Automated Identification of Web Communities for Business Intelligence Analysis

Michael Chau[1], Boby Shiu[2], Ivy Chan[3], Hsinchun Chen[4]

[1]School of Business, The University of Hong Kong, Pokfulam, Hong Kong
mchau@business.hku.hk,
[2]School of Business, The University of Hong Kong, Pokfulam, Hong Kong
gdpbskw@graduate.hku.hk
[3] Department of Decision Sciences and Managerial Economics, The Chinese University of Hong Kong, New Territories, Hong Kong
ivychan@baf.msmail.cuhk.edu.hk
[4]Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA
hchen@eller.arizona.edu

Abstract

Analysts often search the Web for business intelligence using traditional search engines which provide keyword-based search. Recently, it has been suggested that the incoming links, or backlinks, of a company's Web site can provide useful information about the company's "Web communities". Backlinks refer to other Web pages which have a hyperlink pointing to the company of interest and these pages form a cyber community on the Web. Analysis of these communities can provide useful signals for a company or information about its stakeholder groups, but the manual analysis process can be very time-consuming for business analysts and consultants. In this study, we report the design and evaluation of a tool called Redips that integrates automatic backlink meta-searching and text mining techniques to facilitate users in identifying such cyber communities on the Web for business intelligence purposes. The system architecture of the tool is presented and an experimental study was reported. The experiment results showed that Redips performed significantly better than two benchmark methods, namely backlink search engines and manual browsing.

Keywords: web communities, business intelligence analysis, backlink search, Google, knowledge management, data mining

1. Introduction

Seeing the importance of formulation of competitive strategy to gain competitive advantage, large corporations often engage in strategic planning process that assesses both the internal and external situation to formulate strategy. One of the major steps in the strategic planning process is environment scanning (Bradford et al. 1999). Given the information from the environmental scan, the firm should match its strengths to the opportunities that it has identified, while addressing its weaknesses and external threats.

The ways that analysts find information about a firm's environment are diversified. Traditionally, analysts may manually read the published company reports and other kinds of printed information. Many managers and analysts also use analysis tools to monitor a firm's external environments to obtain information relevant to its decision-making process and strategy planning process (Gilad and Gilad 1988). Recent years have seen the tremendous growth of the Internet and many resources and information are now accessible on the Internet and it has become a major source of business intelligence information.

On the Internet, the specific environment of a firm is indeed its Web communities (Kumar et al. 1998; Reid 2003). The identification of Web communities is important in the business intelligence analysis process. However, the huge size of the Web has made this a difficult task. It is simply impossible for a person to manually browse the entire Web to identify the Web communities of a firm. This has been known as the information overload problem. Search engines have been helping people in searching information on the Internet. However, oftentimes a significant portion of the Web pages returned by search engines are irrelevant or outdated and analysts still have to spend a long time to manually browse the content of each Web page, acquire the overall concept of the set of the search results, and summarize the information. Facing numerous “most relevant” Web pages, the process of identification of Web communities certainly become a time-consuming and mentally exhausting task to complete.

Personal Web spiders – robots that help users search for information on the Web – also have been developed for business intelligence analysis (Chau and Chen 2003; Chen and Chau, 2004). For example, the CI Spider can automate the filtering of irrelevant Web pages and facilitate the analysis of retrieval results using Web searching and clustering techniques (Chau et al. 2001; Chen et al. 2002). While these tools can help users with general analysis (e.g., studying the core business of a firm), they are often limited to the outgoing links of an organization or even just the Web pages within the organization’s Web sites. To study the Web communities of a given organization, it is important to identify and analyze the backlinks (incoming links) of the organization’s Web site. However, no business analysis tools have such backlink search capability.

We try to address these problems using our “Redips” architecture. “Redips” is the reverse spelling of the word “Spider”. Instead of using breadth-first search or best-first search like traditional Web spiders (Chen et al. 2002), Redips is designed to search the Internet “backwards”. When a user inputs the URL of a company into Redips, the tool will search the Internet backwards such that the search results will represent the firm’s environment and the implicit communities of the specified URL. The backlink search results will be fetched real-time to the local computer and Redips will examine the fetched Web pages and perform text analysis and text mining to extract the noun phrases from the stored Web pages. Noun phrases symbolize a vector of themes and topics in the Web pages that analysts can use to easily identify the main areas of interest in the Web communities. Lastly, Redips allows users to visualize the retrieved Web pages in the form of a two-dimensional map using the self-organizing maps (SOMs) technique. The map would help analysts and managers to quickly understand the themes in the set of fetched Web pages and shorten the time of reading the Web pages one by one and summarizing the information.

The remaining of the paper is organized as follows: Section 2 reviews related work in Web communities, business intelligence analysis, and Internet-based analysis tools. Section 3 describes the research questions and the problem of the existing analysis tools. Section 4 outlines the architecture of our analysis tool Redips. Section 5 discusses an experiment conducted to evaluate the proposed tool and the corresponding experiment design. Section 6 presents the experiment results and analyzes the results using statistical analysis. Section 7 concludes our work and discusses our future research directions.

2. Research Background

2.1. Business Intelligence Analysis and Web Communities

The Internet has many well-known explicitly defined communities – groups of individuals who share a common interest, together with the Web pages most popular amongst them (Reid 2003). The Web communities consist of the following stakeholders of the firm: customers, suppliers, competitors, regulators, employees, educational institutions, court and legal institutions, financial institutions, stockholders, public-interest groups, labor unions, political parties, federal, state, local governments, etc. (Schermerhorn 2001). The stakeholders listed here can be classified into two categories: explicit and implicit Web communities.

Explicit communities are the communities that are available to be identified easily on the Internet. Kumar et al. discussed the example of an explicit community of Web users interested in Porsche Boxster cars, such as the Porsche newsgroup, or resource collections in directories in search engines, such as the Yahoo directory (Kumar et al. 1998). Explicit communities are easy to be identified and analysts can simply use manual method to find a firm's explicit communities by browsing the firm's newsgroup, or the category in which the firm fall into in the directory like Yahoo on the Internet.

Implicit communities are relatively more difficult to find using manual browsing method. According to Kumar, implicit communities refer to the distributed, ad-hoc and random content-creation related to the common interests on the Internet (Kumar et al. 1998). The pages often have links to each other, but the common interests of implicit communities are sometimes too narrow and detailed for the resource pages or the directories to develop explicit listings for them. As a result, it is more difficult to find the implicit communities of a firm. In identifying the explicit and implicit communities of a firm, it is often assumed that the content pages created by these communities would provide hypertext links back to the firm's homepage for reference (Reid 2003).

2.2. Internet-based Analysis Tools

The simplest Internet-based analysis tool may be just a Web browser like the Internet Explorer. Using a manual browsing method, an analyst only needs to enter a firm's competitor's URL in the browser and then manually browse the information for further analysis. This manual browsing method is common to analysts. It is simple as many people are experienced in Internet surfing nowadays. Manual browsing also ensures the quality of the information collected and alleviates the problem of garbage in, garbage out, thus improving the quality of knowledge discovered.

However, the process of manual browsing is very time-consuming and mentally exhausting. Data collection is the most time-consuming task in typical analysis projects, accounting for more than 30% of the total time spent (Prescott and Smith 1991). It is not practical for analysts to go through the Web sites of all stakeholders of a company in detail. To make the problem worse, many Web pages are updated weekly, daily or even hourly. It is almost impossible for analysts to manually collect the most updated versions of every Web page for analysis.

To address these problems, Web analysis tools have been developed to do more than simple browsing. In this section, we will review some Web-based analysis tools that are related to Web searching and in particular Web community extraction.

2.2.1. General-purpose Search Engines and Backlink Search Engines

Many different search engines are available on the Internet. Each has its own characteristics and employs its preferred algorithm in indexing, ranking and visualizing Web documents. For example, Google (www.google.com) and AltaVista (www.altavista.com) allow users to submit queries and present the Web pages in a ranked order, while Yahoo! (www.yahoo.com) groups Web sites into categories, creating a hierarchical directory of a subset of the Internet. A Web search engine usually consists of the four main components: spiders, indexer, retrieval and ranking, and user interface. Spiders are responsible for collecting documents from the Web using different graph search algorithms. The indexer creates indexes for Web pages and stores the indices into database. The retrieval and ranking module is used for retrieving search results from the database and ranking the search results. The user interface allows users to query the search engine and customize their searches.

Another type of search engines is the meta-search engines, such as MetaCrawler (www.metacrawler.com) and Dogpile (www.dogpile.com). These search engines do not keep their own indexes. When a search request is received, a meta-search engine connects to multiple popular search engines and integrates the results returned by these search engines. As each search engine covers different portion of the Internet, meta-search engines are useful when the user needs to get as much of the Internet as possible.

In addition to general searching, analysts can also use “backlink searching” to research a firm’s Web communities that consist of the important stakeholders of the firm, including customers, suppliers, competitors, regulators, etc. Backlink searching can identify the communities of these stakeholders who generally have on their Web pages a hyperlink that point to the URL of the firm. Some general search engines also provide the feature of backlink searching. In these search engines, the indexer will, in addition to performing regular indexing, also index the links of each Web page collected. The information on these links is stored in the search engine’s database, so it is possible for users to search for all links that point to a given Web page. One example is the Google search engine. Google allows users to use the reserved word “link” as an operator in the query. The query “link:siteURL” shows the users pages that point to a given URL. For example, the query “link:www.google.com” will return pages that contain a hyperlink to Google’s home page. AltaVista and MSN Search (search.msn.com) also have a similar feature and a similar “link:” operator that finds pages with a link to a page with the specified URL text. Yahoo (www.yahoo.com), HotBot (www.hotbot.com), Alexa (www.alexa.com), and AlltheWeb (www.alltheweb.com) are other examples of search engines that provide backlink search.

Unlike general-purpose meta-search engines such as those discussed above, no meta-search engines are available for searching backlinks in the current search engine market.

2.2.2. Text mining Tools

Text mining, also known as text data mining (Hearst 1997) or knowledge discovery from textual databases (Feldman and Dagan 1995), refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents (Tan 1999). Text mining is as well an extension of data mining or knowledge discovery from structured databases (Fayyad 1996). Text mining is a fascinating multidisciplinary field, including the knowledge from information retrieval, textual information analysis, information extraction, and information clustering.

Text mining tools help analysts better understand the retrieved Web document set from the Internet, identify interesting Web documents more effectively, and gain a quick overview of the Web documents’ contents. This saves the manual browsing time of reading the entire set of Web pages. Analysts only have to examine the categories which are of the firm’s own interest.

As information on the Internet is mainly in the form of text, e.g. HTML hypertext documents or PDF documents, text mining and textual information analysis become popular in the literature of Internet-based analysis tools. Textual information analysis is mainly based on natural language processing and has to index the source Web documents for analysis. Many techniques of indexing the source documents and extracting key concepts from text have been proposed in the recent few years. One of the proven techniques is automatic indexing algorithm, which has been shown to be as effective as human indexing (Salton 1986). Automatic indexing algorithms can be based on either single words or phrases. The Arizona Noun Phraser (AZNP) is one example of phrase-based indexing tool (Tolle and Chen 2000). The tool extracts all the noun phrases from each Web documents, based on part-of-speech tagging and linguistic rules.

After the documents are indexed, further analysis like document classification and clustering can be applied. Document classification is one form of data analysis that can be used to categorize the documents into a predetermined set of document classes or concepts (Han and Kamber 2001). Web documents are categorized based on the predefined library science classification methods in this approach. Since the classes or concepts are provided, the classification step is also known as supervised learning. This contrasts with unsupervised learning (or clustering), in which the classes are not known, and the number or set of classes to be learned also may not be known in advance. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. In text mining, the classes or clusters would have category labels defined based on the keywords or phrases that appear in the Web documents collected. The facts that document clustering generates the categories automatically based on the documents make the category labels of clustering more specific, descriptive, and meaningful with respect to the cluster contents.

One of the clustering approaches is Kohonen self-organizing map (SOM). Self-organizing map classifies documents into various categories automatically determined during the clustering process, with the underlying neural network algorithm (Kohonen 1995). This approach clusters documents into various topics that are automatically generated in real time using neural network algorithms (Kohonen 1995; Chen et al. 1998). Every document is assigned to its corresponding regions in the two-dimensional graphical map displayed to the user. Every region contains similar documents under the same topic while those regions with similar topics are located close to each other on the self-organizing map.

3. Research Question

The huge size of Internet has been the source of many problems for business intelligence analysis and the current Internet-based tools help very little in searching the Web communities of an organization. In this research, we try to answer the research question on whether backlink search can be used to identify an organization's Web communities more efficiently and effectively when compared with other existing methods. We also study the usability of the proposed implementation.

4. Architecture of Redips

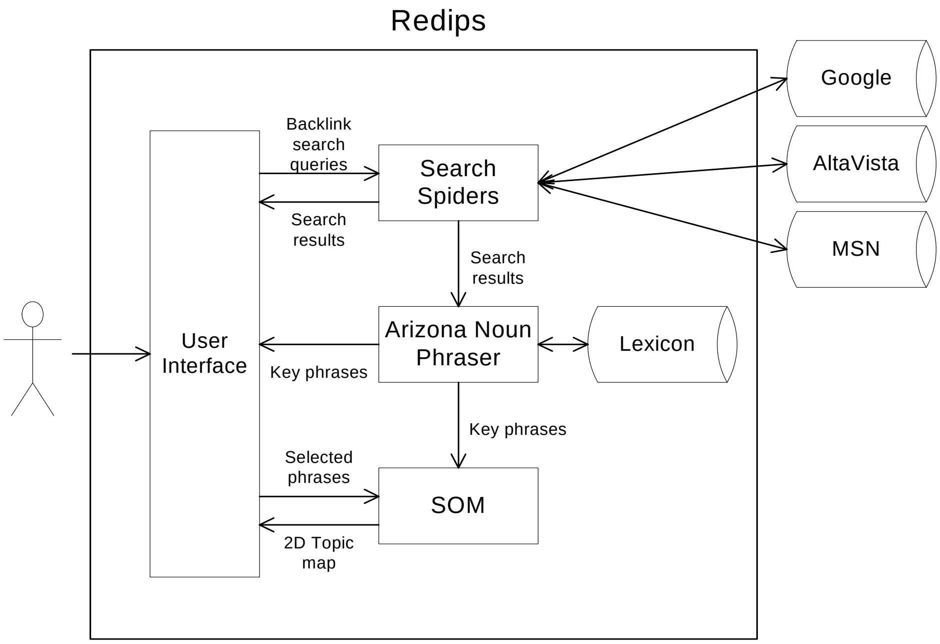


Figure 1: Redips System Architecture

To study the research question, we proposed the Redips architecture as shown in Figure 1. A user session with Redips starts with querying the selected search engines based on the given URLs. The spider then fetches the URLs returned from those Web pages. After collecting the required number of Web pages, further analysis will be performed. Noun phrases will be extracted from them, which allow the user to know what key concepts are related to the Web sites and keywords specified. The concepts can also be visualized in a two-dimensional map, which categorizes the Web pages by collecting them into regions, each of which represents a concept. All these functionalities allow the user automatically to collect information more effectively and represent it in a more meaningful way.

Redips is implemented based on the MetaSpider system developed in our previous research (Chen et al. 2001). The main modules include the user interface, spider, Arizona noun phraser, and self-organization maps. User interface is the first point of contact between the user and the system. Spider fetched the URLs returned from those search engines. Arizona noun phraser is a natural language processing tool to do the key phrase extraction from Internet text. Self-organization maps visualize the concepts in a two-dimensional map. The modules will be discussed in the subsections below.

4.1. Spider

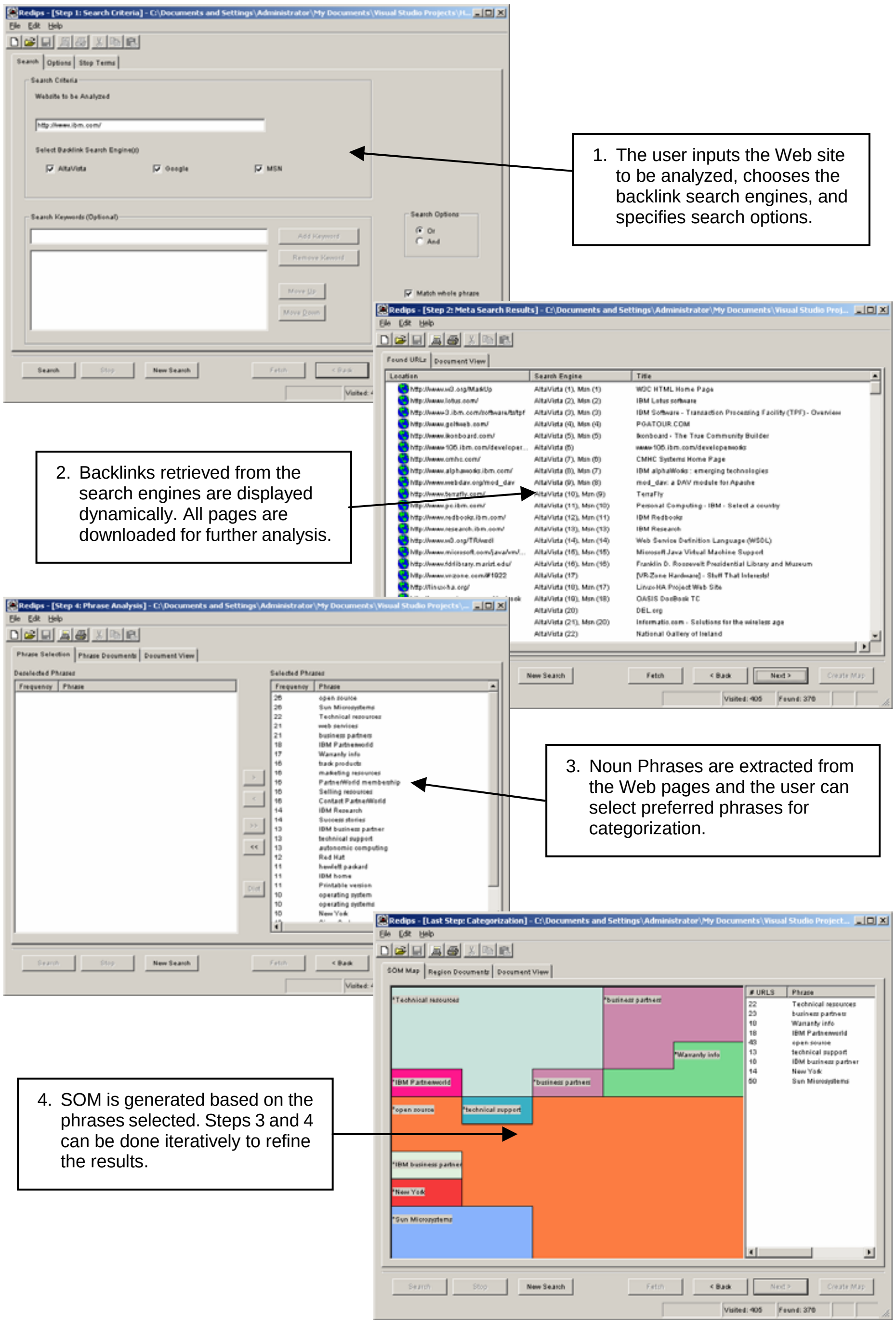
Redips has the ability of meta-searching, which leverages the capabilities of multiple backlink search engines and provides a simple, uniform user interface that alleviate the information overload and low precision issues (Selberg and Etzioni 1997; Chau et al. 2001). Meta-backlink-searching can improve search performance by sending queries to multiple backlink search engines and collating only the highest-ranking subset of the returns from each backlink search engine. Currently, Redips connects to three backlink search engines: Google Web APIs service, Altavista, and MSN Search. More backlink search engines may be easily added in our architecture. Unlike other meta-searching tools, which show only the URLs and page summaries to the user, Redips will fetch the full text of the URLs returned by the underlying backlink search engines and perform post-retrieval filtering and analysis.

4.2. Post-retrieval Analysis

After the documents are fetched from the Web, the Arizona Noun Phraser (AZNP), developed at the University of Arizona AI Lab, is executed to perform “noun-phrasing” by extracting high-quality phrases from the documents (Tolle and Chen 2000). The frequencies of occurrences of the phrases are also recorded. Arizona Noun Phraser helps analysts to evaluate of Web communities’ link in a short time and provides an overview of the entire document set to the user. In addition, the self-organization maps (SOM) algorithm is employed to automatically cluster the Web pages collected into different regions on a two-dimensional map. The map creates an intuitive, graphical display of important concepts contained in textural information (Lin et al., 1991). The SOM aims to visualize the pattern and relationship across Web documents that reveal the business relationships between the firm’s stakeholders. Compared with textual reports, SOM maps would be able to draw the attention of managers and analysts and allow them to quickly understand the overview of the Web communities identified. This would shorten the overall analysis time and the decision making time, which is very important in the today’s fast-changing business world. Readers are referred to Chen for the technical details of the two components (Chen et al. 2001).

4.3. Sample User Session

When using Redips, a user should first enter the Web site to be analyzed and the backlink search engines to be included. A sample user session with Redips is shown in Figure 2. In this example, the Web site entered is http://www.ibm.com/, the homepage of the IBM Web site. Optionally, the user can enter the keyword(s) to be included in the returned Web pages. The user may also specify some other search options. In this step, the user can define the intelligent analysis objectives, e.g. the firm, information source, topic, in the analysis process. After starting the search, multiple threads will be generated to start getting Web pages from the Internet. The URLs returned by the search engines will then be displayed. The user can browse these pages for exploratory, preliminary research in this step.


Redips - [Step 1: Search Criteria] - C:\Documents and Settings\Administrator\My Documents\Visual Studio Projects\H...
File Edit Help
Search
Options
Stop Terms
Add Keyword
Remove Keyword
Move Up
Move Down
Search
Stop
New Search
Fetch
< Back

Search Criteria

Website to be Analyzed: http://www.ibm.com/

Select Backlink Search Engine(s): ☑ AltaVista ☑ Google ☑ MSN

Search Keywords (Optional)

Search Options: ● Or ○ And

☑ Match whole phrase

1. The user inputs the Web site to be analyzed, chooses the backlink search engines, and specifies search options.

Redips - [Step 2: Meta Search Results] - C:\Documents and Settings\Administrator\My Documents\Visual Studio Proj...
File Edit Help
Found URLs
Document View
New Search
Fetch
< Back
Next >
Create Map

Location	Search Engine	Title
http://www.w3.org/MarkUp	AltaVista (1), Msn (1)	W3C HTML Home Page
http://www.lotus.com/	AltaVista (2), Msn (2)	IBM Lotus software
http://www-3.ibm.com/software/ts/tpf	AltaVista (3), Msn (3)	IBM Software - Transaction Processing Facility (TPF) - Overview
http://www.golfweb.com/	AltaVista (4), Msn (4)	PGATOUR.COM
http://www.ikonboard.com/	AltaVista (5), Msn (5)	Ikonboard - The True Community Builder
http://www-106.ibm.com/developer...	AltaVista (6)	www-106.ibm.com/developerworks
http://www.cmhc.com/	AltaVista (7), Msn (6)	CMHC Systems Home Page
http://www.alphaworks.ibm.com/	AltaVista (8), Msn (7)	IBM alphaWorks : emerging technologies
http://www.webdav.org/mod_dav	AltaVista (9), Msn (8)	mod_dav: a DAV module for Apache
http://www.terrafly.com/	AltaVista (10), Msn (9)	TerraFly
http://www.pc.ibm.com/	AltaVista (11), Msn (10)	Personal Computing - IBM - Select a country
http://www.redbooks.ibm.com/	AltaVista (12), Msn (11)	IBM Redbooks
http://www.research.ibm.com/	AltaVista (13), Msn (13)	IBM Research
http://www.w3.org/TR/wsdl	AltaVista (14), Msn (14)	Web Service Definition Language (WSDL)
http://www.microsoft.com/java/vm/...	AltaVista (15), Msn (15)	Microsoft Java Virtual Machine Support
http://www.fdrlibrary.marist.edu/	AltaVista (16), Msn (16)	Franklin D. Roosevelt Presidential Library and Museum
http://www.vrzone.com/#1922	AltaVista (17)	[VR-Zone Hardware] - Stuff That Interests!
http://linux-ha.org/	AltaVista (18), Msn (17)	Linux-HA Project Web Site
...ook	AltaVista (19), Msn (18)	OASIS DocBook TC
	AltaVista (20)	DEL.org
	AltaVista (21), Msn (20)	Informatio.com - Solutions for the wireless age
	AltaVista (22)	National Gallery of Ireland

Visited: 405 Found: 378

2. Backlinks retrieved from the search engines are displayed dynamically. All pages are downloaded for further analysis.

Redips - [Step 4: Phrase Analysis] - C:\Documents and Settings\Administrator\My Documents\Visual Studio Projects\...
File Edit Help
Phrase Selection
Phrase Documents
Document View
>
<
>>
<<
Plot
Search
Stop
New Search
Fetch
< Back

Deselected Phrases

Frequency	Phrase

Selected Phrases

Frequency	Phrase
26	open source
26	Sun Microsystems
22	Technical resources
21	web services
21	business partners
18	IBM Partnerworld
17	Warranty info
16	track products
16	marketing resources
16	PartnerWorld membership
16	Selling resources
16	Contact PartnerWorld
14	IBM Research
14	Success stories
13	IBM business partner
13	technical support
13	autonomic computing
12	Red Hat
11	hewlett packard
11	IBM home
11	Printable version
10	operating system
10	operating systems
10	New York

3. Noun Phrases are extracted from the Web pages and the user can select preferred phrases for categorization.

# URLS	Phrase
22	Technical resources
23	business partners
10	Warranty info
18	IBM Partnerworld
48	open source
13	technical support
10	IBM business partner
14	New York
50	Sun Microsystems

Visited: 405 Found: 378

4. SOM is generated based on the phrases selected. Steps 3 and 4 can be done iteratively to refine the results.

Figure 2: Example of a User Session with Redips

The user can then instruct the system to download the complete pages from the URLs returned by the search engines (only titles and URLs are available in the previous step). The search results are displayed dynamically during the search. The user can explore the results and browse the content of any of the Web pages collected. Noun phrases are also extracted from the Web pages and analyzed. The frequency of appearance of each noun phrase is displayed and the user can browse the pages that contain any particular noun phrase. A categorization map, known as the Self-Organizing Map (SOM), is generated based on the noun phrases selected. The Web pages are categorized into different regions on the map, based on the noun phrases they contain. In our example in Figure 2, several categories of Web communities of IBM are identified on the map, such as IBM Partnerworld, Sun Microsystems, open source, and technical resources.

5. Experiment Design

The research goals are to answering the research questions as described in Section 3 and in know if Redips outperformed other Internet-based analysis tools. We also plan to evaluate the effectiveness and efficiency of different methods in performing both document retrieval and document categorization tasks in the business intelligence analysis process, our project performed a comparative user study to contrast our tool with two traditional business intelligence analysis approaches: backlink search engines and manual browsing.

5.1. Experimental Tasks

Redips, designed to facilitate and integrate both document retrieval and document categorization, is not directly applicable to the traditional evaluation methodologies that treat document retrieval and document categorization completely separately. As a result our project uses the evaluation framework based on theme identification that enables us to measure the performance of the combination of the systems' retrieval and categorization features (Chen et al. 2001; Chau et al. 2001; Chau et al. 2003). The evaluation would ask the test subjects to identify the major themes related to the Web communities of a certain firm's Web site. This is similar to the open-ended "soft" queries created by the National Institute of Standards and Technology (NIST) for the TREC (Text Retrieval Conference) ad hoc tasks (Voorhees and Harman 1998).

The experiment tasks were constructed based on the new evaluation framework. A theme was defined as "A short noun phrase" that "summarizes a specific aspect of Web communities". Noun phrases like "consulting firms", "business intelligence", "java technology", "financial consulting" are the examples of themes in the experiment tasks. The theme-based framework enables us to evaluate traditional methods with Redips, measure the performance of the combination of the systems' retrieval and categorization features in relation to the Web communities. The performance measurements like precision and recall are further discussed in the Section 5.3.

The test subjects had to image themselves as a consultant who has been hired to do some research or investigation on the online Web communities of a certain company, e.g. IBM. The URL of the Web site of the company was presented to the subjects. The subjects had to start the Redips system and started the search by entering the URL in the user interface. Following the search session from meta-search results to noun phrase selection, a two-dimensional SOM map would be displayed to the subjects, providing an overview of categories of the Web communities. The subjects would summarize all the displayed information and write down a number of themes, which are related to the specific aspect of Web communities. The themes should give an overview about the Web communities of the firm. At the end of the experiment, subjects were required to fill out questionnaires about the user experience of the three analysis approaches.

The experiment tasks of backlink search engines and manual browsing approach are similar, except the analysis tools involved in the tasks. Test subjects, when they are using backlink search engines approach, were presented with three search engines: Google, Altavista, and MSN search. The search engines had to be used to collect an organization's Web communities' information. For the manual browsing and searching approach, the subjects were asked to freely explore the contents in the given Web site using an Internet browser for the Web communities' information. Both these two benchmark approaches also required the test subjects to write down a number of themes about the Web communities of the firm, similar to that of using the Redips approach.

5.2. Experiment Design and Hypotheses

The experiment intended to compare Redips's performance with the backlink search engine and manual browsing approach in business intelligence analysis based on Web communities. The following hypotheses were constructed in the experiment:

Hypothesis 1. Redips achieves a higher precision and recall than backlink search engines for searching Web communities of a firm.
Hypothesis 2. Redips achieves a higher precision and recall than manual browsing/searching for searching Web communities of a firm.

The hypotheses constructed were tested using a total of six organizations: IBM, Microstrategy, Sun Microsystems, Inc., Morgan Stanley, Eiffel Software, and the Boston Consulting Group. The firms were selected as diversified as possible to minimize the effects on the experiment results. Two pilot studies were conducted in order for us to refine the experimental tasks and experiment design. During the real experiment, thirty subjects, mostly junior students from the business school from an English university in Asia, were recruited and each subject was required to search for the Web communities of three out of the six different firms using the three different analysis approaches. Rotation was applied such that the order of analysis approaches and business firms tested would not bias the experimental results.

A graduate student majoring in library and information management was invited as the independent expert judge for this experiment. The judge manually browsed the firm's Web communities' Web sites, and individually summarized the Web pages into a number of themes. The themes formed the basis for evaluation and measurement of performance, which is further discussed in the next subsection.

5.3. Performance Measure

The experiment examined both quantitative and qualitative data for the experiment analysis. Our primary interests for quantitative data were in the performance and efficiency of the analysis approaches. Performance was evaluated by theme-based precision and recall, whereas efficiency was measured by the analysis time used by the subjects. Precision rate was the proportion of retrieved material actually relevant. Recall rate was the proportion of relevant material actually retrieved. They were calculated using the following formulae.

$$\text{precision} = \frac{\text{number of correct themes identified by the subject}}{\text{number of all themes identified by the subject}}$$

$$\text{recall} = \frac{\text{number of correct themes identified by the subject}}{\text{number of correct themes identified by expert judges}}$$

Analysis time was recorded as total duration spent on the analyzing, including both the response time of the system and subjects' browsing and analysis time. Time period during which the subjects wrote their answers on the answer sheet were included in the analysis time as well. User search logs recorded major observations of user behaviors, as well as the user's think aloud disclosure during the experiments. Questionnaires were also used to collect qualitative data to evaluate and compare the three analysis approaches.

6. Experiment Results and Analysis

6.1. Performance

Performance in terms of search effectiveness was one of our primary interests in the experiment and was evaluated by theme-based precision and recall. The test statistics was used to test the research hypotheses

1 and 2 in Section 5.2 and the main statistics of the 3 analysis approaches, Redips, backlink search engines, manual browsing, were summarized in Table 1.

Table 1: Experiment results

	Redips	Backlink search engines	Manual browsing
Precision	0.598	0.468	0.422
Recall	0.390	0.237	0.262
Time (seconds)	204	168	128

According to the Table 1, the mean precision (0.598) and mean recall (0.390) of Redips were larger than that of backlink search engines. Pairwise t-tests were conducted to measure the statistical significance of the differences between the analysis approaches and the results revealed that Redips's mean precision is not significantly higher than backlink search engines (p = 0.108), while Redips's mean recall is statistically significantly higher than that of backlink search engines (p = 0.0005). From the statistical results, Hypothesis 1 was accepted. A conclusion that Redips achieved higher precision and recall than backlink search engines for searching Web communities of a firm was confirmed.

Redips not only achieved a better performance than backlink search engines, but also did well compared with manual browsing. The mean precision and mean recall were all statistically better than that of manual browsing (p = 0.0044 and 0.0008, respectively). Hypothesis 2 was also confirmed.

We suggest that Redips excelled in precision and recall, when compared with the two benchmark approaches, due to several reasons. First, Redips has the ability of meta-searching, as described earlier in the paper, leverages the capabilities of multiple backlink search engines and provides a simple, uniform user interface. This would improve the recall rate of the system and thus the overall performance. Redips also allows users to enter the keywords to be included in the returned Web pages. This feature helped to increase the quality of the result set of Web pages, generating themes that were more related to the Web communities, and thus increasing the precision.

Second, clustering techniques like Arizona Noun Phraser and SOM map helped users narrow down the search scope and focuses on the interested Web communities. When reviewing the noun phrases and the categorization map generated by the system, the user can click on any interested Web communities that have been extracted from the full-text Web documents in order to discover a subset of Web documents that focus on the interested Web communities. This helped the user to decide if the Web communities were of the interests of the firm and improved the precision.

Third, the interactive user interface and visualization component, e.g. sorting and iterative map creation, make it possible for the users to focus on the interested Web communities and improve the search results. The SOM map technique used in Redips played an important role as well. SOM map was a summary in which the resulting Web communities had some support, e.g. the most frequent Web communities. The results helped the user to focus their attention on the most frequent Web communities that might usually of the greatest interest of the firm.

6.2. Efficiency

Another important measurement was the efficiency, which was evaluated by the analysis time of the subjects. The results were shown in Table 1. The mean analysis time of Redips was 204 seconds, i.e. 3.4 minutes, which was higher than both the backlink search engines (2.8 minutes) and manual browsing approaches (2.13 minutes), with a p-value of 0.0563 and 0.00166 respectively. The higher analysis time was not significant compared with backlink search engines, but significant compared with manual browsing. We observed that Redips spent a lot of time in fetching the full text of the URLs returned by the underlying backlink search engines and performing post-retrieval filtering and analysis. The users as well had comments on the analysis time used and recommend improvements in the time issues.

6.3. *Questionnaire Results*

The questionnaire was designed primarily to discover users' attitudes and subjective experience with the analysis approaches – Redips, backlink search engines, and manual browsing. The questions in the questionnaire evaluated and compared the analysis approaches on five dimensions, including user interface, usefulness of the information retrieved, subjects' level of certainty about their answers, user satisfaction of the analysis experience, and the amount of knowledge obtained after the analysis. The results of the questionnaire showed that Redips scored higher in ease of use (3.77) than backlink search engines (2.6) and manual browsing (3.2). The differences were statistically significant at the 5% level in both cases. Overall, our participants found that it is easier to search Web communities of a firm using Redips than using backlink search engines or using manual browsing. The subjects generally ranked Redips as the best approach among the three.

7. Conclusions and Future Directions

This paper, on seeing a demand for a tool for strategic business analysis applications, proposes the Redips architecture to help analysts work more efficiently in business intelligence analysis. An experiment was conducted to confirm the improvements of the new tool. We found that Redips achieved higher precision and recall than backlink search engines and manual browsing for searching Web communities of a firm. The results show that the combination of backlink searching and advanced analysis techniques can be used in Web community identification and analysis.

The applications of Redips in strategic planning process and business intelligent process are extensive. Redips helps analysts to do a comprehensive analysis of an organization's environment and find information about the environment in the Internet. With additional knowledge about the environment of organizations and the firm's Web communities, e.g. suppliers, customers, competitors, regulators, and pressure groups of the firm, the knowledge would form the basis of strategy formulation in the strategic planning process, in turn, creating and sustaining superior performance of the firm. While our study has shown the feasibility of the proposed approach, we believe that many other Web retrieval and analysis techniques can possibly be implemented on client-side search tools to improve efficiency and effectiveness in the study of Web communities in business intelligence analysis.

Acknowledgements

The Redips project has been supported in part by the following grants:

- HKU Seed Funding for Basic Research, "Using Content and Link Analysis in Developing Domain-specific Web Search Engines: A Machine Learning Approach," February 2004 – January 2006.
- HKU Seed Funding for Basic Research, "Searching the World Wide Web Backwards for Business Intelligence Analysis," January 2005 – December 2006.
- NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management", April 1999 – March 2002.

We would like to thank the domain experts and the subjects for their participation in our experiment.

References

1. Bradford, R. W., Duncan, P. J., and Tarcy, B. *Simplified Strategic Planning: A No-Nonsense Guide for Busy People Who Want Results Fast!* Chandler House Press, Worcester, 1999.
2. Chau, M. and Chen, H. "Personalized and Focused Web Spiders," in *Web Intelligence*, 1st ed. N. Zhong, J. Liu, Y. Yao (eds.), Springer-Verlag, Berlin, February 2003, pp. 197-217.
3. Chau, M., Zeng, M., and Chen, H. "Personalized spiders for Web search and analysis," *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01)*, ACM Press, New York, 2001, pp. 79-87.
4. Chau, M., Zeng, D., Chen, H., Huang, M., and Hendriawan, D., "Design and Evaluation of a Multi-agent Collaborative Web Mining System," *Decision Support Systems*, Special Issue on Web Retrieval and Mining, 35(1), 2003, pp. 167-183.
5. Chen, H. and Chau, M., "Web Mining: Machine Learning for Web Applications," *Annual Review of Information Science and Technology*, 38, 289-329, 2004.
6. Chen, H., Chau, M., and Zeng, D. "CI Spider: a tool for competitive intelligence on the Web," *Decision Support Systems* (34), 2002, pp. 1-17.
7. Chen, H., Fan, H., Chau, M., and Zeng, D. "MetaSpider: meta-searching and categorization on the Web," *Journal of American Society for Information Science & Technology* (52:13), 2001, pp. 1134-1147.
8. Chen, H., Houston, A. L., Sewell, R. R., and Schatz, B. R. "Internet browsing and searching: user evaluations of category map and concept space techniques," *Journal of the American Society for Information Science* (49:7), 1998, pp. 582-603.
9. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. "From data mining to knowledge discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining*, 1st ed. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), MIT Press, Cambridge, Mass., 1996, pp. 1-36.
10. Feldman, R. and Dagan, I. "Knowledge discovery in textual databases (KDT)," *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, AAAI Press, Montreal, Canada, August 20-21, 1995, pp. 112-117.
11. Gilad, B. and Gilad, T. *The Business Intelligence System*, AMACOM, New York, 1988.
12. Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001.
13. Hearst, M. A. "Text data mining: Issues, techniques, and the relationship to information access," *UW/MS workshop on data mining*, July 1997, presentation notes.
14. Kohonen, T. *Self-Organizing Maps*, Springer-Verlag, Berlin, 1995.
15. Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. "Trawling the Web for Emerging Cyber-communities," *Proceedings of the 8th International World Wide Web Conference*, 1998.
16. Lin, X., Soergel, D., and Marchionini, G. "A self-organizing semantic map for information retrieval," *Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'91)*, ACM Press, New York, 1991, pp. 262-269.
17. Prescott, J. E. and Smith, D. C. "SCIP: who we are, what we do," *Competitive Intelligence Review*, 1991.
18. Reid, E. O. F. "Identifying a Company's Non-Customer Online Communities: a Proto-typology," *Proceedings of the Hawaii International Conference on System Sciences*, Big Island, Hawaii, January 6-9, 2003.
19. Salton, G. "Another look at automatic text-retrieval systems," *Communications of the ACM* (29:7), 1986, pp. 648-656.
20. Schermerhorn, J. R. *Management*. John Wiley & Sons, Inc., Hoboken, 2001.
21. Selberg, E. and Etzioni, O. "The MetaCrawler architecture for resource aggregation on the Web," *IEEE Expert* (12:1), 1997.
22. Tan, A. H. "Text Mining: The state of the art and the challenges," *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, 1999.
23. Tolle, K. and Chen, H. "Comparing noun phrasing techniques for use with medical digital library tools," *Journal of the American Society for Information Sciences* (51:4), 2000, pp. 352-370.
24. Voorhees, E. and Harman, D. "Overview of the sixth Text Retrieval Conference (TREC-6)," in *NIST Special Publication 500-240: The Sixth Text Retrieval Conference (TREC-6)*, 1st ed. Voorhees, E. and Harman, D. (eds.), National Institute of Standards and Technology, Gaithersburg, MD, 1998, pp. 1-24.