Accepted for presentation at International Symposium on Research, Development & Practice in Digital Library, Nov. 18-21, 1997, Tsukuba Science City, Japan

Searching for Parts and Services on the Web

Quan Zhu, Fangqi Hu, Ke-Thia Yao, Peter Will

Information Sciences Institute, University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292. USA

Abstract

The World Wide Web is a rich repository of information. However, the Web is primarily designed for human users, and maintains the majority of its information in textual form. This makes reliable and efficient ways of retrieving and extracting that information from the Web difficult. Textual information has some structure, but it is not readily accessible through traditional database techniques. Its enormous size and exponential growth make the problem even more difficult. This paper describes a method that uses statistical correlation analysis techniques in order to discover useful information by recognizing certain patterns present in the Web documents. In particular, we apply our method to electronic commerce[1] to discover taxonomies of parts and services, and to discover companies providing these parts and services.

***Keywords**:* Internet, WWW, Electronic Commerce, Search Engine, Agent, Information Retrieval, Data Mining, Resource Discovery, Pattern Recognition, Correlation Analysis.

1 Introduction

Digital libraries are not just large, online collections of documents with search and indexing capabilities. Digital libraries should be customizable, *task-oriented* information repositories. This means that digital libraries should provide tools which help work with information in ways that assist their tasks and goals. For example, users do not necessarily come to a library with the goal of retrieving specific information. Rather, their goal is often to find out where they should be focusing attention. This is important in intelligence analysis, both military and commercial, where one needs to review large bodies of documents from a standpoint of identifying significant organizations, activities, events, etc. It is also of importance in electronic commerce, where one needs to identify things such as potential sources of supply or potential customers.

Thus, in many different applications, one critical generic task facing users is that of making sense of a space of documents. This paper describes one technique that can assist in that task, which makes use of statistical correlation analyses to extract information about a set of documents. We do not claim that this technique is sufficient in and of itself, but argue that it is useful and that it is potentially promising as part of a *suite* of tools for investigating information spaces. We will discuss other components of that suite briefly in the conclusion of this paper.

Not surprisingly, the techniques described in this paper are oriented toward documents published on the World-Wide Web (WWW). There is little doubt by now that the WWW is on its way to becoming the largest information database ever created. It is safe to say that in the not so distant future, most things that can be converted into digital format will be found on the WWW. The popularity of the WWW is partly due to the fact that there is little restriction on how the Web documents should be structured or formatted. But this feature of the WWW has also made the automatic processing of information much more difficult, especially as the amount of online information is experiencing its exponential growth. Current generation of commercial web tools provide a useful start for searching and organizing the Web, but they are not enough.

In addition, as we have mentioned, the users want the digital library to provide value-added, task-oriented information to the users. When users access a digital library to retrieve documents, they are not interested in the document per se, but in the information that is contained in the documents in support of the information analysis task at hand.

[1] In many respects the challenges and issues facing electronic commerce are similar to that of the digital library domains, see [1] for more discussion.

A useful practical instance of this kind of task arises in electronic commerce. The examples that will be used throughout this paper are drawn from that domain, and were explored under the auspices of a research project at ISI on digital libraries and electronic commerce, called DASHER [5], which is concerned with federated Web applications and information repositories relating to parts and services. One objective of the DASHER research project has been to improve upon the success of an earlier project, the ISI FAST Electronic Broker System [6], by developing extensions to the existing FAST *sourcing* and *quoting* services. From a digital library perspective, this is part of an attack upon the generic information space analysis problem; electronic commerce is just a useful special case.

2 Information Space Analysis: Problems

Consider the problems of making sense of the WWW information space on product information today. When users want to locate information on certain product parts, such as who makes them, what applications the parts are used in, etc., they can use one of the search engines to perform a keyword search using their knowledge of this product. However, as anyone who has used one of these commercial search engines can testify, search results returned using just one or two keywords often comprise hundreds of thousands of replies. The need to use more keywords to narrow the search calls for more intimate knowledge from the users who initiate the search. Frequently, it is unclear how to refine a search. For example, how would users refine an initial search on the keyword, *battery,* to include just companies manufacturing batteries?

Another approach adopted by Internet search companies, like Yahoo™[7], is to provide a directory-like service to the WWW to help users locate relevant information quickly. In a sense, the setup of the Yahoo directory provides Internet users with a well-structured database with indexed information. Potentially a directory service can create a directory for each category of goods. However, the Yahoo information database requires a team of human indexers. This is labor intensive and therefore, cannot keep up with the dynamic and constantly changing nature of what is available on the WWW. Moreover, because contents of schemes like Yahoo's are limited to those who choose to submit, they are potentially incomplete. For example, in Yahoo's Accelerator directory under the Physics subtree, the world's largest accelerator, CERN, is missing, although other accelerators, like Fermi, are present.

Another issue is that Yahoo introduces a bias that can be inappropriate for many purposes, because its approach necessarily treats companies equally in listing them. For example, in the Yahoo's "PC hardware" directory there are hundreds of entries. An user unfamiliar with the personal computer industry has no way to differentiate a major company from a minor one.

Recent developments in *search agents* have produced tools to automate well-defined Web browsing processes. These agents go directly to pre-specified Web pages, fill out the various forms if necessary, and return the collected information to the user. However, most of these agents [2, 3, 4] either require *a priori* knowledge of the layout of Web sites or make regularity assumptions about the way Web sites present information, both of which limit their ability to deal with new sites.

For example, the LiveAgent [4] product of AgentSoft provides a macro creation mechanism that allows the user to record mouse clicks and key presses. Then, the LiveAgents use these macros to act as surrogates for the user on the Web to retrieve information. Such agents are unable to search sites they have not previously been shown. The Shopbot [2] system is more flexible. It is able to extract price information of products from Web sites that it has not seen, but the class of Web sites it is able to extract information is narrow and the system makes relatively strong assumptions on the how vendors list their product items.

In general, these agents are useful in relieving human readers of repetitive browsing tasks, and they do provide a value-added service that traditional Internet search engines do not. However, the scalability of these search agents is potentially limited by their over-reliance on the structural layout of Web pages. Most of the information found on the Web is unstructured text.

An alternative approach to this search process is to use well-structured product information Web sites such as PartNet™ [8], Thomas Publishing[9], or IndustryNet™ [10]. To learn more about the sub-categories of products, a user can refer to these manufacturing catalog and engineering handbook sites. However, this approach introduces two other problems. First, another level of knowledge is needed about the location of these catalogs (whether online or in printed form) and/or the names of the handbooks. There is still a need for a "one-stop" search engine that integrates across the catalogs. Such an engine would take an initial simple product inquiry from the user, discover on its own more detailed information related to this product category, ask the user to narrow the search using the new information, and repeat this process, if necessary, to further narrow the search. The second problem is that, in general, these are commercial services each with its own proprietary information hierarchy structure. Companies pay a fee to be listed on these services. These services have similar shortcomings as Yahoo's Internet Directory, in the sense that they often do not reflect the dynamic nature of the Web, they are labor intensive to set up and maintain, and their information could be subject to biases unavoidably introduced by the nature of the approach.

To sum up, the World Wide Web is a rich information repository, but the current search tools do not do an adequate job of accessing it. The Web keyword search engines do not provide enough guidance to adequately focus the search to narrow the results to a manageable relevant subset. The general Web directories and specialized product directories are labor intensive to maintain, and not likely to keep pace with the growth of the Web. They are often incomplete, because they only contain information of those who chose to submit or to pay. They introduce inappropriate biases by not ranking the information according to its relevance. Search agents are limited by the dependence of their information extractors on structured Web pages. Most Web pages are unstructured.

What users need is a tool suite that is able to provide focused information to support the specific information analysis task they are trying to address. In order for such tools to be successful, they must be general, fast and user friendly. They must be applicable to a wide range of Web sites and not be hand-tailored to extract fixed and static information. The Web browsing is interactive by design. Users tend not to tolerate tools that take a long time to complete. Finally, the tool should not require the users to have detailed knowledge of the information space to refine the search.

3 Approach

In the following sections we discuss a statistical correlation analysis technique that uses samples of Web pages randomly collected from the WWW to help users construct a detailed view of certain categories, such as product or service information. It does not rely on any *a priori* knowledge or regularity assumptions about Web sites. The result of the search shows an expanded view of the category, which can be further expanded by repeating the search using any particular sub-category returned from the previous search.

Our view of searching the Web for relevant information is that one should treat the entire World Wide Web as the source for knowledge and information in the form of a textual database. This does not mean that it is necessary to visit every Web page in order to convince oneself that relevant information is not missed. In most cases, information relating to various commercial products and service is often described and mentioned in multiple places on the Web. This is simply because there are many different users of the same product, and more than one company may produce a certain product. Therefore, one only needs to *sample* a part of the "Web space" in order to locate enough relevant information for most cases. Since most Web information is in natural language text, the more difficult issue here is how to extract only the relevant information for a product that a user is interested from these Web pages. Another issue of importance is how to organize the search results so that the more relevant information is presented first.

Our approach can be divided into three steps: (1) information retrieval, (2) information extraction, and (3) information mining. The purpose of step one is to focus the attention of the search to a "semantically coherent" subset of Web pages. Web contains enormous amounts of information on a multitude of subjects. A word that has a particular meaning in one subject may have a completely different meaning in another. For example, the word "gear" in the mechanical engineering domain refers to a class of mechanical part typically characterized by its diameter and its number of teeth. But, in the fishing domain "gear" refers to fishing equipment, such as rods and reels. To properly mine semantic information from the Web pages, the Web pages must refer to approximately the same subject.

In our current implementation, the traditional keyword-based approach is used to perform the information retrieval step. We use existing commercial Web search engines to gather Web pages that simply contain the set of keywords that describe a subject that we are interested in. This is a somewhat simplistic approach to gathering Web pages that refer to approximately similar subject. However, it is interesting to observe that it provides a good enough sample of Web documents to pass on to the next step.

The next step is to extract appropriate information from the natural language text of the retrieved Web pages. In our case the information to extract is the taxonomy of a particular part and the companies that sell that part. Each Web page needs to be parsed and analyzed to extract categories of the taxonomy and names of companies.

Finally, in the last step the extracted information is treated as a statistical sample gathered from a population. Then, statistical correlation analysis is performed to filter and rank the extracted information to determine the relevant categories and companies.

Since there is no particular order to the Web documents that are returned from a Web indexing engine (e.g., when searching for “batteries," one has no reason to expect that the indexing engine will return battery-related Web pages from Japan before or after those from the US), therefore we assume Poisson statistics can be applied for sampling. Poisson statistics deals with events that are not highly correlated, which in our case applies to the sampling of loosely or uncorrelated Web pages. If a search finds a particular result N times, then the error of the result is:

$$\sigma = \frac{\sqrt{N}}{N}$$

Taking into account this error, then statistically speaking, e.g., a result that appears in our search 100 times (uncertainty is 10%) is very likely to be more statistically significant than a result that appears only 50 times (uncertainty is 14%). This conclusion is not likely to be reversed even if more Web pages are analyzed. Thus, from a given sample, we can obtain reasonably reliable relative rankings of the frequencies of reference for given terms. This statistical significance does not guarantee practical significance, nor can such an algorithm infer the reason for relative rankings. We are not making such claims. We do, however, claim that (particularly in concert with other analysis tools) such rankings provide a very useful heuristic for focusing initial attention and stimulating further analysis.

The first two steps of our approach would initially appear to be quite difficult to implement. Indeed, they would be difficult, *if* we had to perform them perfectly. That would require a deep semantic understanding of subject area being searched, and sophisticated natural language understanding capabilities. However, because of the highly redundant nature of the Web and the statistical correlation done in the final step, we can afford to use relatively simple techniques. Details of the method are best illustrated in the following two examples, where we perform two specific type of searches.

4 Example Applications

As a first example, we show how this technique applies to extracting information about product subcategories from document spaces formed by queries about the category. As a second set of examples, we show how the same technique helps a user locate relevant companies that produce or market certain products. The result of the search produces a ranked list of companies, which serves as a guide to investigating candidate companies in this product category. Companies that are listed higher, having engendered more references are more likely to to have some characteristic worth investigating than those that are listed lower.

4.1 Search for Types of Parts

Our first application is a Web search method that discovers the “ontology” of product categories, or, in other words, the subcategories of product under certain categories, by analyzing Web pages that contain the keyword description of the product. This helps users who want to learn more about a product, but know only the general category name of the product, but not the detailed description. For example, consider a user interested in learning more about the subject “battery," and wanting to use the WWW to find out how many kinds of batteries are out there.

Following our three step approach, we first retrieve a set of Web pages that may contain battery category information. We piggyback on existing search engines to collect Web pages containing the **seed keywords**, “battery” or “batteries.” In this case the usage of the word, battery, on the Web seems to be mostly consistent. The Web pages typically use the word, battery, to refer to a cell carrying an electric charge. Any usages that are not semantically consistent are treated as noise, and they are removed in step three. In the event the Web uses the word, “battery,” inconsistently, then the keyword search may be augmented with additional words to focus the search, such as the words, voltage and electricity. The additional words may inadvertently rule out useful Web pages. But, because of the highly redundant nature of the Web, there are typically enough useful Web pages left for the extraction step.

To extract categories from the Web pages, we concentrate on sentences that contain the seed keyword, batteries. Words that appear to be descriptive are extracted as candidates for category headings. Candidates for descriptive words are identified as words between the seed keyword and a common stop word such as *of, for, if, which, use, take*, etc., which appeared earlier in the same sentence. Beginning of the sentence can also be used as a boundary if no stop word were found. This information extraction method is highly approximate. It may reject true descriptions and admit spurious ones. Again, we rely upon the redundancy of the Web to provide enough true descriptions, and statistical analysis methods to reject the false positives.

In step three, statistical analysis is applied to the distribution and correlation of these descriptive words to produce a ranked list of the most probable descriptions of sub-categories for the product. Since the number of types of products under certain product category is usually quite limited, one is only interested in perhaps the top 20 (or less) most used descriptive keywords. Poisson statistics [11] can be applied here to judge whether enough statistical information has been collected to construct a meaningful view of the most often used descriptive keywords. The Poisson statistic maybe computed dynamically as the descriptive words are extracted. This allows the information retrieval and extraction processes to terminate as soon as enough statistics has been gathered.

Figure-1 shows the search results using this method for searching *battery* types. The horizontal axis is the *frequency*, or number of times a particular type of battery was found during the search. We want to emphasize that the results shown in Figure-1 come directly from the search program, which was given only one input parameter: the keyword, “battery.” No manual or semi-manual filtering has been applied to the results. “Noise results," or any highly unlikely product types, are listed, if there are any. However, as one sees in Figure-1, the top returns from this search method are all highly relevant results.

Additional examples are shown in Figure-2 & 3. In Figure-2, we show the search results for types of *motor*, and in Figure-3, results are listed for types of *capacitor.*

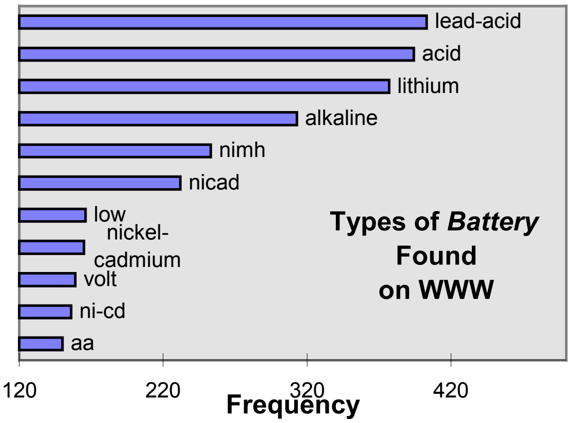

Figure-1: Frequency distribution of descriptive words appearing before the word *battery* found in Web pages randomly collected. Notice that 10 of 11 are relevant product categories of batteries.

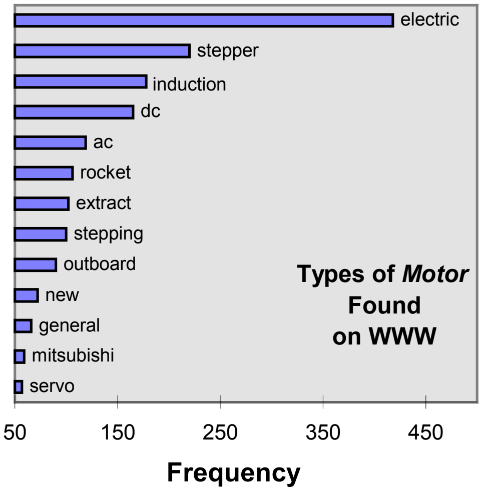

Figure-2: Frequency distribution of descriptive words appearing before the word *motor* found in Web pages randomly collected. Notice that 9 of 13 are relevant product categories of motors, and 2 out 13 refer to names for manufacturers of motor vehicles.

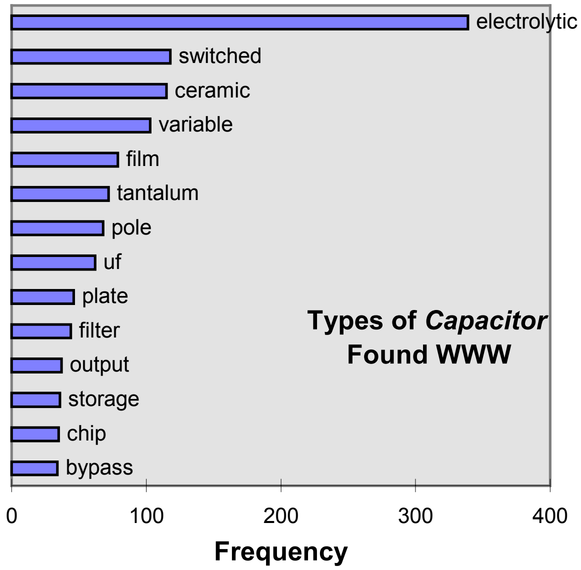

Figure-3: Frequency distribution of descriptive words appearing before the word *capacitor* found in Web pages randomly collected. Notice that all top results are relevant product categories of capacitors.

4.2 Search for Manufacturers and Services

Our second application learns the names of the commercial companies that make certain products, by identifying company names in Web pages that also mention the names of the products. These companies could be original manufacturers of certain products, or they could be retailers of products of a certain category. Here the retailers can be viewed as services who perform the tasks of locating the various manufacturers for certain product categories, taking the risk (sometimes) of paying for the products before selling them, or provide other value-added services many original manufacturers would otherwise unable to provide. The methodology of this Web search is similar to the one mentioned above, except this time one looks for keywords that resemble the description of company names. Also, the search is not limited to the vicinity of the product name, but the entire Web document.

Again, this method is based on the observation that the association of company names and the product names is highly redundant on the Web. In other words, even though company names that have nothing to do with certain products might be mentioned in some Web pages, overall, the top names should dominate over irrelevant ones if one scans through enough Web documents. Also, one can afford to ignore names that do not end with business types such as "Inc.", "Incorporated," "LLP," etc. because of this redundancy. We believe this is the only efficient approach since otherwise one would have to rely on a huge lookup table that lists all the company names in the world, and large number of keyword combinations would need to checked against this lookup table, thus making the search highly inefficient.

Figure-4 shows the results of this search for companies that are associated with *microprocessors*. Again, we want to point out that the final results listed in the Figure-4 are the direct results of the search; they were not checked against any special databases to ensure that these are real company names. Still, as we see in Figure-4, it turns out that this was not needed: the results not just show *real* companies that are associated with microprocessors, but most of them are highly important manufacturers.

More examples of this search are shown in Figure-5 through Figure-7 for *batteries, DRAM,* and *Memory Products.* It is interesting to note that the *DRAM* search resulted in mostly memory manufacturers, whereas *Memory Products* search resulted in mostly memory retailers.

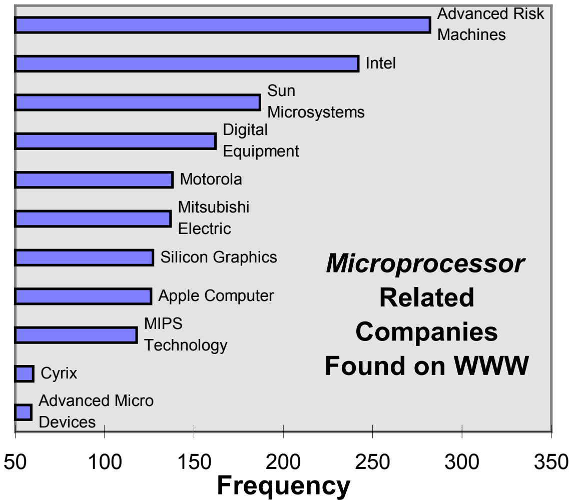

Figure-4: Companies related to microprocessors found on the WWW. Results are obtained after analyzing Web pages randomly collected that contained the word “microprocessor."

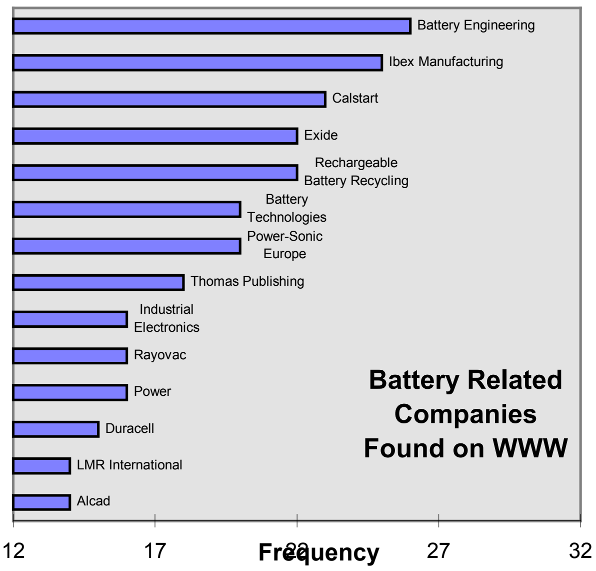

Figure-5: Companies related to batteries found on the WWW. Results are obtained after analyzing Web pages randomly collected that contained the word “battery."

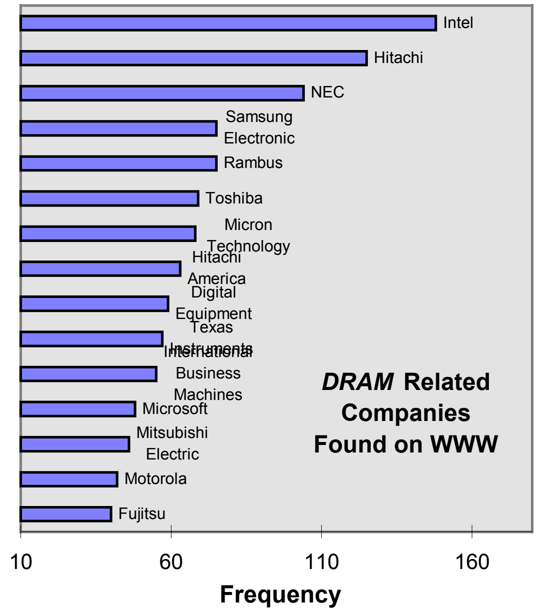

Figure-6: Companies related to DRAM found on the WWW. Results are obtained after analyzing Web pages randomly collected that contained the word “DRAM."

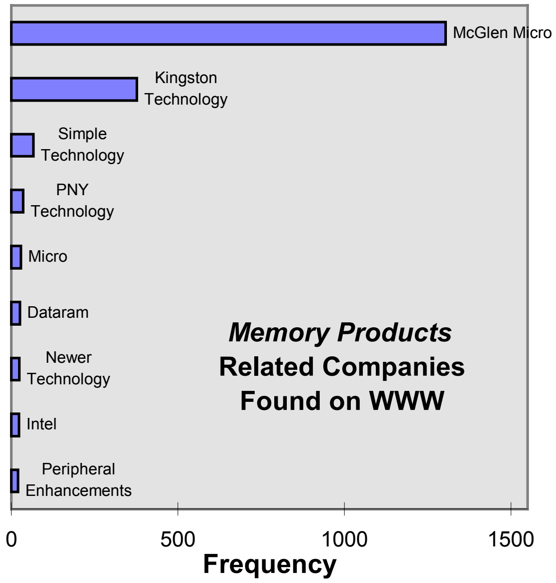

Figure-7: Companies related to memory products found on the WWW. Results are obtained after analyzing Web pages randomly collected that contained the word “memory products."

5 Discussion

The World Wide Web is a rich repository of information. However, automatic extraction of knowledge from this repository is hampered by the fact that most of the information is stored in natural language text. In this paper, we described an automatic method that is able to extract ontological taxonomies of parts (and, by extension, services), and help identify companies related to these parts and services.

This method is general in several ways. It does not rely on deep semantics and natural language understanding capabilities, which tend to be difficult to incorporate and fragile. To the extent that it relies on natural language processing capabilities at all, it makes use of techniques for text extraction (a simpler problem than text understanding); these have received a great deal of attention in communities focused upon message process and are becoming increasingly robust and efficient. The heart of our approach relies upon pattern recognition and statistical correlation techniques, which tend to be more robust, less sensitive to noise or irrelevant results, and tend to scale better. It does not rely on the structure or layout of the Web pages, so it is applicable to a much wider class of Web pages.

We have demonstrated a range of applicability of this method by using it to determine parts taxonomies and to find relevant companies for the parts. The method has no knowledge of parts *per se*; therefore, it does not make any assumptions on the particular part or type of parts that is being searched.

The search method described in this paper has been implemented in multi-threaded JAVA code. Each thread is responsible for scanning for patterns from one URL. The search query runs average about 20 minutes on a SUN Ultra. There are a couple of ways to speed up the query runs. First, because of the multi-threadedness of the code, the runs are CPU bound. With better tuning of the code and a good JAVA compiler, we believe dramatic speedups are achievable. Second, we have started harvesting Web pages to store and index them locally, a technique used by commercial search engine to improve performance. With locally cached Web pages, each query typically returns within a few seconds.

A common complaint about Web keyword search engines is that they are too difficult to use. Specifying one or two keywords is often not enough. Typically, the users need to provide additional terms to narrow the focus of the search. In contrast, our method is able to provide useful information just based on the one or two keywords. The method uses information extraction techniques to relieve the users of having to scan through all the documents. Then, it uses statistical correlation analysis to rank the extracted information by its relevance.

The effectiveness of our searches demonstrates the possibility that additional applications (not necessarily specific to electronic commerce) based on our method can be designed to perform a variety of other highly focused Web searches. Furthermore, these specialized applications can be made to interact with each other, and be grouped together to form a Web search interface that can facilitate the kind of "meta query" that most traditional library users are familiar with. An analysis environment can be setup that supports "add-ons" of modules that perform various other types feature extractions such as geographical locations, dates, financial figures, etc.

There are a number of extensions that are required to go from the specific results described in this paper to a generically-useful suite of tools for analyzing information spaces. High on our list is developing capabilities for recursively analyzing findings to support a progressive search. For example, when "McGlen Micro" was found to be by far the most frequently-cited company associated with memory products (see Figure-7), one would be interested in finding out what are the primary memory products associated with this company, and why it is so frequently cited. (As it turns out, McGlen's notebook computer memory is usually the cheapest.)

Another need that we see are capabilities for characterizing the search space. For example, it would be desirable to have tools that can partition a set of retrieved documents according to multiple taxonomies, both the ones our methods infer and others, and can display those partitionings together with frequency information of the sort we have described in this paper. This would help users understand what kinds of documents were available to them and how to prioritize their examination of those documents. Such capabilities can be further augmented by tools that can record and organize findings (a very simple example being bookmark lists that users can structure in outline and/or sub-topic format according to the issues that come up in their analysis of the information space). There is a closely related need for tools that help users in planning investigations (e.g., multiple question searches). The work described in this paper represents an initial step toward this larger suite of capabilities that are needed to help digital library users make sense of an information space.

6 References

[1] Nabil Adam and Yelena Yesha. Electronic Commerce and Digital Libraries: Towards a Digital Agora. *ACM Computing Surveys* **28**(4), December 1996.

[2] Robert Doorenbos, Oren Etzioni, and Daniel Weld, A Scalable Comparison-Shopping Agent for the World-Wide Web. In *Proceedings of Autonomous Agents*, 1997.

[3] Erik Selberg and Oren Etzioni. Multi-Service Search and Comparison Using the MetaCrawler, In *Proceedings of the 1995 World Wide Web Conference.*

[4] Bruce Krulwich. Automating the Internet: Agents as User Surrogates. *IEEE Internet Computing*, Vol. 1, No. 4, July/August 1997.

[5] David Benjamin, David Grossman, Paul Postel, Curt Powley, Peter Will, Ernesto Brodersohn, and Rupal Fadia. Federated Services in Electronic Commerce: Architecture and Issues. ISI Research

Report, ISI-RR-96448, July 1996. Submitted to IEEE Expert special issue on Electronic Commerce.

[6] Craig Milo Rogers, Anna-Lena Neches and Paul Postel. Finding What You Want to Buy Using the Web, In WWW Fall Conference, 1994. URL=http://info.broker.isi.edu

[7] *Yahoo, Inc*., <http://www.yahoo.com>.

[8] *PartNet (NTEC Information Systems),* <http://www.part.net/>.

[9] *Thomas Publishing Company,* <http://www.thomasregister.com/>.

[10] *Industry.net (Nets, Inc.),* <http://www.industry.net>.

[11] W. Feller. *Probability Theory and Its Applications*, vols. 1 & 2. New York: John Wiley, 1971.