IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4494

A Method to Determine the Fraction of the Scientific Literature Within a Domain That is Available on the World Wide Web

Ketil Perstrup
Department of Computing, Copenhagen University, Universitetsparken 1, DK-2100 Copenhagen East, Denmark, Ph: +45 35 32 14 00, Fax: +45 35 32 14 01, ketil@diku.dk

ABSTRACT

The World Wide Web is often seen as a way for scientists and professionals to disseminate information to people outside their own communities. The current World Wide Web search engines make it hard to locate information for scientific and professional use, and search engines focused on one domain have been suggested as a solution. The success of a search engine focused on one domain depends on whether there are enough documents available on the World Wide Web to make the engine useful. To answer this question we propose a method to determine how large a fraction of the literature within a domain that is available on the World Wide Web. We use the method to determine fraction of the computer science literature that is available on the Web and estimate that 18% of the computer science papers published in 1997 is available on the World Wide Web. We also discover that a large amount of unpublished papers is available as well. We believe that with this amount of information available a search engine focusing on computer science documents on the World Wide Web could be useful in the work of scientists, students, and professionals.

INTRODUCTION

Having all the literature in your field available at your fingertips appeals to scientists and professionals in a way that is easy to understand. The idea has fascinated researchers at least since Vannevar Bush described his Memex in 1945 (Bush 1945). The World Wide Web has been seen as a way to fulfill this dream and as a way for researchers and professionals to disseminate their work to persons outside their own communities (Barrie & Presti 1996). While the Web seems to be popular among scientists and professionals as a way to present their work, searchers using the Web search engines have problems locating this information. The problems that occur when searchers use the Web search engines to search for information for scientific and professional use have at least three causes:

1. Web search engines have problems keeping up with the size and dynamics of the Web,
2. No Web search engine cover all of the Web (Lawrence and Giles 1998), and
3. Web search engines will retrieve information from a multitude of domains (Brake 1997).

We believe that one way to make it easier to find information intended for scientific and professional use is to create focused search engines providing in-depth coverage of the Web for specific domains, possibly with tools tailored to the domain as suggested by Brake. This paper shows how to estimate the fraction of the papers published in a domain that is available on the Web. We can use this estimate to determine whether it is feasible to use existing scientific documents on the Web as a basis for a search engine focused on domain.

There have previously been some attempts to quantify the size of the Web, such as the study by Lawrence & Giles (1998), but we are unaware of any previous attempt to measure the amount of information available for specific purposes. As a first attempt to study this, we investigate how large a fraction of the computer science literature that is available on the World Wide Web. The reason we have chosen to survey computer science literature is that we expect computer scientists to have had an early exposure to the Web. We therefore assume that trends in the way scientists make publications available on the Web will show up among researchers in computer science quite early compared to researchers from other domains. The method described in this paper is used to estimate the fraction of the papers published within a domain that is available on the World Wide Web. We do that by examining papers and publication lists that are available from researchers' personal Web pages. We will discuss the potential problems with the method and the resulting estimates later.

WHY CREATE A FOCUSED SEARCH ENGINE?

The Web search engines try to cover the entire Web, and this causes at least three problems for the Web search engines and their users:

1. The Web search engines have problems keeping up with the size and dynamics of the Web
2. No Web search engine cover all of the Web and
3. Web search engines will retrieve information from a multitude of domains.

This causes problems for searchers when they try to find information for scientific or professional use.

The main problem for the Web search engines is perhaps to keep up with the size and dynamics of the Web. The disparity between what is needed and what can be achieved technically makes this hard. Descriptions of the performance of operational Web search engines are scarce, but the search engine described by Brin & Page (1998) is able to collect and index roughly 600K of data

per second or 49 megabytes a day. Lawrence and Giles (1998) estimate that by December 1997 there were 320 million documents on the World Wide Web. These documents contain on the average 6.5 kilobytes of text, not including pictures, sounds, and other non-text material (Brin & Page 1998), and they have an estimated average lifetime of 44 days (Kahle 1997). A search engine would have to store approximately 1983 gigabytes and it would have to transport and index 45.1 gigabytes a day to index the Web by December 1997. This means that the search engine described by Brin and Page can only collect and index about 0.1 percent of what is needed.

The resource problems also seem to cause the search engines to leave out large parts of the Web from their indexes. Lawrence and Giles (1998) estimates that the largest Web search engines cover at most 34% of the Web. This means that information that is available in only one location, such as information from researchers' personal Web pages, has a high probability of not being indexed by the major search engines. The problem is made worse by the fact that many documents that are made available by researchers are made available in formats other than HTML, which means that the Web search engines do not index them.

The Web search engines often retrieve information from a large number of sources and domains, since the Web search engines try to cover everything on the Web. This means that searchers that use the Web search engines have to sift through documents from several domains (Brake 1997) to find what they are looking for.

These observations indicate that the Web search engines are far from ideal when used to search for information intended for scientific or professional use. We believe that one way to improve the situation would be to create focused search engines providing in-depth coverage of Web documents for specific domains. Several researchers are investigating the problems associated with this type of search engines (Perstrup et al 1997) (Chakrabarti et al 1999). Focused search engines have at least two possible benefits. The first is that a focused search engine would need to index a much smaller set of documents than the general Web search engines so the resource requirements for a focused search engine is lower. In addition, searchers using a focused search engines not have to worry about sifting through documents from other domains. This should make it easier to locate the information that is sought. We hope that these benefits will make focused search engines cheaper to run and easier to use. The success of a focused search engine depends on whether there are enough documents available on the World Wide Web to make the engine useful. To answer this question we propose a method to determine how large a fraction of the literature within a domain that is available on the World Wide Web.

DEFINITIONS

In this paper, we will use the term *a paper* for any document with scientific information such as a book, a chapter of a book, an article in a journal or a technical report. We will use the term *a published paper* for any paper that, according to the author, has been published in a scientific context. We will say that a paper is *available* if the full text of a paper is available on the Web and *unavailable* if it is not. Available and unavailable papers will be called *referenced papers* if one of the authors has a publication list available on the Web that includes the paper. Finally, a paper that is available but have not yet been published, such as a paper submitted for publication, will be called *an unpublished paper*.

METHOD

We will determine the fraction of computer science papers that is available on the Web by collecting a random sample of computer scientists and a list of their papers. We then examine the papers and note whether they are available on the Web.

The researchers in this study was chosen by selecting fifty institutes randomly from the list of computer science institutes at Yahoo (Yahoo 1998) and then randomly selecting a fifth of the researchers from each institute. For each researcher we recorded whether they had any personal Web pages, whether the Web pages contained a list of papers, and whether the list was complete. To determine how many of the researchers' papers that were available, we divided the papers from each researcher's list of papers into seven groups: Papers published in 1998, 1997, 1996, 1995, 1994, before 1994 and unpublished papers. For each group we recorded the number of papers published and the number of papers that was available. In all cases we used the researcher's statements of whether, when, where and how the papers had been published.

RESULTS

The geographic location of the fifty institutes we selected for this study is shown in table 1. We had to exclude eight of the institutes because we were unable to obtain the information we needed. The causes for this was:

- In six cases, we could not find information about which researchers were associated with the institute,
- In one case, we could not connect to the server listed at Yahoo despite repeated attempts over three weeks, and
- In one case, we were unable to translate the information from Japanese.

Table 1 The countries in which the institutes were located. The column named No. of institutes *is the number of institutes from each country. Numbers in parentheses is the number of institutes discarded because we were unable to retrieve information about the researchers at the institute.*

Country	No. of institutes	Country	No. of institutes
Austria	2	Japan	1 (1)
Australia	2	Mexico	0 (1)
Canada	1	New Zealand	1
Germany	1	Poland	1
United States	24 (4)	Romania	0 (1)
Spain	1	Turkey	1
Greece	1	United Kingdom	3 (1)
Hong Kong	1	South Africa	1
Italy	1		

In one case, we had to locate the server at the computer science institute through the server at the university, since the URL from Yahoo was not valid. The results presented in this paper is based on information about the researchers associated with the remaining 42 institutes.

Researchers Covered by the Study

Table 2 Number of researchers that had personal Web pages, referenced one or more of their papers, made complete publication lists available, and made one or more papers available.

Researchers that	Number of researchers	Percent of all researchers
Had personal Web pages	167	76%
Referenced one or more of their papers	91	43%
Made complete publication lists available	69	32%
Made one or more papers available	57	27%

Table 3 The number of papers in each group and how many that were available and unavailable.

	Unpub-lished	Jan-Jun 1998	1997	1996	1995	1994	Before 1994	All groups
Available	65	24	120	100	91	95	132	627
Unavailable	100	74	130	160	173	209	1249	2095
Listed	165	98	250	260	264	304	1381	2722
% available	*39%*	*24%*	*48%*	*38%*	*34%*	*31%*	*10%*	*23%*

The 42 institutes that we selected had 1063 researchers associated with them. We randomly selected a fifth of the researchers from each institute for a total of 214 researchers. Table 2 shows how many researchers had personal Web pages, how many that referenced one or more of their papers, how many that made complete publication lists available, and how many that made one or more papers available.

Papers Covered by the Study

We collected the papers in this study by retrieving publication lists from the researchers' personal Web pages. We found 2722 references to papers with 627 papers (23%) of those available. The references were divided into groups based on whether the paper had been published in 1998, 1997, 1996, 1995, 1994, before 1994, or it was unpublished. Table 3 shows the number of papers that was referenced and how many of those that were available in each group.

Two things should be noted: First, our data shows a remarkably large number of publications in 1994 compared to other years. We cannot explain this but, as we will see later, shows up elsewhere. Second, compared to previous years, there seems to be a decrease in the number of published and available papers in 1998. We do not have data from all of 1998 since the data was collected in June 1998, so the decrease may be a byproduct stemming from the time of year that the data were collected.

The Papers Available and Their Authors

There are large individual differences in how many papers researchers make available. Table 4 compares the number of papers published by each researcher with the number of papers that is available. The table shows data for papers published by researchers that have made complete publication lists available and compares the number of papers published and the number of papers made available in 1994 and 1997. There seems to be to types of researchers: Those that make (almost) everything they publish available and those that make nothing available. The table also shows that the number of researchers that make everything available has increased from 1994 to 1997.

Estimating the Fraction of Papers Available

The data allows us to estimate the fraction of the papers published in computer science that is available on the Web. If we assume that the researchers in this study are typical for the domain, then the fraction of papers from computer science that is available on the Web is the number of papers observed to be available divided by the total number of papers written by all researchers. Unfortunately, many of the researchers have not made full publication lists available so we also need an estimate for the number of publications made by these researchers. To find this, we assume that the average number of papers published by the researchers is the same as the average number of papers published by the researchers that has made complete publication lists available. With this assumption, the estimated total number of papers published by all researchers is the number of researchers times the average number of papers published by researchers with complete publication lists. This results in the following formula:

percentage_available = *papers_available* / (*researchers* · *avg_papers_pr_researcher*)

where

percentage_available is the estimate for the fraction of papers that is available

papers_available is the number of papers that is available.

researchers is the number of researchers (214) and

avg_papers_pr_researcher is an estimate of the average number of papers published by the researchers.

Table 5 shows the number of papers published by the 69 researchers with complete publication lists available, the average number of papers published, and the estimate for the fraction of papers available.

Table 4 Number of researchers that published a specific number of papers in 1994 and 1997 compared to how many of the papers made available. The table only includes data from researchers that made complete publication lists available.

		Number of papers published in 1994							
		0	**1**	**2**	**3**	**4**	**5**	**6**	***Sum***
	0	0	0	2	12	9	10	4	*37*
	1		0	1	2	6	0	1	*10*
Number	**2**			2	0	2	0	1	*5*
of papers	**3**				4	3	0	0	*7*
available	**4**					5	1	0	*6*
	5						3	0	*3*
	6							1	*1*
	Sum	*0*	*0*	*5*	*18*	*25*	*14*	*7*	*69*

		Number of papers published in 1997							
		0	**1**	**2**	**3**	**4**	**5**	**6**	***Sum***
	0	0	0	2	7	3	0	0	*12*
	1		2	1	1	0	0	0	*4*
Number	**2**			11	3	0	0	0	*14*
of papers	**3**				19	1	0	0	*20*
available	**4**					15	0	0	*15*
	5						3	0	*3*
	6							1	*1*
	Sum	*0*	*2*	*14*	*30*	*19*	*3*	*1*	*69*

HOW WELL DOES THE STUDY COVER THE DOMAIN?

Before we go on, we would like to examine the extent to which we have covered the computer science literature in this study. The ACM Electronic Guide to Computing Literature (ACM 1998) lists 24645 different authors of publications in 1996. Our study examines the Web pages of 214 of those corresponding to 0.9 percent of the authors in the ACM Guide. Table 6 lists the number of papers published 1980–1996 according the ACM Guide. The table shows that our study of 2722 papers examines approximately 1 percent of the number of papers included in the ACM Guide with the exact fraction depending on which group we examine. It is interesting to

Table 5 Total number and average number of papers by the researchers with complete publication lists and the resulting estimate for the fraction of all papers that are available.

	Unpub-lished	Jan-Jun 1998	1997	1996	1995	1994	Before 1994
Papers listed	146	83	217	222	215	278	1172
Avg. no. of papers	2.12	1.20	3.14	3.22	3.12	4.03	16.99
Percent available	*14%*	*9%*	*18%*	*15%*	*14%*	*11%*	*4%*

note that the data from the ACM Guide show the same increase in the number of papers published in 1994 as our data. We are unable to explain this phenomenon.

DISCUSSION

The Web seems to be a popular way for researchers to present themselves and their work. As it can be seen from table 2, 76% of the researchers in this study have personal Web pages and 27% of them make at least some of their work available on the Web. Given the number of researchers that present their work on the Web we must conclude that computer science researchers consider the Web an important medium for presenting their work. However, there are large individual differences between the researchers. Less than half (43%) of the researchers had personal Web pages that referenced any of their papers but 51 (24%) of the researchers made every paper they published in 1997 available and several researchers made unpublished papers available as well.

Table 5 shows that a substantial and increasing fraction of the papers that are published by computer science researchers seems to be available on the Web. The table indicates that the amount of papers available on the Web has increased from 11% to 18% between 1994 and 1997. Table 4 shows that the number of researchers that make everything they publish available is increasing. The data have been collected over a short range of years, so extrapolation must be done with some caution. There is, however, a clear increase in the fraction of papers available on the Web for the range of years examined.

There are some potential problems with the method used to estimate the number and fraction of publications available. The results of this survey might include a bias in the selection of researchers that favor US researchers, since we selected researchers from a US Internet service. Beyond the introduction of a US bias, we have excluded researchers from institutes that are not on the Web. Since the Internet has long been a research tool for computer scientists, we expect most computer science institutes to be on the Web. We therefore assume that the number of computer science institutes that we may have excluded is small, leading to an estimate that is slightly too high. On the other hand, there might be sources for a researcher's papers other than their personal Web pages such as Web pages for scientific journals, co-authors' Web pages, and the Web pages for research groups. We have only examined researchers' personal Web pages so we may have left out papers that are available from these sources because the researchers themselves have not referenced them and that would cause our estimates to be too low. Finally, we have only studied researchers at universities although there are several companies with highly regarded research groups so this study cannot be used to infer the behavior of researchers outside universities.

CONCLUSION

We have examined the possibility of using documents from on the Web as a basis for a Web search engine focused on documents from one domain, since the general Web search engines are not good for searching for documents for scientific and professional use. Using the method described, we estimate that the fraction of the papers written by computer scientists that is available on the Web has grown from approximately 11% in 1994 to 18% in 1997. Our results also indicate that more researchers are making their papers available on the Web. Better tools for Web publishing are appearing in standard word processing programs so we expect a continued increase in the fraction of papers available on the Web. A large number of unpublished papers that has been submitted or accepted for publication are also available on the Web. The large number of papers that is available means that we believe that that a search engine focusing on computer science papers available from researchers' personal Web pages would be a useful tool for scientific and professional work.

ACKNOWLEDGEMENTS

I would like to thank Erik Frøkjær and Kasper Hornbæk for helpful comments and for encouraging me to publish the results of this investigation. I would also like to thank the anonymous reviewers of the original paper for drawing my attention to parts of the paper that needed further work.

REFERENCES

ACM 1998, *The ACM Electronic Guide to Computing Literature*, ACM, April 1998.

Barrie J. M. & D. E. Presti 1996, The World Wide Web as an Instructional Tool, in *Science*, October 18 vol. 274 pp. 371–372.

Brake, D. 1997, Lost in Cyberspace, in *New Scientist*, June 28, vol. 154.

Bush, Vannevar 1945 As We May Think in *The Atlantic Monthly*, vol. 176, no. 1 July 1995, pp. 101–108.

Chakrabarti, S. 1999 M. van den Berg & Byron Dom Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery *in Proceedings of the Eighth International World Wide Web Conference*. <URL: http://www8.org/w8-papers/5a-search-query/ crawling/>, Toronto 1999.

Kahle, Brewster, 1997, Preserving the Internet, in *Scientific American*, March 1997, pp. 72–73.

Lawrence, S. 1998 & C. L. Giles, Searching the World Wide Web in *Science*, April 3 vol. 280 pp. 98–100.

Perstrup, K 1997 E. Frøkjær, M. Konstantinovitz, T. Konstantinovitz & J. Varming A World Wide Web-based HCI-library Designed for Interaction Studies *in Proceedings of the third ERCIM Workshop on User Interfaces for All*, INRIA Lorraine, 1997, pp. 137–142.

Yahoo 1998 *Home > Science > Computer Science > Institutes* <URL: http://www.yahoo.com/ Science/Computer_Science/ Institutes>, June 1998.

Table 6 The number of papers in this study compared to the number of papers in the ACM Electronic Guide to Computing Literature.

	Unpub-lished	Jan-Jun 1998	1997	1996	1995	1994	Before 1994	All groups
Our study	165	98	250	260	264	304	1381	2722
ACM guide	–	–	–	18763	16786	21927	225736	283212
% studied	–	–	–	*1,4%*	*1,6%*	*1,4%*	*0,6%*	*1,0%*