DOI 10.1007/s10115-003-0112-y
Springer-Verlag London Ltd. © 2004
Knowledge and Information Systems (2004) 6: 230–241

Knowledge and
Information Systems

Discovering and Analyzing World Wide Web Collections[1]

Sougata Mukherjea

Verity Inc., Sunnyvale, CA, USA

Abstract. With the explosive growth of the World Wide Web, it is becoming increasingly difficult for users to discover Web pages that are relevant to a topic. To address this problem we are developing a system that allows the collection and analysis of Web pages related to a particular topic. In this paper we present the system's overall architecture and introduce the focused crawler used by the system. We also discuss the various techniques we use to allow the user to analyze and gain useful insights about a collection. Finally, we present some statistics on the collections.

Keywords: Authorities; Focused crawling; Graph algorithms; Hubs; Site graph analysis

1. Introduction

The World Wide Web (WWW) is undoubtedly the best source for getting information on any topic. Therefore, more and more people use the Web for *topic management* (Amento et al, 1999): the task of gathering, evaluating and organizing information resources on the Web. Users may investigate topics both for professional or personal interests.

Generally the popular portals or search engines like Yahoo and Google are used for gathering information on the WWW. However, they are not suitable for users who use the Web to gather detailed information on a topic of interest, for of a number of reasons:

- The explosive growth of the Web poses basic limits of scale for today's generic search engines. Thus much relevant information may not have been gathered and some information may not be up to date. For serious understanding of a topic having the latest information is essential.
- Most queries generate a large number of results and the search engines show these results as pages of scrolled lists. Going through these pages to retrieve the relevant

[1] Some sections of this paper have appeared in the Proceedings of the WWW9 Conference (in Amsterdam) and the Hypertext 2000 Conference (in San Antonio).

Received 8 Mar 2002
Revised 5 Aug 2002
Accepted 10 Sep 2002

information is tedious. More sophisticated techniques for organizing and analyzing the information are needed.

- The search engines retrieve documents containing the user-specified keywords. Some relevant documents may not be retrieved because they may not contain that specific word. For example, if a user is interested in *Wireless Technology*, searching using those words won't retrieve pages that have information on *WAP* or *Bluetooth* even though they are very relevant. Therefore, a more intelligent way of gathering the Web pages of interest is needed.
- Most corporate intranets have a large amount of information that is of interest. These cannot be collected by the search engines because they are under the firewall. It will be of great help to corporate users if both the intranet and Internet information can be gathered, organized and presented in a single unified view.

Because of these problems, there has been much awareness recently that for serious Web users focused *portholes* are more useful than generic portals (Chakrabarti et al, 1998). Therefore, systems that allow the user to collect and organize the information related to a particular topic and allows easy navigation through this information space is becoming essential.

We have built a *specialized search engine (SSE)* for the collection and organization of information on the Web related to a particular topic. The system uses a focused crawler to gather relevant pages. It also allows the user various ways to analyze the collection. This paper discusses the various features of the system. The next section cites related work and Section 3 describes the architecture of the system. Section 4 explains our focused crawler, which collects information from the WWW about a particular topic. Section 5 discusses the various analysis techniques that are present in the system to allow the user to gain useful insights about the information space. In Section 6 we present some statistics on the collections. Finally, Section 7 concludes the paper.

2. Related Work

2.1. World Wide Web Topic Management

In recent times there has been much interest in collecting Web pages related to a particular topic. Focused crawlers for collecting topic-specific Web pages are presented in Hersovici et al (1998), Chakrabarti et al (1999) and Aggarwal et al (2001). Our focused crawler is similar to these systems. However, we have incorporated some heuristics to improve performance. These are explained in Section 4.

Mapuccino (formerly WebCutter) (Maarek and Shaul, 1997; Hersovici et al, 1998; Ben-Shaul et al, 1999) and TopicShop (Terveen and Will, 1998; Amento et al, 1999) are two systems that have been developed for WWW topic management. Both systems use a crawler for collecting Web pages related to a topic and use various types of visualization to allow the user to navigate through the resultant information space. While Mapuccino presents the information as a collection of Web pages, TopicShop presents the information as a collection of Web sites. We believe that it is more effective to present the information at various levels of abstraction depending on the user's focus.

2.2. Analyzing Web Page Collections

Modifications of the traditional information retrieval clustering techniques are being applied to the WWW. For example, Pirolli et al (1996) and Pitkow and Pirolli (1997) describe research at Xerox Parc that uses clustering to extract useful structures from the Web. Similarly, in this paper we use various criteria to group Web pages for a particular topic at various levels of abstraction.

In an attempt to impose some structure on the WWW, Kleinberg (1998) defined two types of Web pages which pertain to a certain topic: authority and hub. Authority pages are authorities on a topic and therefore have many links pointing to them. On the other hand, hubs are resource lists: they do not directly contain information about the topic, but rather point to many authoritative sites.

Kleinberg's algorithm has been refined in CLEVER (Chakrabarti et al, 1998) and Topic Distillation (Bharat and Henzinger, 1998). Both of these algorithms augment Kleinberg's link analysis with textual analysis. On the other hand, Flake et al (2002) identifies Web communities based purely on connectivity. A slightly different approach to find hubs and authorities is SALSA (Lempel and Moran, 2000). A good overview of various link analysis techniques to find hubs and authorities and suggestions for improvements are presented in Borodin et al (2001). We believe that these algorithms are very important for a specialized search engine. We determine the hub and authority sites for a topic as well as the hub and authority pages. We have also made a minor modification to remove a drawback of the previous algorithms. This is explained in Section 5.

3. System Architecture

The specialized search engine for topic management consists of two main components:

- *Focused crawler*. The focused crawler collects pages from the WWW that are relevant to a particular topic. Various information about the collected pages is stored in an Oracle database. The pages are then indexed by a text search engine to enable users to search the collection using keywords. The crawler is discussed in the next section.
- *Collection analyzer*. The analyzer is the run-time component of the system that allows the search and analysis of a collection. It is an *Enterprise Java Beans* (http://java.sun.com/products/ejb/) running inside a *BEA Weblogic Application Server* (http://www.bea.com/products/weblogic/server/). The analyzer is discussed in Section 5.

Generally a user will access the system as a stand-alone application. The system can also run as a *portlet* in a portal that is developed using *Weblogic Portal* (http://www.bea.com/products/weblogic/portal). The user interface of both the stand-alone application and the portlet is developed using *Java Server Pages* which accesses the Analyzer EJB. Sometimes it would be helpful if other enterprise applications could be integrated with SSE. To enable this integration easily, SSE can also be run as a Web service. (Weblogic has tools that enable EJBs to run as Web services.) Thus other applications can integrate with SSE by communicating using *Simple Object Access Protocol (SOAP)* messages.

4. Focused Crawling

4.1. Basic Focused Crawling Technique

For collecting WWW pages related to a particular topic the user has first to specify some *seed URLs* relevant to the topic. These URLs may be from the corporate intranet as well, so that the resultant collection may have relevant pages from both the Internet and the intranet. The crawler can also issue a request with user-specified query terms to a popular Web search engine and augment the seed URLs with the search engine results. The SSE crawler downloads the seed URLs and creates a *representative document vector (RDV)* based on the frequently occurring keywords in these URLs. For best results the user should augment the RDV with some keywords that are very relevant to the collection.

The crawler then downloads the pages that are referenced from the seed URLs and calculates their similarity to the RDV using the vector space model. If the similarity is above a threshold, the pages are added to the collection and the links from the pages are added to a queue. The crawler continues to follow the out-links until the queue is empty or a user-specified limit is reached. The crawler also determines the pages pointing to the seed URLs. (A query *link:u* to search engines like AltaVista and Google returns all pages pointing to URL *u*). These pages are downloaded and if their similarity to the RDV is greater than a threshold, they are indexed and the URLs pointing to these pages are added to the queue. The crawler continues to follow the in-links until the queue is empty or a user-specified limit is reached.

After crawling, the collection consists of the seed URLs as well as all pages similar to these seed URLs that have paths to or from the seeds. Web pages that are not retrieved by a Web search engine because they don't have the user-specified query term may be retrieved by the focused crawler if they have paths to or from the seed URLs. Therefore, we believe that this collection is a good source of information available on the WWW for the user-specified topic.

4.2. Using Heuristics to Improve Performance

The main bottleneck of a crawler is the time spent in downloading Web pages. Besides network congestion, a crawler needs to follow the convention of issuing one download request to a site per 30 seconds. Therefore, downloading many pages from a single site may take a long time. For a focused crawler only Web pages relevant to the topic are important. So many of the downloaded pages may have to be discarded. In fact using our basic focused crawler for topics as diverse as *World Cup Soccer*, *Information Visualization* and *Titanic* we found that less than 50% of the downloaded pages were relevant.

If we could determine that a page will be irrelevant without examining the contents of the page, we could avoid downloading the page, thereby improving performance. We analyzed the crawler logs to figure out whether there was a particular type of page that was found to be irrelevant after downloading. Based on the log analysis we developed two heuristics to improve crawler performance.

4.2.1. Nearness of the Current Page to the Linked Page

If a Web site has information related to several topics, a page in the Web site for one of the topics may have links to pages relating to the other topics or to the main page of the site for ease of navigation. For example, the page http://www.discovery.com/guides/history/titanic/Titanic/titanic.html, a page relevant to *Titanic*, has a link to http://dsc.discovery.com/, the main page of the Discovery channel, a page not related to the topic. However, since most Web sites are well organized, pages that are dissimilar do not occur near each other in the directory hierarchy.

Therefore, when we are examining the pages linked to or from a Web page, we need to download the linked page only if it is near the current page. The determination of nearness will be an optimization between the number of irrelevant downloads and the number of relevant pages that are not downloaded. If we use a strict criterion to determine the nearness between two pages, the number of downloads will be lower, but we may miss some relevant pages. On the other hand, a lenient criterion to determine

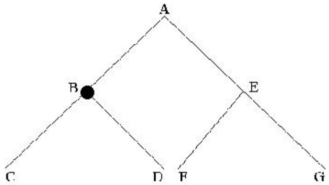


Fig. 1. Determining the nearness of two pages in the directory hierarchy.

nearness will retrieve all the relevant pages but at the cost of increasing the number of downloads.

Figure 1 shows how the SSE crawler determines nearness. Suppose a page in the directory *A/B* is the current page. Then pages in the same Web site are considered to be near (and therefore downloaded) if and only if they belong to the directories shown in the figure. Thus pages in the parent directory (*A*) as well as any children directories (*C,D*) are considered to be near. Sections of sibling directories (*E,F,G*) are also downloaded. After crawling several Web sites, we found that this definition of nearness gives the best result. It should be noted that if a page has a link to or from a page in another Web site, we have to download the page (unless it is in the stop URL list). Also note that if a URL contains any of the topic keywords, the page will always be downloaded. So all pages from http://www.titanicmovie.com will be downloaded for the *Titanic* collection.

4.2.2. Irrelevant Directories

Because Web sites are well organized, generally most pages in the same directory have similar themes. Thus all pages in http://www.murthy.com/txlaw/ talk about tax laws. One of these pages, http://www.murthy.com/txlaw/txwomsoc.html, was retrieved by a query to Google with the keywords *'World Cup soccer'* since it talked about visa issuance to the Women's World Cup Soccer tournament. However, none of the other pages of the directory are relevant to the collection on *World Cup soccer*. However, in the basic crawler all these pages were downloaded, only to be discarded after determining the similarity to RDV.

To avoid this problem, during crawling, we keep a count on the number of pages that have been indexed and ignored for each directory. If more than *25* pages of a directory are downloaded and *90%* of those pages are ignored, we do not download any more pages from that directory.

4.2.3. Evaluation

Table 1 shows the comparison between the basic crawler and the enhanced crawlers for three collections: *Information visualization*, *World Cup soccer* and *Titanic*. The following statistics are shown:

- *% Download* = 100 * (Number of pages downloaded by enhanced crawler)/ (Number of pages downloaded by basic crawler). For all three collections there is a significant decrease in the number of URLs that were downloaded. Besides reducing the network overhead, the amount of time needed to wait between successive requests to the same server is also decreased.
- *% Nearness* is the number of pages not downloaded because the linked page was not near the original page. A significant number of pages were not downloaded using this heuristic.

Table 1. Effectiveness of the heuristics in improving crawling performance

Collections	Information visualization	World Cup soccer	Titanic
% Download	73.54	66.27	77.4
% Nearness	21.9	27.59	15.7
% Rejected directories	4.56	6.14	6.9
% Relevant pages missed	4.26	4.04	2.56
Average score pages missed	0.261	0.279	0.254

- *% Rejected directories* is the number of pages not downloaded because a significant number of pages from the directory the page belonged to were already rejected after downloading. Unfortunately, only a small number of pages were ignored using this criterion. Maybe extending the criterion to determine irrelevant sites instead of irrelevant directories may be more useful.
- *% Relevant pages missed* = 100 * (Number of relevant pages indexed by the basic crawler but ignored by the enhanced crawler)/(Total number of pages indexed by the basic crawler). This statistics shows the number of relevant pages that were ignored by the enhanced crawler. It indicates that most of the pages that were not downloaded by the enhanced crawler were not relevant to the topic and were not indexed by the base crawler also.
- *Average score pages missed* is the average score of the relevant pages indexed by the basic crawler but ignored by the enhanced crawler. Even if the enhanced crawler misses some relevant pages, for the crawler to be acceptable none of these pages should be very important to the collection. To measure the importance of the missed pages, we calculate the average similarity scores of these pages. In our crawler all pages whose similarity to the representative document vector is greater than 0.25 are indexed (we use a vector space model where the similarity of a page is a number between 0 and 1). Since the average score of the missed pages is close to 0.25, it shows that these pages were not the most important pages of the collection.

Thus, our enhancements were able to significantly reduce the download time without missing many relevant pages of the collection.

5. Analyzing a Collection

The specialized search engine allows the user to search a collection. Other techniques of analysis are also provided.

5.1. Viewing Information at Various Levels of Abstraction

For most topics the crawler will be downloading thousands of Web pages. To allow the user to understand the information space better, the system groups the pages together at various levels of abstraction. Figure 2 shows the abstraction hierarchy. We use the following techniques to form the hierarchy:

1. *Website directory structure.* The Web pages are considered to be the leaves of the hierarchy. Most Web sites organize the pages into a meaningful directory hierarchy. Thus www.cnn.com/WORLD contains world news and www.cnn.com/SHOWBIZ contains entertainment news. Further, www.cnn.com/ SHOWBIZ/Music and www.cnn.com/SHOWBIZ/Movies contain news about music and movies respectively. Therefore, we use the directory structure of a particular Web site to organize the pages at

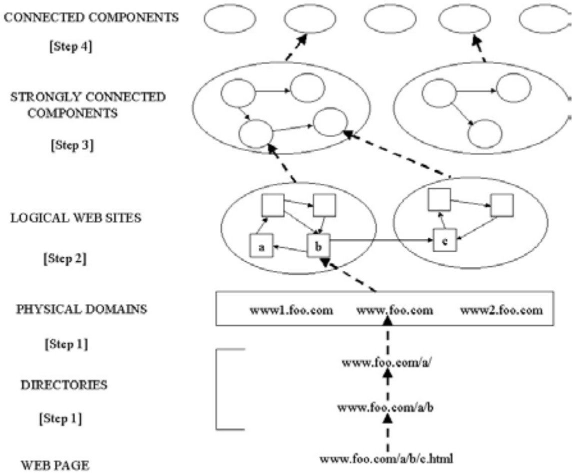


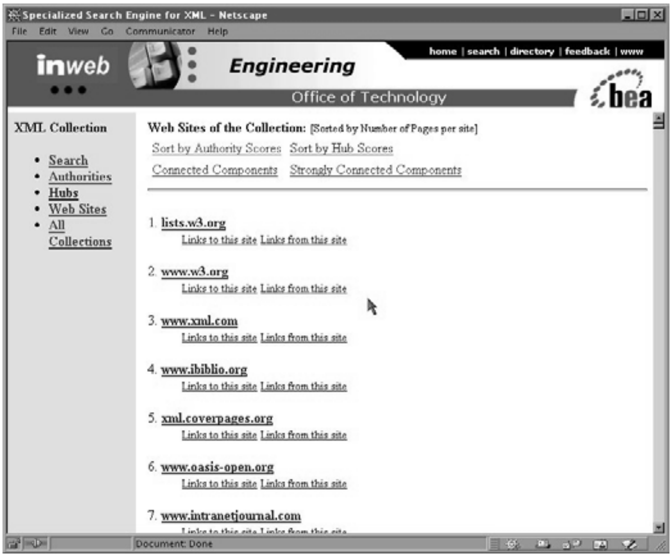
Fig. 2. Abstraction hierarchy of a collection. The disconnected connected components are at the top and the Web pages are the leaves.

the lowest level. At the end of this process the pages are grouped into their respective physical domains like www.cnn.com.

2. *Logical Web sites.* Many large corporations use several Web servers based on functionality or geographic location. For example, www.nec.com and www.nec.co.jp are the Web sites of NEC Corporation in the USA and Japan respectively. Similarly, shopping.yahoo.com is the shopping component of the Yahoo www.yahoo.com portal. Therefore, we group together related physical domains into logical Web sites at the next level. The physical domains are grouped into logical Web sites as follows:

   - We first analyze the domain's URL. The URL is separated into dot-separated tokens. Thus www.nec.co.jp is split into *www*, *nec*, *co* and *jp*. We then determine the main token among these tokens using some simple heuristics. In our example, *nec* is the main token. Physical domains that have the same main token are grouped into a logical Web site. This process will group www.nec.com and www.nec.co.jp together as well as shopping.yahoo.com and www.yahoo.com.
   - Just analyzing the URLs will not group together the domains cnn.com and cnnfn.com since they have different main tokens. For this we have to analyze the IP addresses. The addresses are also separated into dot-separated tokens. If the first three tokens of two IP addresses are the same, we can assume the domains belong to the same corporation. For example, the IP addresses of both cnn.com and cnnfn.com start with 207.25.171. So we can assume that they are in the same logical Web site. However, note that just analyzing the IP addresses is not sufficient to group geographically distributed Web sites of the same corporation. For example, www.nec.com and www.nec.co.jp have very different IP addresses.

Most readers generally want to know what is available on a given site, not on a single page. Therefore, logical Web sites are a basic unit of information in the specialized search engine. The user can start from the Web sites of a collection and navigate up and down the abstraction hierarchy. Figure 3 shows the logical Web sites for a collection on *XML*. They are sorted by the number of pages per site. The figure shows that site lists.w3.org has many relevant pages for XML. It is also possible to sort by the authority or the hub scores of the site.


Specialized Search Engine for XML - Netscape
File Edit View Go Communicator Help
home | search | directory | feedback | www
inweb
Engineering
Office of Technology
hea
Document: Done

XML Collection

- Search
- Authorities
- Hubs
- Web Sites
- All Collections

Web Sites of the Collection: [Sorted by Number of Pages per site]

Sort by Authority Scores Sort by Hub Scores

Connected Components Strongly Connected Components

1. lists.w3.org
   Links to this site Links from this site
2. www.w3.org
   Links to this site Links from this site
3. www.xml.com
   Links to this site Links from this site
4. www.ibiblio.org
   Links to this site Links from this site
5. xml.coverpages.org
   Links to this site Links from this site
6. www.oasis-open.org
   Links to this site Links from this site
7. www.intranetjournal.com
   Links to this site Links from this site

Fig. 3. The logical Web sites for a collection on *XML*. They are sorted by the number of pages per site.

3. *Strongly connected components.* The above two steps create a collection of Web sites for a particular topic. For many topics there will be hundreds of Web sites; therefore further abstraction will be useful. To determine the related Web sites, we create a site graph. This is a directed graph with the Web sites as the nodes. If a page in Web site *a* has a link to a page in Web site *b*, then an edge is created from node *a* to *b* in the site graph. We then calculate the *strongly connected components* in the site graph. Since each pair of nodes of a strongly connected component is reachable from each other, these nodes can be considered to be related. Thus they can be grouped together. Note that the strongly connected component is represented by the site with the highest authority score inside the component.
4. *Connected components.* The procedure to determine the strongly connected components creates a *component graph*. The nodes of the graph are the strongly connected components that were discovered. If there was an edge between any two Web sites belonging to different strongly connected components, then there will be an edge between the corresponding nodes in the component graph. At the final stage of forming the abstraction hierarchy we consider the component graph as an undirected graph and determine the connected components. Nodes in the same connected component form a cluster.

Figure 2 explains how the logical Web sites are grouped into strongly connected and connected components. Web sites in the same strongly connected component have bidirectional paths among each other; for example, sites *a* and *b*. On the other hand, sites in the same connected component have unidirectional paths only; for example, in Fig. 2 there is a path from *b* to *c* but not *c* to *b*. Thus, sites in the same strongly connected component can be considered to be more similar than sites in the same connected component. Obviously, sites belonging to different connected components have no links between them.

It should be emphasized that the procedure to form the abstraction hierarchy is linear and thus applicable to collections with a large number of Web pages. The directory structure and the logical Web sites can be determined by an analysis of the URLs and IP addresses. On the other hand, the algorithms for finding the strongly connected and connected components are $O(n + e)$, where n and e are the number of nodes and edges in the graph.

5.2. Hubs and Authorities

To determine the hubs and authorities we have implemented a modified version of Kleinberg's original algorithm (Kleinberg, 1998) incorporating the improvements suggested in Chakrabarti et al (1998) and (Bharat and Henzinger, 1998). This algorithm was used to determine both the hub and authority pages as well as the hub and authority sites. (To find the sites the algorithm was applied on the site graph.)

5.2.1. The TKC Effect

As mentioned in Lempel and Moran (2000), one problem with Kleinberg's algorithm and all algorithms derived from it is the *tightly knit community (TKC) effect*. A tightly knit community is a small but highly interconnected set of sites. Roughly speaking, the TKC effect occurs when such a community scores high in link-analyzing algorithms, even though the sites in the TKC are not authoritative on the topic, or pertain to just one aspect of the topic. As an example, consider a collection C which contains the following two communities: a community y, with a small number of hubs and authorities, in which every hub points to most of the authorities; and a much larger community z, in which each hub points to a smaller part of the authorities. The topic covered by z is the dominant topic of the collection, and is probably of wider interest on the WWW. Since there are many z-authoritative sites, the hubs do not link to all of them, whereas the smaller y community is densely interconnected. The TKC effect occurs when the sites of y are ranked higher than those of z.

Our study indicates that the TKC effect mostly occurs when a large corporation has many physical domains and pages in these domains point to each other. Although the previous algorithms ignore links between pages in a domain, they do not ignore links between the different physical domains within a logical Web site. For example, Lempel and Moran (2000) found that the top authorities found by Kleinberg's algorithm for a collection on *movies* were all from the go.msn.com logical Web site. Similarly, most of the top authorities for a collection on *Java* was part of the EARTHWEB Inc. network.

In our system, the links between pages in a domain, as well as the links between the different physical domains within a logical Web site, are ignored. Therefore, pages that have many links to/from other pages in the same logical Web site will not have a high hub/authority score. As a result, the proposed algorithm does not suffer from the TKC effect in most cases. Note that to avoid the TKC effect Lempel and Moran (2000) proposed the SALSA approach to identify hubs and authorities. However, this approach suffered from another deficiency as highlighted in Borodin et al (2001). Our algorithm avoids this problem as well as the TKC effect.

Another advantage of identifying logical Web sites is evident when we determine the hub and authority sites. If we just identified the hub and authority domains many domains within the same logical site may have higher scores because of the TKC effect. This is shown in Table 5.2.1. The top 10 authority physical

domains of the *wireless technology* collection all belong to the Mobile Lifestreams corporation (their IP addresses start with 195.82.119) and have high scores just because they link extensively to each other. It is obvious that the logical Web sites are more meaningful. The logical domain for Mobile Lifestream is not even a top authority since

Table 2. The authority physical domains and logical Web sites for a collection on *wireless technology*

Authority physical domains	Authority logical Web sites
www.mobile4mobile.com	www.nokia.com
www.links2mobile.com	www.ericsson.com
www.mobilemms.com	www.wapforum.com
www.mobileems.com	www.digimob.com
www.yes2wap.com	www.palowireless.com
www.mobilesms.com	www.wirelessinanutshell.com
www.mobileplanners.com	www.cellular.co.za
www.mobileserviceeng.com	www.bluetooth.com
www.wirelessclueless.com	www.nttdocomo.com
www.mobileretailers.com	www.w3schools.com

Table 3. The number of logical Web sites, strongly connected and connected components for different collections

Collections	Intelligent agents	Web service	Wireless technology	XML
Logical Web sites	616	428	884	1688
Strongly connected components	486	319	749	1382
Connected components	61	7	18	27

other sites do not point it to extensively! Note that if we considered links between the physical domains within the Mobile Lifestreams logical Web site, pages from that corporation will be the major hubs and authorities of the wireless collection and important pages from other Web sites will have lower scores.

6. Features of a Collection

We have developed various collections on topics both of professional interest (for example *peer to peer*, *Web services* and *wireless technology*) as well as personal interest (for example, *Titanic* and *World Cup soccer*). These collections have given us some insights about the structure of the collection information space.

Table 5.2.1 shows the number of logical Web sites, strongly connected components and connected components that were discovered for four collections: *intelligent agents*, *Web service*, *wireless technology* and *XML*. It shows that the number of strongly connected components is not much smaller than the number of logical sites for all three collections. It seems to indicate that there are not many cyclical referencing among the Web sites; therefore they could not be grouped into strongly connected components. However, Table 5.2.1 also indicates that the number of connected components is small. Thus most of the Web sites in a collection have paths to or from other sites.

Table 6 shows the percentage of total pages and sites of the collection in the strongly connected and connected components with the most pages. An interesting observation is that most of the sites and pages of a collection are grouped into the one connected component. For example, for the *XML* collection as many as 99.54% of the pages and 99.29% of the sites are in one connected component. Obviously, this component accounts for the hubs and authorities of the collection. Most of the other components have one or two sites that are not very important to the collection. Table 6 also shows that a majority of sites and pages are also grouped into one strongly connected component. For example, for the Web service collection, a strongly connected component contains 73.38% of the pages and 24.77% of the sites. The main connected component of all

Table 4. The Percentage of Pages and Sites in the main Strongly Connected and Connected Components for different collections

Collections	Intelligent agents	Web service	Wireless technology	XML
% of pages in main SCC	47.43	73.38	62.76	72.03
% of sites in main SCC	20.45	24.77	15.27	18.31
% of pages in main CC	79.07	99.27	99.3	99.54
% of sites in main CC	91.23	97.92	97.74	99.29

collections consists of the main strongly connected component as well as many other strongly connected components, which are very small, containing one or two sites. These sites have links to or from sites in the main strongly connected component. The above observations allow us to determine the approximate graph structure of a collection. Each collection has a strongly connected component containing a large number of sites. Some sites have links to sites in the strongly connected component and some sites have links from that component. Together they form a large connected component. There are also some smaller connected components not connected with the main component. Thus the graph structure of a collection is similar to the *bow-tie structure* of the overall Web as determined in Broder et al (2000).

7. Conclusion

In this paper, we have presented a specialized search engine for Web topic management. SSE can be used as a stand-alone application, a portlet or a Web service. The system uses a crawler to gather information related to a particular topic from the WWW. The enhancements to the crawler have resulted in improved performance by reducing the number of pages that need to be downloaded by more than 20%, while missing only a few insignificant pages. SSE allows the user various ways to analyze a collection. For example, it allows the user to examine the information at various levels of abstraction. We have developed a technique of grouping physical domains into more meaningful logical Web sites. Our algorithm to determine hubs and authorities ignores links between the logical Web sites. Because of this, the algorithm does not suffer from the tightly knit community effect, a problem common in Kleinberg's algorithm. In this paper we have also tried to determine the graph structure of a collection. As the WWW grows bigger, systems like SSE will become essential for retrieving useful information.

References

Aggarwal CC, Al-Garawi F, Yu PS (2001) Intelligent crawling on the World Wide Web with arbitrary predicates. In Proceedings of the 10th international World-Wide Web conference, Hong Kong, May 2001, pp 96–105

Amento B, Hill W, Terveen L, Hix D, Ju P (1999) An empirical evaluation of user interfaces for topic management of web sites. In Proceedings of the ACM SIGCHI '99 conference on human factors in computing systems, Pittsburgh, PA, May 1999, pp 552–559

Ben-Shaul I, Hersovici M, Jacovi M, Maarek YS, Pelleg D, Shtalheim M, Soroka V, Ur S (1999) Adding support for dynamic and focussed search with Fetuccino. In Proceedings of the 8th international World-Wide Web conference, Toronto, Canada, May 1999, pp 575–588

Bharat K, Henzinger M (1998) Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the ACM SIGIR '98 conference on research and development in information retrieval, Melbourne, Australia, August 1998, pp 104–111

Borodin A, Roberts GO, Rosenthal JS, Tsaparas P (2001) Finding authorities and hubs from link structures on the World Wide Web. In Proceedings of the 10th international World-Wide Web conference, Hong Kong, May 2001, pp 415–429

Broder AZ, Ravi Kumar S, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener JL (2000) Graph structure in the web. In Proceedings of the 9th international World-Wide Web conference, Amsterdam, Netherlands, May 2000, pp 309–320

Chakrabarti S, Dom B, Gibson D, Kleinberg J, Raghavan P, Rajagopalan S (1998) Automatic resource compilation by analyzing hyperlink structure and associated text. Computer Networks and ISDN Systems (special issue on the 7th international World-Wide Web conference, Brisbane, Australia) 30(1–7):65–74

Chakrabarti S, van den Berg M, Dom B (1999) Focussed crawling: a new approach to topic-specific web resource discovery. In Proceedings of the 8th international World-Wide Web conference, Toronto, Canada, May 1999, pp 545–562

Flake GW, Lawrence S, Giles CL, Coetzee FM (2002) Self-organization and identification of web communities. IEEE Computer 35(3):66–71

Hersovici M, Jacovi M, Maarek YS, Pelleg D, Shtalheim M, Ur S (1998) The Shark-Search algorithm: an application: tailored web site mapping. Computer Networks and ISDN Systems (special issue on the 7th international World-Wide Web conference, Brisbane, Australia) 30(1–7):317–326

Kleinberg JM (1998) Authorative sources in a hyperlinked environment. In Proceedings of the 9th ACM-SIAM symposium on discrete algorithms, May 1998

Lempel R, Moran S (2000) The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In Proceedings of the 9th international World-Wide Web conference, Amsterdam, Netherlands, May 2000, pp 387–401

Maarek Y, Shaul IZB (1997) WebCutter: a system for dynamic and tailorable site mapping. In Proceedings of the 6th international World-Wide Web conference, Santa Clara, CA, April 1997, pp 713–722

Pirolli P, Pitkow J, Rao R (1996) Silk from a sow's ear: extracting usable structures from the Web. In Proceedings of the ACM SIGCHI '96 conference on human factors in computing systems, Vancouver, Canada, April 1996, pp 118–125

Pitkow J, Pirolli P (1997) Life, death and lawfulness on the electronic frontier. In Proceedings of the ACM SIGCHI '97 conference on human factors in computing systems, Atlanta, GA, March 1997, pp 383–390

Terveen L, Will H (1998) Finding and visualizing inter-site clan graphs. In Proceedings of the ACM SIGCHI '98 conference on human factors in computing Systems, Los Angeles, CA, April 1998, pp 448–455

Correspondence and offprint requests to: Sougata Mukherjea, IBM India Research Lab, New Delhi, India. Email: smukherj@in.ibm.com