarna, Talas
	(The stemming did not recognise the difference between tele (as in telephone) and tala (the Swedish word for speech))		
484	Atlanta: Atlantas	4651	Althin (Swedish company)
		4648	Denmark (abbrev)/Danm: Danma, Danmar

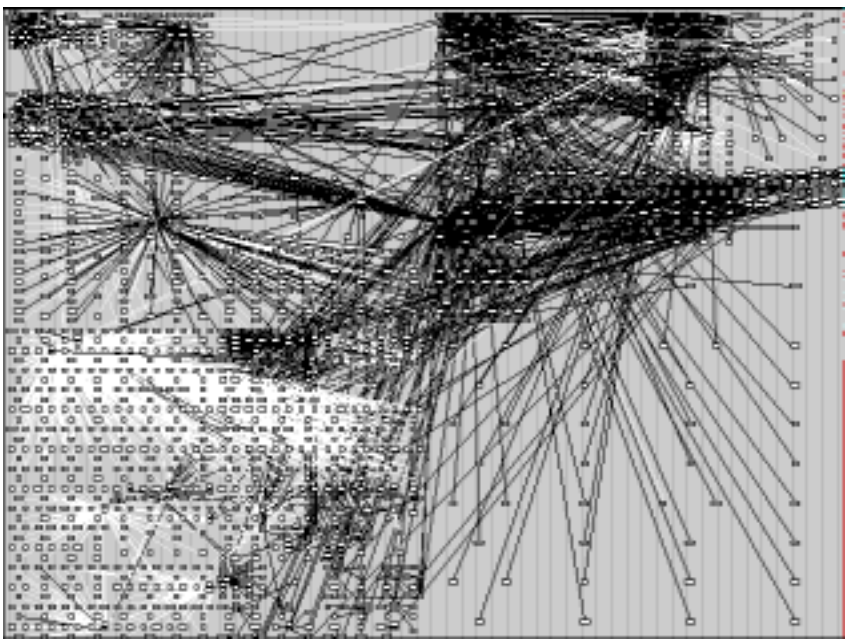
This semi-automatic technique of identifying Web site topics is an example of filtering and summarisation, i.e., eliminating unwanted information and representing a large amount of information with a smaller one (Nielsen, 1999). Interesting and quick overviews

could be provided by the Web site if techniques such as the ones demonstrated here were applied as a service to visitors. In a more automated version, it could also provide important high-level information about trends in the development of a Web site over time by comparing lists of most common topics on a regular basis.

5.3Mapping Site Structure

How can a measure provide an overview of the structural aspects of a Web site? One way is, of course, to compile a graph representing the inter-document linking. Such an approach would provide a very detailed model, but also one where advanced technology for information visualisation would be necessary in order to represent the complexity of the information space (Young, 1996). The purpose of such a model would, further more, be to represent the full complexity of the hyperspace in order to, for example, support navigation through the use of sophisticated virtual reality technology. The use of this type of information visualisation techniques, where the full complexity of the structure is conserved, can automatically lead to information overload, as can be seen from Figure 2, where we have shown the link structure of the informatics.gu.se Web site at Göteborg University. Projects such as Apple's Project-X have made attempts to alleviate this problem without reducing the amount of detailed information available (Young, 1996).

Figure 2: The link structure of the informatics.gu.se site at Department for Informatics, Göteborg University.



How can we present a picture which does not so easily lead to information overload, but instead provides an overview where information on purpose has been discarded in order to provide the overview? Chakrabarti et al. (1999) analyse link structures to establish hubs of authority. Li (1998) uses a similar technique as the basis for constructing a search engine index. If, however, the aim is to detect and visualise change to a Web site, then the exact structure of the Web directory is not of the utmost importance. It is also important to be able to reflect changes. One first reduction of complexity is to look at the directory structure in which the html files are stored. This, of course, implies that the data is analysed as a tree structure, and not as a graph. For large sites, this, however, can also lead to problems of information overload. We have, therefore, adopted the CyberGeo maps concept as a visual

mechanism for providing overview of Web site structure (Holmquist et al., 1998). Cyber-Geo maps provide a means for monitoring changes to the names and dates of files and directories by visualising the site's directory and document structure.

Figure 3: Five CyberGeo maps of Göteborgs Posten showing the changes in directories for the Web site from September 23rd to November 4th 1996. New files and directories, i.e., with changed name and/or time stamp are shown as white circles. Files with name and time stamp which are identical to the previous search will for each time be drawn in a darker shade.

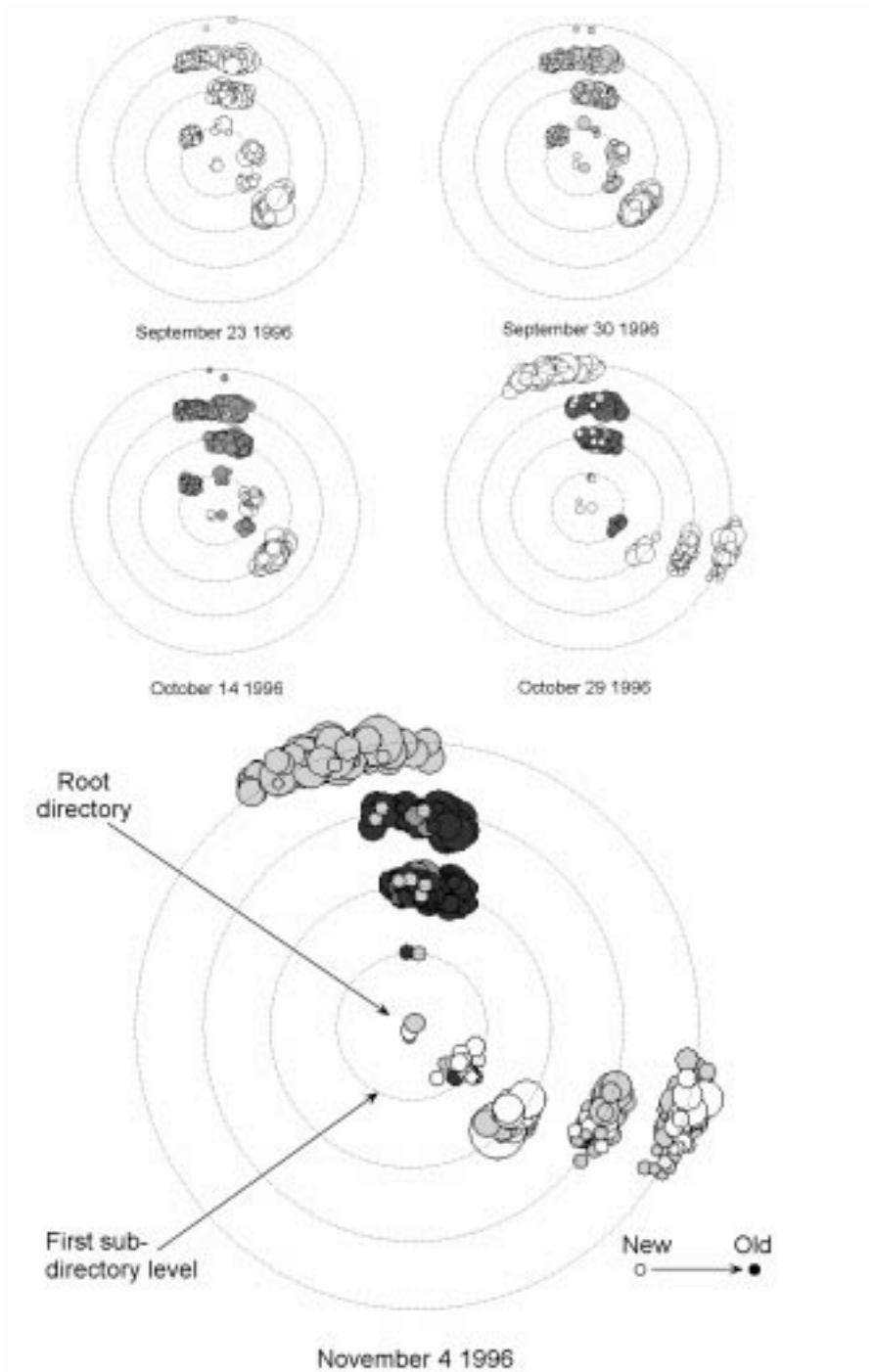
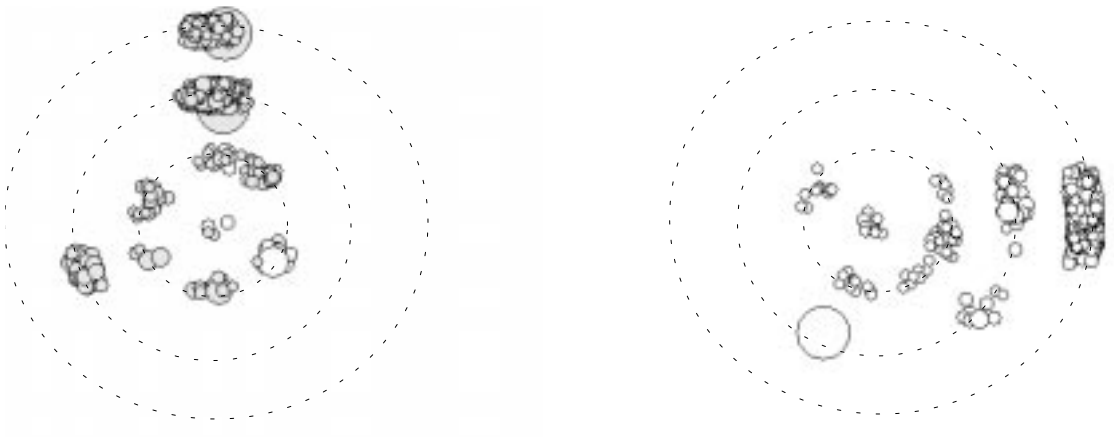


Figure 3 shows five consecutive CyberGeo maps for Göteborgs Posten. On a given map, each of the many discs represents a document. Applying a simple hash function on the document file name each disc is placed in a circle around the centre. The size of a document is represented by the size of the disc.First time a document is encountered, it is represented as a white disc. If a particular document has not been changed or deleted since previous measurement, the disc representing the document will assume a slightly darker shade of grey. Documents, which are unchanged will gradually, change from white to black as each measurement is made. The innermost of the concentric circles is the directory root. Documents in the same sub-directory are placed in an outer circle within ten degrees of its parent directory. This approach to presenting a high-level view of the Web site obviously only provides a simplified view. The mapping technique does not take into consideration that a Web site potentially is a complex graph with links across and within documents (see Figure 2), and potentially with documents only serving as aliases or references to other documents somewhere else in the tree structure. It assumes that it makes sense to view the Web site as a tree structure.

The reorganisation of the Göteborgs Posten site mentioned in Section 5.1 between sample II and IV can very easily be detected from the CyberGeo map sequence in Figure 3. However, Figure 4 illustrates even more clearly how a re-organisation of a Web site will be visualised by the CyberGeo map. Here, two consecutive CyberGeo map pictures for the Omgroup Web site (`www.omgroup.com`) illustrate a major redesign, where the document representations are almost turned 90° clockwise as a result of almost all documents being renamed.

Figure 4: Example of a CyberGeo map shift due to major revamp of Web site. The radical change in the CyberGeo map of this site represents the fact that the Web site was revamped, amongst others resulting in most directories and files being renamed. The large area on the second circle on the right map figure represents the company's annual report.



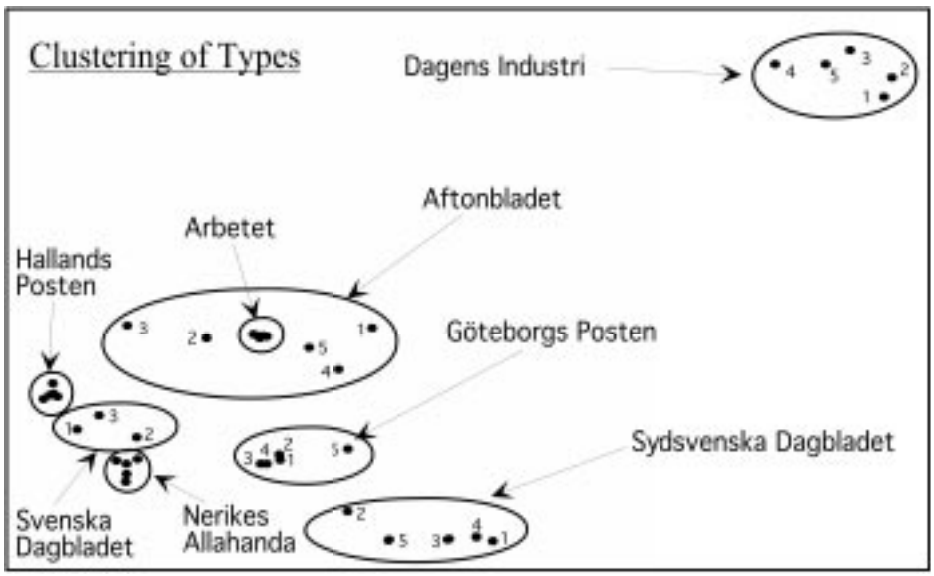
5.4Comparing Several Sites

Having analysed how we can summarise the contents and topics of a Web site, amongst others, supporting the comparison of two sites, we now turn to the issue of comparing a number of Web sites to each other with respect to the language used. At this level, all of the eight newspaper sites were analysed in relation to each other by a cluster analysis on lexical equality of both types (see Figure 5), and tokens. We did not find significant differences between clustering tokens and types, so only the type-based approach is shown. Based on the list of types, the clustering algorithm will group together observations based on the degree of commonality between the observation and all other observations. The clustering

algorithm visualises the percentages of lexical equality between sites in two-dimensions. There are no axes in the figure, only relations, e.g. the upper right corner is the least equal to the lower left corner. Each sample is represented as a plot, and each of the sites were sampled five times, except Svenska Dagbladet where the webmaster blocked out our robot using the robots exclusion standard (Koster, 1997) during the last two samples.

As seen in Figure 5, none of the sites changed their language significantly during the period, and the samples are therefore within a small region in the clustering. In particular, the language on the Web sites for Arbetet, Hallandsposten, and Nerikes Allehanda changed very little over the period, which is not surprising since none of the other variables collected change significantly either. The variation in tokens and types within one particular site was in most cases less than the variation between sites, making it possible to represent each site as a region (See Figure 5). As clearly demonstrated by Figure 5, Dagens Industri was quite different from the others Web sites. The reason for this is most likely that it is a financial newspaper that, due to the limited scope, used a different language compared to the others. Arbetet and Aftonbladet are both associated with the Swedish Social Democratic Party. Although only the editorial in Aftonbladet has a distinct political flavour, it seems as though the language of the two newspapers is much the same. These two sites were quite similar in terms of tokens used, but as Figure 5 shows an even greater similarity between the two when clustering types, i.e. comparing the two lists of distinct words. Techniques such as clustering frequency lists of tokens or types can provide a valuable overview of the relative similarities between a number of Web sites.

Figure 5: Clustering on lexical quality of types for the newspapers at five occasions.



5.5Surveying A Messy World

One of the main problems in surveying the World Wide Web is that the HTML "programming" environment does not enforce strong typing. If a particular statement is incorrect, the Web browser will just skip it, making it difficult for the author to ensure correct markup. This, amongst others, results in an extremely "messy" environment to survey. Table 9 shows to the left the set of correct HTML tags at the time of the survey, and to the right examples of markup encountered during the sampling. The table does not indicate any relationship between the proportion of correct and incorrect tags. It is basically the list of tag types divided into correct and incorrect ones. With the emergence of authoring technologies sup-

porting the automatic generation of HTML documents, this problem could be expected to diminish over time.

Table 9: Listing of the HTML tags found during the survey, distinguishing between the correct and incorrect ones.

Correct HTML Tags	Correct HTML Tags	Correct HTML Tags	Incorrect HTML Tags	Incorrect HTML Tags	Incorrect HTML Tags
a	Em	option	abcsubmit	gave	ps
address	Font	p	adress	grin	rfpicr
align	Form	param	ahref	hrnoshade	shift
applet	Frame	pre	aired	htm	silence
area	Frameset	right	aircheck	http	stations
b	h[1..6]	script	and	il	svd
base	head	select	are	imgsrc	tdalign
basefont	hr	small	bgsound	inputtype	tdcolspan
blink	html	strong	bodybgcolor	is	tdnowrap
blockquote	i	sub	border	it	tdwidth
e	img	table	brt	jberg	textareawrap
body	input	td	by	jpd	thank
br	left	textarea	ceneter	krone	tsfinfo
caption	li	th	centre	m	valign
center	kbd	title	clear	marquee	vk
cite	link	tr	color	means	w
code	map	tt	colw	moore	wbr
dd	menu	u	embed	name	were
dir	meta	ul	emp	nobr	width
div	noframes	width	front	of	wireless
dl	ol		fontsize	palign	www
dt					

6. Discussion

This paper discusses important lessons we can learn about the contents of and changes to the fastest growing technology in the world, the World Wide Web, through employing semi-automatic retrieval and aggregation of Web sites. The paper investigated this question by presenting and discussing results from an experimental survey where Web robots supported the collection and analysis of 82 Swedish Web sites over a seven-week time-span from September 23 to November 4 1996. It should be noted here that during the course of the experiment a couple of the sites banned robots from accessing data.

We have demonstrated how changes to the contents of Web sites can be detected through analysing changes to the basic parameters, such as number of bytes or link errors. CyberGeo maps provided a technique for mapping and monitoring changes to the Web site architecture. Clustering lexical equality of tokens and types across the 8 Swedish newspaper sites and representing the result in a two dimensional plot illustrated linguistic similarities and differences across Web sites. The results did not only demonstrate the sampling instruments' ability to detect changes, but also that a closer calibration of the instrument must be conducted over a longer time-span than seven weeks in order to amplify the sensitivity to significant changes. For example, the samples showed an increase in number of tokens of 9.5% over the seven weeks. If scaled to one year this amounts to around 70%. We recognise that this is a significant change, but compared to what? In general, a sampling period of only seven weeks most likely proved to be too short for obtaining results showing large variations.

In general, compiling frequency lists has provided us with much deeper material about the contents of the Web sites, compared to the analysis conducted by Bray (1996). This, however, is associated with a much smaller sample. Common readability formulas such as Coleman-Liau grade level and Bormuth grade level could be used to further analyse the sampled texts. These indexes determine a readability grade level, based on characters per word and words per sentences and are therefore relatively easy to calculate. Word processors, such as, Microsoft Word uses these types of indexes in their grammar-checking facilities. Also, a further analysis of differences and similarities between the language used in newspaper Web sites compared to the printed papers could further enlighten the discussion of new media (Tesitelová, 1992; Hagman and Ljungberg, 1995; Palmer and Eriksen, 1999; Eriksen et al., 2000).

In order to detect statistically significant differences between Web sites within the six sectors, we conducted a number of discrete statistical analyses using contingency tables and Chi-Square tests. This analysis was based on the measures showed in Section 3, and showed some relative trivial results, such as a dependency between site size and number of link errors. This analysis revealed a number of dependencies between, for example, the size of a site and the number of link errors. It did, however, not show any further results in terms of significant differences between sites from different sectors. The reason for this could be the very short sampling period, and the absence of substantial contextual parameters which we could not establish without a more qualitative approach.

What are the possible implications of this research? All of the techniques employed can be operationalised further and made subject for increased automation, and thus provide support for increasingly automatically generated sophisticated maps of the Web. As the Web increases in size and complexity, there will be an increased need for supporting overview and navigation. The unstructured nature of the Web, which has been identified as one of its primary assets, is however, recognised as beginning to cause severe navigation problems. There is no centrally maintained classification scheme guiding people when they wish to access information. The Web can be navigated using a bottom-up search through one of the Web search engines (Type 1) or top-down through index pages providing taxonomies for Web contents (Type 2) (Gudivada et al., 1997). Because of the dynamic nature of the Web, there is considerable overhead associated with locating and maintaining links to desired resources on the Web. The more extensively an individual uses the Web, the greater this overhead will be. Software agents have been promoted as a means of supporting navigation (Maes, 1994; Krulwich, 1997; Maglio and Barrett, 1997; Schneiderman and Maes, 1997). Current efforts suggest semantic markups supporting specific markup types characterising Web contents (Khare and Rifkin, 1997; Wired, 1997; Lassila, 1998; Rein, 1998). This structuration of the Web would support standardisation of text types within domains and supporting increased automatic analysis of Web page contents, for example within electronic commerce (Glushko et al., 1999).

Currently, the issues of electronic commerce and the management of knowledge are at the fore of both research and public debate. The electronic market is increasingly becoming associated with the Internet infrastructure and the Web most often serves as the preferred front end. Companies are finding the Web a productive place to market products, and it is important to understand how the company Web site can play an important role (Smithson, 1999). For potential customers, the ability to easily obtain substantive information about the contents of a Web site is considered as important as being able to locate a shop on the high-street. Similarly, if for example, a car company wishes to know where to market its most recent model, it would be extremely useful to know where cars are intensively discussed and where they are not. Topic spotting on selected Web sites could provide an index to the contents of the site. This is why software agents have been suggested as a means of mediating in an electronic market (Glushko et al., 1999; Maes et al., 1999)

The management of knowledge has also frequently been linked to Internet and Web based systems supporting people in sharing knowledge and interacting (Scarbrough et al., 1999). If people in organisations stores textual traces of their activities in intranets, there

will be an inherent need for semi-automatic techniques that provide an overview of and access to this information (Stenmark, 1999; Robertson et al., 2000).

In both these cases, there will be substantial challenges involved in creating patterns and structure from unstructured bodies of text with the purpose of answering questions or connecting people. Data mining structured databases involves very complex calculations, and data mining the unstructured textual traces on the Web is an even greater challenge (Etzioni, 1996; Kawano and Hasegawa, 1998; Chakrabarti et al., 1999; Guan and Wong, 1999). We believe that one viable strategy is to apply relatively simple techniques for abstracting databased on pragmatic heuristics. Nielsen (1999) argues for the need for functionality supporting: (1) Aggregation - showing a single unit that represents a collection of smaller ones; (2) Summarisation - representing a large amount of information with a smaller one; (3) Filtering by eliminating unwanted information; and (4) Elision which is example-based representation. The techniques demonstrated in this paper provides examples of all of these categories, based on techniques from computational linguistics and information visualisation with filtering being the central technique. For example, the CyberGeo maps summarises an entire complex Web site in a relatively simple image and indicates if substantial changes has been made.

The results and approach presented in this article have provided one perspective with regard to the discussion of how to conduct quantitative surveys of the Web. It is because the Web is a highly dynamic and interactive information space, we must apply state-of-the-art computational power to study it. This will increasingly present itself as a challenge, and so far it has mainly been an issue for the professional IT community and not one carried out in a wider context. We are, however, increasingly living in Cyberspace, and will increasingly have to rely on high-level representations in order to make sense of the world and navigate it. Only time will tell to what degree most people will accept an agent-based human-computer interaction paradigm as opposed to a more conventional direct manipulation paradigm (Schneiderman and Maes, 1997)

7. Acknowledgements

This research is partly funded by the Swedish Transport and Communications Research Board (Kommunikationsforskningsberedningen) grants for the Internet Project (http://internet.informatik.gu.se). A big thanks to Leif Grönquist and Peter Ljungstrand for assistance on computational linguistics and to Michael Mandahl for bright ideas. Thanks to Maxine Robertson for extensive proof-reading. Thanks to the two anonymous reviewers providing us with valuable critique. Also thanks to all our other colleagues in the Internet Project for constructive comments.

8. References

Allen, R. B.: User models: theory, method, and practice. International Journal of Man-machine Studies, vol. 32, no. 5, pp. 511-543, 1990.

Ambrosini, L., V. Cirillo, and A. Micarelli: A Hybrid Architecture for User-Adapted Information Filtering on the World Wide Web. In User Modelling: Proceedings of the Sixth International Conference, UM97, Vienna, ed. Anthony Jameson, Cécile Paris, and Carlo Tasso. Springer Verlag, pp. 59-61, 1997.

Barry, B. and M. Batty: The Electronic Frontier - Exploring and mapping cyberspace. Futures, vol. 26, no. 24, pp. 699-712, 1994.

Berners-Lee, T., R. Calliau, A. Luotonen, H. F. Nielsen, and A. Secret: The World-Wide Web. Communications of the ACM, vol. 37, no. 8, pp. 76-82, 1994.

Bray, T.: Measuring the Web. In Proceedings of the 5th International World Wide Web Conference, Paris, France, 1996.

Chakrabarti, S., B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg: Mining the Web's Link Structure. IEEE Computer, vol. 32, no. 8, pp. 60-67, 1999.
CommerceNet: Nielsen Internet Demographic. http://www.commerce.net/nielsen/(1996)
Dodge, M.: Mapping the World Wide Web. GIS Europe, vol. 5, no. 9, pp. 22-24, 1996.
Dodge, M.: The Geography of Cyberspace, 1999.
Drew, N. S. and R. J. Hendley: Visualising Complex Interacting Systems. In CM SIGCHI '95 Proceedings of the Conference on Human Factors in Computing Systems, Denver, Colorado, USA, 1995.
Eichmann, D.: Ethical Web Agents. In Proceedings of the 2nd International World Wide Web Conference, Chicago, USA, 1994.
Eriksen, L. B., C. Ihlström, and J. Palmer: News. In Planet Internet, ed. Kristin Braa, Carsten Sørensen, and Bo Dahlbom. Lund: Studentlitteratur, 2000.
Etzioni, O.: The World-Wide Web: Quagmire or Goldmine? Communications of the ACM, vol. 39, no. 11, pp. 65-68, 1996.
Fielding, R. T.: Maintaining Distributed Hypertext Infostructures: Welcome to MOMspider's Web. In Proceedings of the 1st International World-Wide Web Conference, Geneva, Switzerland, 1994.
Garofalakis, J., P. Kappos, and D. Mourloukos: Web Site Optimisation Using Page Popularity. IEEE Internet Computing, vol. 3, no. 4, pp. 22-29, 1999.
Girardin, L.: Mapping the virtual geography of the World-Wide Web. In Proceedings of the 5th International World Wide Web Conference, Paris, France, 1996.
Glushko, R. J., J. M. Tenenbaum, and B. Meltzer: An XML Framework for Agent-Based E-Commerce. Communications of the ACM, vol. 42, no. 3, pp. 106-114, 1999.
Greenhalgh, C.: Analysing movement and world transitions in virtual reality tele-conferencing. In Proceedings of the Fifth European Conference on Computer Supported Cooperative Work, Lancaster, UK. Kluwer Academic Publishers, pp. 313-328, 1997.
Guan, T. and K. F. Wong: KPS - a Web Information Mining Algorithm. In Proceedings of the 8th World Wide Web Conference, Toronto, Canada. Elsevier, Amsterdam, 1999.
Gudivada, V. N., V. V. Raghavan, W. I. Grosky, and R. Kasanagottu: Information Retrieval on the World Wide Web. IEEE Internet Computing, vol. 1, no. 5, pp. 58-68, 1997.
Guice, J.: Looking Backward and Forward at the Internet. The Information Society, vol. 14, no. 3, pp. 201-211, 1998.
Hagman, J. and J. Ljungberg: Brute Facts versus Institutional Facts of Language as a Foundation for IR. In Proceedings of the 5th European-Japanese Seminar on Knowledge Bases and Information Modeling, Sapporo, Japan, 1995.
Hannemyr, G.: An Even Briefer History of the Internet. Oslo University. Department for Informatics. http://www.ifi.uio.no/~inint/emne01a.htm, 1998.
Hill, W. C., J. D. Hollan, D. Wroblewski, and T. McCandless: Edit wear and read wear. In Proceedings of the ACM 1992 Conference on Human Factors in Computing Systems, Monterey, CA. ACM Press, pp. 3-9, 1992.
Hoffman, D. L., W. D. Kalsbeek, and T. P. Novak: Internet and Web use in the U.S. Communications of the ACM, vol. 39, no. 12, pp. 106-108, 1996.
Holmquist, L. E., H. Fagrell, and R. Busso: Navigating Cyberspace with CyberGeo Maps. In 21st Information systems Research seminar In Scandinavian, August 8-11, Sæby Søbad, Denmark, ed. Peter-Axel Nielsen, Niels Jacob Buch, and Lars Bo Eriksen. Aalborg University, 1998.
Huhns, M. N. and M. P. Singh: Mobile Agents. IEEE Internet Computing, vol. 1, no. 3, pp. 80-82, 1997.
Jain, A. K. and R. C. Dubes: Algorithms for Clustering Data. NJ: Englewood Cliffs: Prentice-Hall, 1988.
Kawano, H. and T. Hasegawa: Mondou: Interface with Text Data Mining for Web Search Engine. In Thirty-First Hawaii International Conference on System Sciences (HICSS-31), Big Island Hawaii, ed. Ralph Sprague Jr. IEEE, 1998.

Khare, R. and A. Rifkin: XML: A Door to Automated Web Applications. IEEE Internet Computing, vol. 1, no. 4, pp. 78-87, 1997.
Koster, M.: The Robots Exclusion Protocol. http://info.webcrawler.com/mak/ projects/ robots/exclusion.html), 1997.
Krulwich, B.: Automating the Internet: Agents as User Surrogates. IEEE Internet Computing, vol. 1, no. 5, pp. 34-38, 1997.
Kumar, S. R., P. Raghavan, S. Rajagopalan, and A. Tomkins: Trawling the web for emerging cyber-communities. In Proceedings of the 8th World Wide Web Conference, Toronto, Canada. Elsevier, Amsterdam, 1999.
Lassila, O.: Web Metadata: A Matter of Semantics. IEEE Internet Computing, vol. 2, no. 4, pp. 30-37, 1998.
Leiner, B. M., V. G. Cerf, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. Wolff: A Brief History of the Internet. http://info.isoc.org/internet-history/brief.htm, 1997.
Leonard, A.: Bots: The Origin of New Species. San Francisco: HardWired, 1997.
Li, Y.: Toward a Qualitative Search Engine. IEEE Internet Computing, vol. 2, no. 4, pp. 24-29, 1998.
Maes, P.: Agents that reduce work and information overload. Communications of the ACM, vol. 37, no. 7, pp. 31-40, 1994.
Maes, P., R. H. Guttman, and A. G. Moukas: Agents that Buy and Sell. Communications of the ACM, vol. 42, no. 3, pp. 81-91, 1999.
Maglio, P. P. and R. Barrett: How to Build Modeling Agents to Support Web Searchers. In User Modelling: Proceedings of the Sixth International Conference, UM97, Vienna, ed. Anthony Jameson, Cécile Paris, and Carlo Tasso. Springer Verlag, pp. 5-16, 1997.
Mitchell, W. J.: City of Bits: Space, Place and the Infobahn. The MIT Press, USA, 1995.
Mukherjea, S. and J. D. Foley: Visualizing the World-Wide Web with the Navigational View Builder. Computer Networks and ISDN System: Special Issue on the 3rd International Conference on the World-Wide Web '95, no. April, 1995.
Nielsen, J.: User Interface Directions for the Web. Communications of the ACM, vol. 42, no. 1, pp. 65-72, 1999.
Oard, D. W.: The State of the Art in Text Filtering. User Modeling and User-Adapted Interaction: An International Journal, vol. 7, no. 3, pp. 141-178, 1997.
Olsen, K. A., R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams: Visualization of a Document Collection with Implicit and Explicit Links - The Vibe System. Scandinavian Journal of Information Systems, vol. 5, pp. 79-95, 1993.
Palmer, J. W. and L. B. Eriksen: Digital Newspapers Explore Marketing on the Internet. Communications of the ACM, vol. 42, no. 9, pp. 33-40, 1999.
Pitkow, J.: The WWW User Population: Emerging Trends. Slides presented at GlobeCom 1997 GVU Center. Georgia Institute of Technology. www.gvu.gatech.edu/user_surveys, 1997.
Rein, L.: XML-Enabled Tools. IEEE Internet Computing, vol. 2, no. 3, pp. 16-19.
Resnick, P. and H. Varian (1997): Recommender systems. Communications of the ACM, vol. 40, no. 3, pp. 56-58, 1998.
Robertson, M., C. Sørensen, and J. Swan: Managing Knowledge With Groupware: A Case Study of a Knowledge-Intensive Firm. In Thirty-Third Hawaii International Conference on System Sciences (HICSS-33), Maui, Hawaii, 2000.
Scarbrough, H., J. Swan, and J. Preston: Knowledge Management: A Literature Review. Issues in People Management. London: Institute of Personnel and Development, 1999.
SCB: Statistisk Årsbok 96. Stockholm: Statiska Centralbyrån, 1996.
Schneiderman, B. and P. Maes: Direct manipulation vs Software Agents: Excerpts from debates at IUI 97 and CHI 97. Interactions. November-December, pp. 42-61, 1997.
Schubert, C., R. Zarnekow, and W. Brenner: A Methodology for Classifying Intelligent Software Agents. In ECIS98, Aix-en-Provence, ed. Walter R.J. Bates, vol. I, pp. 304-316, 1998.

Senanayake, C. R.: WebSpy: Towards Supporting Persistence of Interest on the Web. Masters Dissertation. Computer Science, Aston University, 1998.
Smithson, S.: The 1999 World Wide Web 100. Department for Information Systems. The London School of Economics and Political Science, 1999.
Sørensen, C.: Where Have You Been Today? Investigating Web Navigation Support. In 21st Information systems Research seminar In Scandinavian, August 8-11, Sæby Søbad, Denmark,, ed. Peter-Axel Nielsen, Niels Jacob Buch, and Lars Bo Eriksen. Aalborg University, 1998.
Stenmark, D.: Capturing Tacit Knowledge Using Recommender Systems. In IRIS 22, Jyväskylä, Finland, ed. Timo Käkölä. Department of Computer Science. Jyväskylä University, 1999.
Tauscher, L. and S. Greenberg: Revisitation Patterns in World Wide Web Navigation. In ACM SIGCHI '97 Proceedings of the Conference on Human Factors in Computing Systems, Atlanta, Georgia. ACM Press, 1997.
Tesitelová, T.: Quantitative linguistics. Amsterdam: John Benjamins Publishing Company, 1992.
Tufte, E. R.: The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphics Press, 1983.
Whittaker, S., L. Terveen, W. Hill, and L. Cherny: The dynamics of mass interaction. In Proceedings of the ACM 1998 Conference on Computer Supported Cooperative Work, Seattle, WA. ACM Press, pp. 257-264, 1998.
Wired: Automatable Web. Wired Magazine, August 1997, pp. 41, 1997.
Wooldridge, M. J. and N. R. Jennings: Software Engineering with Agents: Pitfalls and Pratfalls. IEEE Internet Computing, vol. 3, no. 3, pp. 20-27, 1999.
Young, P.: Three Dimensional Information Visualisation. Visualisation Research Group. Centre for Software Maintenance. Department of Computer Science. University of Durham. Computer Science Technical Report, No. 12/96. http://www.dur.ac.uk/~dcs3py/pages/work/documents/lit-survey/IV-Survey/index.html, 1996.
Zakon, R. H. o.: Hobbes' Internet Timeline v4.0. http://info.isoc.org/guests/zakon/ Internet/History/HIT.htm, 1998.