Redips: Backlink Search and Analysis on the Web for Business Intelligence Analysis

Michael Chau and Boby Shiu
School of Business, The University of Hong Kong, Pokfulam, Hong Kong. E-mail: mchau@business.hku.hk, kwshiu@hkucs.org

Ivy Chan
The Hong Kong Community College, The Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong. E-mail: ccivy@polyu.edu.hk

Hsinchun Chen
Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721. E-mail: hchen@eller.arizona.edu

The World Wide Web presents significant opportunities for business intelligence analysis as it can provide information about a company's external environment and its stakeholders. Traditional business intelligence analysis on the Web has focused on simple keyword searching. Recently, it has been suggested that the incoming links, or backlinks, of a company's Web site (i.e., other Web pages that have a hyperlink pointing to the company of interest) can provide important insights about the company's "online communities." Although analysis of these communities can provide useful signals for a company and information about its stakeholder groups, the manual analysis process can be very time-consuming for business analysts and consultants. In this article, we present a tool called Redips that automatically integrates backlink meta-searching and text-mining techniques to facilitate users in performing such business intelligence analysis on the Web. The architectural design and implementation of the tool are presented in the article. To evaluate the effectiveness, efficiency, and user satisfaction of Redips, an experiment was conducted to compare the tool with two popular business intelligence analysis methods—using backlink search engines and manual browsing. The experiment results showed that Redips was statistically more effective than both benchmark methods (in terms of Recall and *F*-measure) but required more time in search tasks. In terms of user satisfaction, Redips scored statistically higher than backlink search engines in all five measures used, and also statistically higher than manual browsing in three measures.

Received January 14, 2006; revised March 11, 2006; accepted March 11, 2006

 ● Published online 20 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20503

Introduction

Business intelligence can be defined as the process of monitoring a firm's external environment to obtain information relevant to its decision-making process (Gilad & Gilad, 1988). In the past, the sources of business intelligence information mainly consisted of published company reports, subscription-based online databases, and other kinds of printed information. However, this practice has changed significantly in the past 10 years. With the advances in information technologies, many resources and information are now accessible on the Web. In late 2004, Google announced that they had indexed more than 8 billion Web pages. The Web has become a large repository of information that could be relevant to a firm's decision making. For example, looking into the Web sites of a firm's competitors can reveal useful information about the firm's competitive environment (Chen, Chau, & Zeng, 2002). The Web sites of other stakeholders of the firm, like customers, suppliers, and pressure groups, can also provide important information about the firm's competitive environment. As the Web sites of these stakeholders often have hyperlinks pointing to each other or are pointed to by the same set of Web sites, they are known as the "Web communities" of the firm (Kumar, Raghaven, Rajagopalan, & Tomkins, 1998; Reid, 2003).

Although the identification of Web communities is important in the business intelligence analysis process, most existing Internet tools have been designed for traditional keyword-based search on the Web. Keyword-based searching can only return search results with a page containing the search keyword, but does not guarantee the relationship between the search result and the firm of interest. Despite the efforts of the search engines to refine search results to a

higher quality, many of the Web pages are still irrelevant or outdated, and analysts have to filter out the unwanted Web pages manually. In addition, content analysis tools are often not available in these search engines. Analysts often have to spend a long time to browse the content of each Web page manually, acquire the overall concept of the set of search results, and summarize the information. This can be a very time-consuming and mentally exhausting process. A tool that automatically identifies and analyzes the Web communities of a firm is therefore highly desired.

In this article, we try to address the existing problems using our "Redips" architecture. *Redips* is the reverse spelling of the word *Spider*. The rationale behind the name is that Redips does not search using breadth-first search or keyword-based search like traditional Web spiders (Chen et al., 2002). Instead, Redips searches the Web *backwards*—when a user inputs the URL (uniform resource locator) of a firm into Redips, the tool will search the Web backwards by searching Web pages that have links pointing to the given URL. The search results will represent the firm's Web communities. The backlink search results will be fetched in real-time to the local computer and Redips will examine the fetched Web pages and perform text analysis to extract the important phrases from the stored Web pages. These phrases symbolize a vector of themes and topics in the Web pages that can be used by analysts to identify the main areas of interest in the Web communities. Moreover, Redips allows users to visualize the retrieved Web pages in the form of a two-dimensional map using the self-organizing map (SOM) technique. The map would help analysts 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 rest of this article is organized as follows. In the next section, we review related work in Web communities, business intelligence analysis, and Internet-based analysis tools. Then we describe the research questions and the problem of the existing analysis tools. This is followed by an outline of the architecture of our analysis tool Redips. A sample user session is presented and we discuss an evaluation study conducted to evaluate the proposed tool and the corresponding experiment design. Next we present the experiment results and analyze the results using a statistical analysis. In the last section, we conclude our work and discuss our future research directions.

Research Background

Business Intelligence Analysis and Web Communities

Facing the challenges of the global marketplace, informed and demanding customers, bargaining suppliers, strategic competitors, and evolving technologies, a business organization is in an environment much more competitive than ever before. To succeed in such a competitive business world, today's business must always keep an eye on what is happening in the industry every day. This would allow them to make decisions to respond and adjust quickly to the changes in the business environment before it is too late. As stated earlier, the term *business intelligence*, also referred to as *competitive intelligence*, *commercial intelligence* or *corporate intelligence*, is used to describe the process of monitoring a firm's external environment to obtain information relevant to its decision-making process (Gilad & Gilad, 1988). The typical business intelligence process consists of a series of activities that involve identifying, gathering, developing, analyzing, and disseminating information (Gilad & Gilad, 1988; Keiser, 1987; Vedder, Vanecek, Guynes, & Cappel, 1999). One of the important steps in the process is to identify the customers, suppliers, competitors, stockholders, public-interest groups, labor unions, political parties, governments or other variables in the environment to be monitored (Schermerhorn, 2001). With the rapid growth of the Internet in recent years, most of this information can be accessible on the Web, including organization/company Web sites, discussion forums, resource directories, or individual Web pages. Moreover, because these stakeholders share a common interest (either in a firm, a product, or a market), they often have hyperlinks pointing to each other. These, together with the Web pages most popular among them, form the Web communities of the firm or the market of interest (Reid, 2003). Web communities thus have become a very important component in business intelligence analysis in the Internet age and have been investigated in previous business intelligence research (Chung, Chen, & Nunamaker, 2003, 2005; Reid, 2003).

Web communities can be classified into two categories: explicit and implicit Web communities. Explicit communities are the communities that can be easily identified on the Internet. Kumar et al. (1998) discussed the Porsche newsgroup as an example of explicit community of Web users interested in Porsche Boxster cars. Such communities are often found in resource collections in Web directories such as the Yahoo directory. Analysts can often use a manual method to find a firm's explicit communities by browsing the firm's newsgroup or through Web directories.

Implicit communities are relatively more difficult to find using a manual browsing method. According to Kumar and colleagues (1998), implicit communities refer to the distributed, ad hoc, and random content-creation related to some common interests on the Internet. These pages often have links to each other, but the common interests of implicit communities are sometimes too specific 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 reasonable to assume that the content pages created by these communities would provide hypertext links back to the firm's homepage for reference (Reid, 2003). Therefore, to find a firm's online communities, it is necessary to find the Web pages that have hyperlinks pointing to the firm's URL, i.e., the inbound links of the firm's Web site. In previous research (Reid, 2003), it has been shown that backlink searching on the Web can be used to

DOI: 10.1002/asi

find the implicit Web communities of a company that are otherwise difficult to identify. A case study of MicroStrategy, an e-business software company, was performed and the results were promising: More than 10 types of stakeholders of the firm were revealed from the backlinks of the firm. However, only a manual approach was used in the study and it was a labor-intensive process.

Internet-Based Business Intelligence Tools

Many tools have been used to assist Web-based business intelligence analysis. The simplest tool may be just a Web browser like Internet Explorer. A browser is a client software program used for searching and viewing various kinds of information on the Web. Using a manual browsing method, an analyst only needs to enter the URL of a firm or its stakeholders in the browser and then manually browse the information for further analysis.

This manual browsing method is common to analysts; as many people are experienced in Internet surfing by now. 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 & Smith, 1991). It is not practical for analysts to go through the Web sites of all the 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 collect manually the most updated versions of every Web page for analysis.

To address these problems, Web-based business intelligence tools have been developed to do more than simple browsing. In the following, we will review the literature and existing tools in Web-based business intelligence analysis. Based on their functionalities, the tools can be classified into three categories, namely Web searching, content analysis, and visualization. In addition, there are also tools that integrate more than one of these functions. The discussion in the following will be based on this taxonomy.

Web search tools. Web search engines are the most popular way that people use to search for information on the Web. Each engine has its own characteristics and employs its preferred algorithm in indexing and ranking Web documents. For example, Google (www.google.com) and AltaVista (www.altavista.com) allow users to submit queries and present the search results in a ranked order, whereas Yahoo (www.yahoo.com) groups Web sites into categories, creating a hierarchical directory of a subset of the Web. A Web search engine usually consists of four main components: spiders, an indexer, retrieval and ranking facility, and user interface (Brin & Page, 1998; Chau & Chen, 2003). Spiders are responsible for collecting documents from the Web using different graph search algorithms. The indexer creates indexes for Web pages and stores the indexes 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 engine is the meta-search engine, 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 (Chen, Fan, Chau, & Zeng, 2001; Selberg & Etzioni, 1997).

Besides server-side Web searching, there are also client-side tools that allow users to perform searching on the Web or downloading of Web sites. For example, Webseeker (www. bluesquirrel.com/products/webseeker) and Copernic Agent (www.copernic.com) are two meta-search tools that run on the client's computer instead of the server side. WebMiner (tribolic.com/webminer), Grab-a-site (www.bluesquirrel.com/products/grabasite), and Teleport (www.tenmax.com/teleport) are all software that helps users to download specified files from given Web sites so that the Web sites' content can be archived and further analyzed more easily.

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. Backlink searching can identify these communities because the stakeholders 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 in addition to performing regular indexing, the indexer will also index the links of each Web page collected. The information on these links is stored into 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 (www.google.com). Google allows users to use the reserved word "link" as an operator in the query. The query "link:siteURL" shows the user's pages that point to a given URL (Google, 2005). For example, the query "link:www.google.com" will return pages that contain a hyperlink to Google's home page. With the Google Web APIs service, a program developed by Google, software developers can also use the backlink search feature directly from their own computer programs (Google, 2004). A special query capability "Back Links" is included in the APIs and likewise, the query prefix "link:" lists Web pages that have links to the specified Web page. Altavista (www.altavista.com) 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. 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 the backlink search feature.

DOI: 10.1002/asi

Unlike general-purpose searching, no meta-search engines are available for searching backlinks in the current search engine market. A meta-backlink search engine may be able to improve retrieval performance just like general meta-search engines; however, this has not been studied to date.

Content analysis tools. After documents are retrieved from the Web, indexing and text mining techniques are often applied to perform further analysis on the documents. Text mining, also known as text data mining (Hearst, 1997) or knowledge discovery from textual databases (Feldman & Dagan, 1995), refers generally to the process of extracting interesting and nontrivial 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, Piatetsky-Shapiro, & Smyth, 1996). Text mining is a fascinating multidisciplinary field, including information retrieval, textual information analysis, information extraction, and information clustering.

Text-mining tools help analysts to 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 that are of the firm's interest.

As the information on the Web mainly contains textual contents, e.g., HTML documents or PDF documents, text mining and textual information analysis are often studied in Internet-based analysis tools. Textual information analysis relies on the indexing of the source Web documents. Many techniques of indexing the source documents and extracting key concepts from text have been proposed in recent years. One of the proven techniques is an 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. Single word indexing allows users to search for documents that contain the search keywords and has been widely adopted in information retrieval systems. The output is a vector of extracted words representing the documents of interest, based on each term's frequencies. Different from single word indexing, phrase indexing outputs a vector of extracted phrases to represent the documents of interest. The underlying motivation to use phrases, especially noun phrases in information retrieval, is that the phrases can convey and represent more precise meaning than single words and as a result, capture a "richer linguistic representation" of document content (Anick & Vaithyanathan, 1997). An analysis tool Arizona noun phraser (AZNP) has been developed based on this concept (Tolle & Chen, 2000). The tool extracts all the noun phrases from each Web document based on a part-of-speech tagging and a set of linguistic rules.

The output of automatic indexing algorithms can often be used in further text mining analysis. Document classification or document clustering can be applied to the noun phrases to deduce patterns and relationship across documents and to derive firm-related knowledge in the analysis project of the firm. Document classification is one form of data analysis that can be built to categorize the documents into a predetermined set of document classes or concepts (Han & Kamber, 2001). Web documents are categorized into predefined classes in this approach. Because 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 objects into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. In text mining, the classes or clusters would have a category label defined based on the keywords or phrases that appear in the Web documents in that category. The fact 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 popular clustering approaches for a client-side Internet analysis tool is Kohonen's (1995) self-organizing map (SOM). The SOM algorithm classifies documents into various categories automatically determined during the clustering process, with the underlying neural network algorithm technology. The algorithm clusters the retrieved documents into different regions and displays the results as a two-dimensional map. More details on its visualization capabilities are discussed in the next subsection.

Visualization tools. Visualization tools are often used to display the document classification or document clustering results to users in an organized and meaningful way using certain graphical representation. A graphical representation of the document clusters helps analysts and managers to better comprehend the returned documents, identify interesting documents more quickly, gain a quick overview of the documents' contents, and acquire knowledge more effectively (Johnson, 1994). Furthermore, a graphical representation summarizes the key results and shortens the time for users to digest the data, information, knowledge, and intelligence in the documents.

Depending on the respective visualization targets, visualization tools can be classified into two categories. The first one is a category of tools that visualizes the document attributes, e.g., document type, location, created date, title, document size, source, topic, and author. The objective is to provide the users with additional information about the retrieved documents. Another category includes tools that utilize interdocument similarities to reduce the multidimensional document space to a two-dimensional or three-dimensional space (clusters) by aggregating similar documents under the same topic. The objective is to provide the users with a quick overview of the whole document collection. Cluster labels are decided based on the words or phrases written in the document collection.

DOI: 10.1002/asi

A variety of representation schemes for document clustering results is available in the current market. Cartia's ThemeScape is an enterprise information mapping application that presents clusters of documents in landscape representation. InXight (InXight Software, Inc., Sunnyvale, CA) also offers a visualization tool known as VizControls that performs value-added postprocessing of search results by clustering the documents into groups and displaying based on a hyperbolic tree representation. Semio Corp's SemioMap (Semio Corp., San Mateo, CA) employs a three-dimensional graphical interface that maps the links between concepts in the document collection.

Kohonen's self-organizing map, as discussed earlier, is not only a clustering algorithm but is also an example of a visualization tool based on interdocument similarities. This approach clusters documents into various topics that are automatically generated in real time using neural network algorithms (Chen, Schufels, & Orwig, 1996; Kohonen, 1995; Lin, 1997). Every document is assigned to its corresponding regions in a two-dimensional graphical map displayed to the user. Every region contains similar documents under the same topic whereas those regions with conceptually similar topics are located close to each other on the self-organizing map.

Visualization tools display graphical representation to the users and help them to understand more fully the set of retrieved documents in a short amount of time, which is especially important in the today's fast-changing business world. Clearly, the visualization tools improve the user experience and form an important component in Internet-based analysis tools. Despite the benefits of visualization tools, many tools discussed in this section are used only in prototype research systems, and only a few tools are put on the commercial market for real business applications. Training and system assistance are needed to improve the effectiveness of clustering approaches and visualization tools for Internet-based analysis tools (Sutcliffe, Ennis, & Hu, 2000).

Integrated tools. In recent years, many tools have been developed to incorporate more than one of the functions of searching, analysis, and visualization. For example, CI Spider conducts a breadth-first search and best-first search on the Web and performs document clustering and visualization on the search results (Chen et al., 2002). Collaborative Spider, an extended version of CI Spider, is a multiagent system designed to improve search effectiveness by sharing relevant search sessions among users (Chau, Zeng, Chen, Huang, & Hendriawan, 2003). The Focused Crawler (Chakrabarti, van den Berg, & Dom, 1999) performs topic-specific searches and classification on the retrieved documents. The Business Intelligence Explorer (Chung et al., 2003, 2005) performs meta-searching on the Web and uses a tree hierarchy and a knowledge map to display the search results.

Many commercial tools are also available. For instance, Convera's RetrievalWare (www.convera.com/products/retrievalware) collects, monitors, and indexes information from text documents on the Web as well as graphic files. Categorization and entity extraction also can be performed on the retrieved documents. Autonomy's products (Autonomy Group, San Francisco, CA) (www.autonomy.com) support a wide range of information collection and analysis tasks, which includes automatic searching and monitoring information sources in the Internet and corporate intranets, and categorizing documents into categories predefined by users or domain experts. Verity's knowledge management products (www.verity.com), such as Agent Server, Information Server, and Intelligent Classifier, also perform similar tasks in an integrated manner.

Research Questions

Current Internet-based tools are often good for information retrieval but lack the functionality to help finding the Web communities of a firm. Backlink searching, which has been suggested as a promising way to identify Web communities (Reid, 2003), is still often performed manually and has not been integrated into business intelligence tools that incorporate other functionalities such as document analysis and visualization. In this article, we address the following research questions: (a) How can we develop an automated tool to identify Web communities of a firm using backlink searching and existing analysis and visualization techniques? (b) Can such an integrated tool facilitate business intelligence analysis on the Web based on searching and analysis of Web communities? (c) How does such a tool compare with existing methods in Web community analysis?

System Architecture of Redips

To answer the above research questions, we propose the Redips architecture as shown in Figure 1. The architecture of Redips is based on the MetaSpider system developed in our previous research (Chau et al., 2001; Chen et al., 2001). The main modules include the spider, the Arizona noun phraser, the self-organizing map, and the user interface.

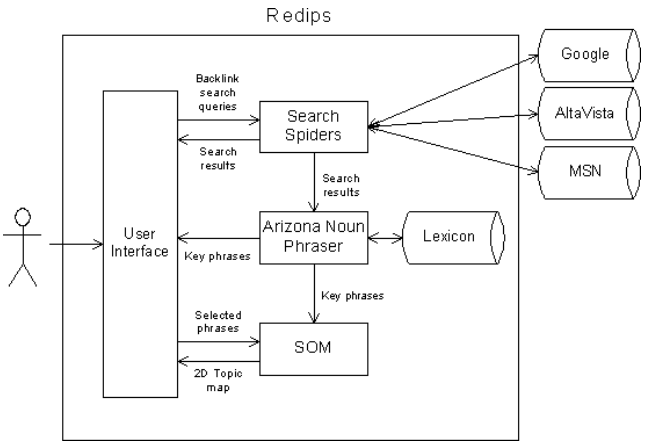

FIG. 1. System architecture of Redips.

DOI: 10.1002/asi

First, the user interface accepts URLs and keywords from users. The spider module formulates and sends the queries to several popular backlink search engines and retrieves their search results. The top *n* pages (*n* may be changed in the search options, with the default value being 20) returned from those search engines form the preliminary search results. Redips then fetches all the actual content of the Web pages based on the URLs in this set of search results. The Arizona noun phraser is a natural language processing tool that performs key phrase extraction on Web pages. Noun phrases are extracted from the documents; this allows users to know what key topics are related to the Web sites and the keywords specified. The self-organizing map visualizes the concepts in a two-dimensional map, which categorizes the Web pages by clustering them into regions, each of which represents a topic. All of these functionalities allow users to automatically collect and analyze information more effectively and represent the information in a more meaningful way.

Backlink Search

Redips has the ability of meta-searching through connecting to different backlink search engines. This approach leverages the capabilities of multiple backlink search engines by providing a simple, uniform user interface. Meta-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, in a way similar to regular meta-searching. In our implementation, Redips connects to three backlink search engines: Google, Altavista, and MSN Search. More backlink-search engines may be easily added. Unlike other meta-searching tools that 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 postretrieval filtering and analysis.

Redips makes the extraction of the implicit Web communities easier. The backlink search engine results are the Web pages that point to the firm's URL. In other words, these are the Web communities of the firm. In addition, the Redips architecture has the following advantages over general backlink search engines: (a) the meta-searching feature is implemented to improve the search coverage; (b) the optional feature allows users to enter keyword(s) to be included in the returned Web pages and to specify other search options; and (c) the filtering feature allows users to filter Web pages in specific domains or specific locations, those from the same Web host, those with any stop terms, or those that do not exist anymore. These additional features help analysts to extract the implicit Web communities of the firm more effectively. These features also alleviate the problem of finding pertinent and useful information from search results.

Content Analysis

The content analysis is conducted in two phases. First, the Arizona Noun Phraser (AZNP; Tolle & Chen, 2000), developed at the University of Arizona, is used to extract key phrases that appear in the documents retrieved and filtered by the spiders. The Arizona Noun Phraser (AZNP) was developed at the University of Arizona with the goal to extract high-quality phrases from textual data (Tolle & Chen 2000). It has three main components: a tokenizer, a tagger, and a noun phrase generator. The tokenizer component is designed to take raw text input (text or HTML files) and create output that conforms to the UPenn Treebank word tokenization rules. The tagger component of the AZNP is a significantly revised version of the Brill tagger (Brill 1993). A lexicon is used by this tagger module. The third major part of the AZNP is the phrase generation component, which converts the words and associated part-of-speech tags generated by the tagger into noun phrases. The AZNP is used to extract the key phrases that appear in the documents retrieved and filtered by the Spider module. The frequencies of occurrences of the phrases are recorded and sent to the User Interface. The frequencies of occurrences of the phrases are recorded and sent to the User Interface. The AZNP helps analysts evaluate the links of Web communities in a short time and provides an overview of the entire document set to the user. Moreover, because the extraction of key phrases is performed automatically, the analysis time is much reduced when compared with manual analysis, especially when the number of files to be processed is large.

In the second phase, the SOM 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 the documents (Lin, Soergel, & Marchionini, 1991; Orwig, Chen, & Nunamaker, 1997). In the algorithm, each document is represented as an input vector of noun phrases extracted by AZNP and a two-dimensional grid of output nodes is created (Chen et al., 1998). The network is trained through repeated presentation of all inputs. Each region in the map is labeled by the phrase that best describes the key concept of the cluster of documents in that region. The size of the color block indicates the relative significance of the term to the documents collected. The relative proximity reveals the distance between the two concepts presented by the respective term. In backlink analysis, the SOM helps cluster these Web pages that link to a firm of interest into the Web communities of the firm by grouping Web pages with similar content into labeled clusters. The maps created by SOM would draw users' attention easily. Users can quickly understand the overview of the Web pages retrieved. This would shorten the overall analysis time and the decision-making time, which is very important in today's fast-changing business world. Furthermore, the Dynamic SOM (DSOM) technique is used in Redips such that the user can select and deselect phrases for inclusion in the analysis and produce a new map on the fly within seconds. New maps can be generated until the users are satisfied with the results. More technical details of the two components and their integration into the architecture can be found in our previous work (Chau et al., 2001; Chen et al., 2001, Tolle & Chen 2000).

DOI: 10.1002/asi

Sample User Session

There are five basic steps in performing a search session using Redips; they include not only the search process but also the business intelligence analysis process, as described in a previous section. In the following, we provide an example to illustrate how a user interacts with Redips.

A user should start using Redips by entering the Web site to be analyzed and choosing the backlink search engines to be included in the search. Redips also provides an optional feature that lets the user enter the keyword(s) to be included in the returned Web pages. A screenshot of the main user interface of the system is shown in Figure 2. In this example, the user has entered the Web site of International Business Machines Corporation (IBM; www.ibm.com) and chosen all of the three backlink search engines. Other search options, like the domains to be included in the search results, the timeout for the spider, the number of simultaneous threads to be used, also can be specified. In other words, the user can define the intelligence analysis objectives, such as the firm, the information sources, and the search topics, in this step. After all the settings have been specified, the user can click the *Search* button to commence the search and the Spider module will start sending requests to the chosen backlink search engines.

After submitting the search requests, the system will collect the backlinks of the given URL from the different search engines. The results will be displayed to the user as a list (see Figure 3). To avoid causing confusion to the user, our system only groups URLs that are returned from more than one search engine, but does not perform any re-ranking of the search results in this step. The user can click on any URL and view the actual Web page. Exploratory, preliminary research can be carried out to identify the firm's Web communities in this step.

After browsing through the list of search results, the user can click on the *Fetch* button to command the system to download from the Web the complete content of the URLs retrieved from the search engines. The pages are displayed dynamically during the fetching process. The user can explore the results and read the content of any of the Web pages collected. The user can also switch to the *Good URL List* to browse the filtered result (Figure 4). In this list, Web pages that no longer exist (e.g., the page has been removed from the Internet; the Web server hosting the page reports a 404 "Page Not Found" error; or the Web server does not respond within a specific timeframe) and those that do not contain the search keyword(s) will be excluded. Overall, this step automates the process of information collection from the Web communities on the Internet.

After the complete contents of all the Web pages have been downloaded to the system, the results are sent to the Arizona Noun Phraser for further analysis. Noun phrases are extracted from the Web pages and analyzed. One should note that noun phrases are extracted without stemming (removal of word suffix). The frequency of appearance of each noun phrase is displayed and the user can browse the

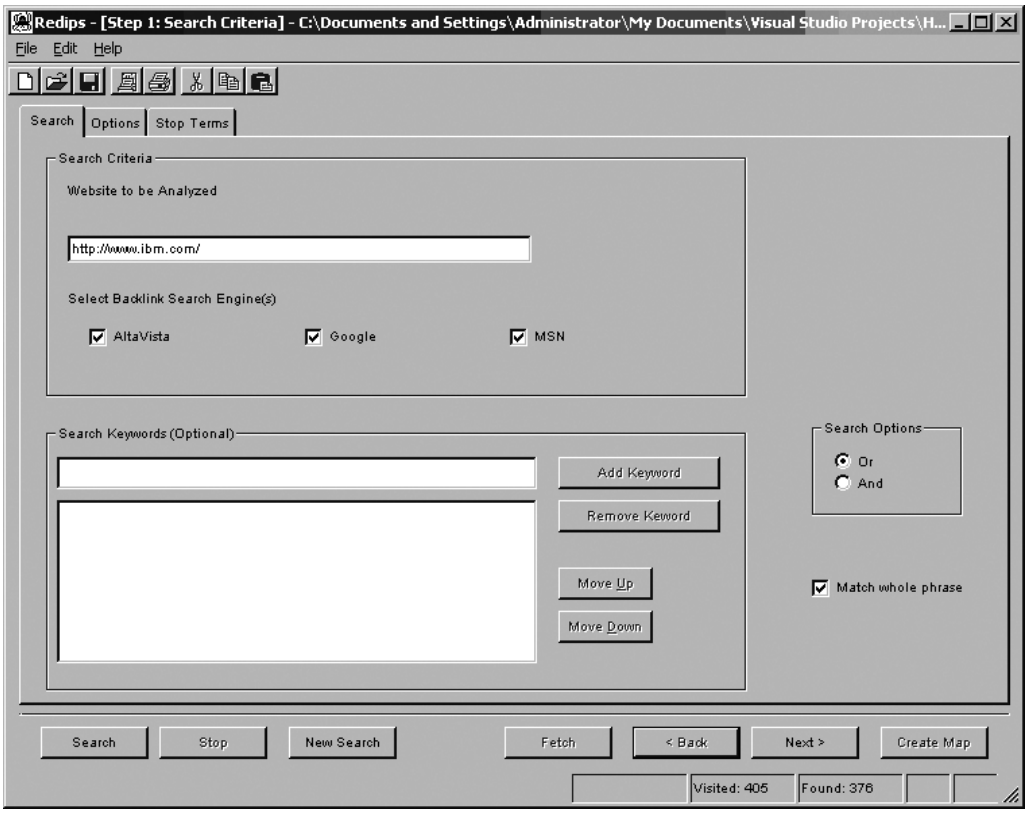

FIG. 2. The main user interface.

DOI: 10.1002/asi

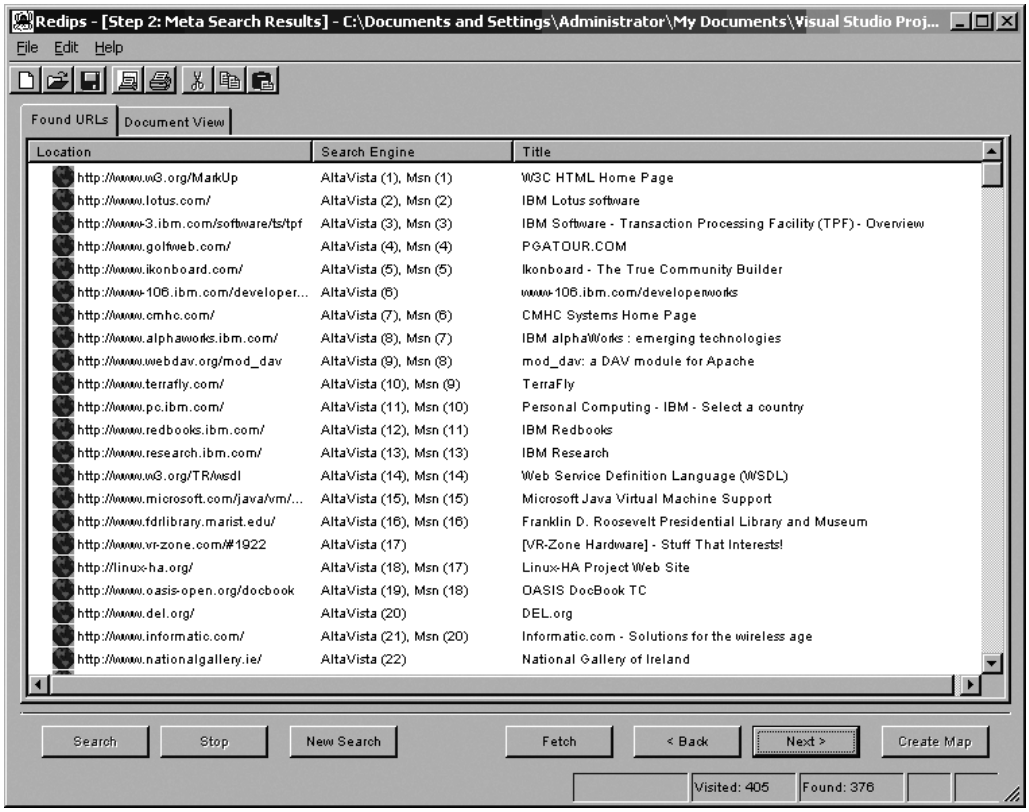

FIG. 3. Preliminary search results returned by the search engines.

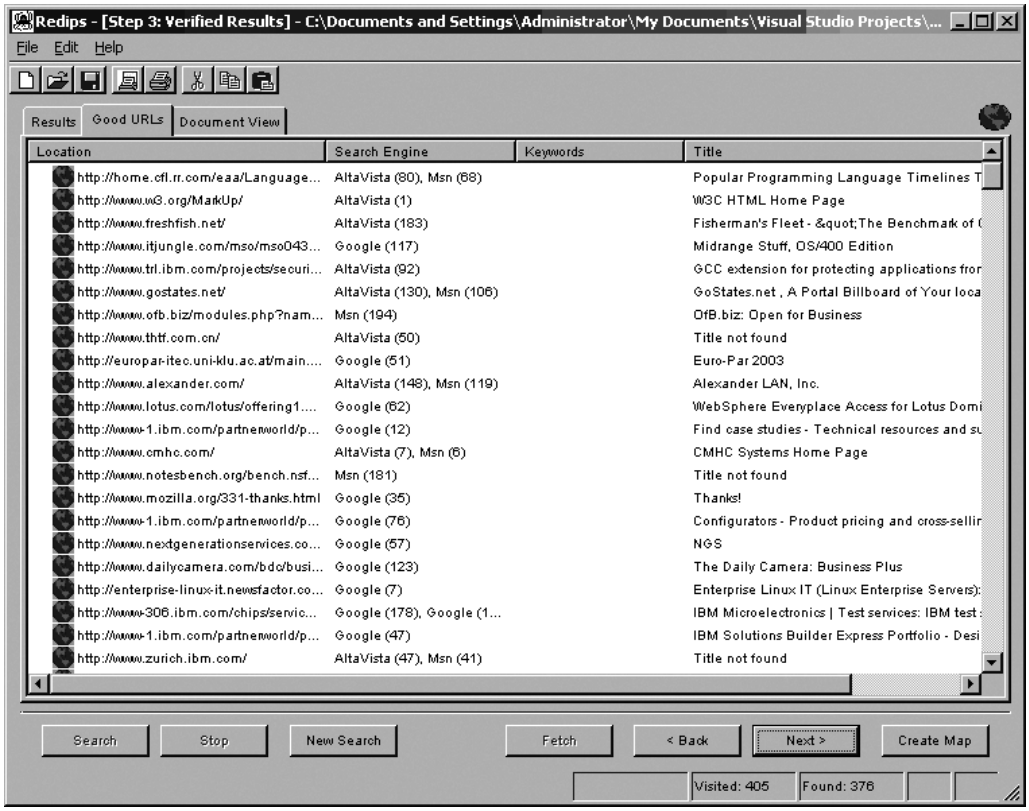

FIG. 4. Fetching the complete content of the Web pages from the Internet.

DOI: 10.1002/asi

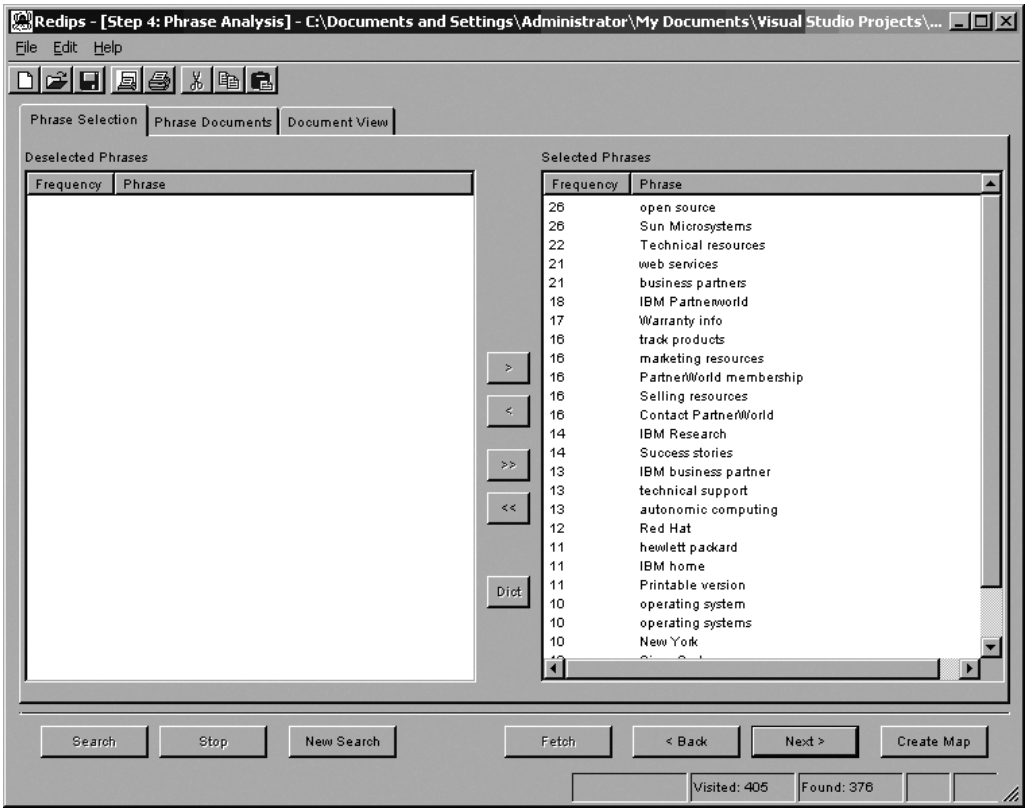

FIG. 5. Noun phrases extracted from the Web pages collected.

pages that contain any particular noun phrase by clicking on the phrase (see Figure 5). In this step, the user can get a quick overview of the most frequent topics that appear in the Web communities of IBM, e.g., open source, Sun Microsystems, technical resources, and Web services. The user can also select what noun phrases are to be included in creating the categorization map in the next step.

A categorization map, the SOM, is generated based on the noun phrases selected in previous step. The Web pages are categorized into different regions on the map, based on the noun phrases they contain (see Figure 6). The SOM summarizes and visualizes the Web links as Web communities in a graphical representation, which can be useful in the business-intelligence analysis process. In this example, a few Web communities of IBM are identified on the map, e.g., IBM Partnerworld, Sun Microsystems, the open source community, and so on. More important communities occupy larger regions (e.g., open source), and similar communities are grouped close to each other on the map. Such information can help the analyst identify the Web communities more easily. The user can also go back to the previous step and choose a new set of noun phrases to refine the map.

Experiment Design and Hypotheses

To study whether Redips outperforms other Internet-based analysis tools and to evaluate the effectiveness and efficiency of different methods in performing document retrieval and categorization in the business intelligence analysis process, we conducted an evaluation study to compare our system with two traditional business intelligence analysis approaches: backlink search engines and manual browsing. In this section, we will describe the design of our experiment.

Because Redips has been designed to facilitate both document retrieval and document categorization, traditional evaluation methodologies that treat document retrieval and document analysis separately cannot be directly applied in our evaluation. In our experiment design, we use the evaluation framework developed based on theme identification (Chen et al., 2001, 2002). This framework enables us to measure the performance of the combination of the systems' retrieval and categorization features. The test subjects would be asked to identify the major themes related to the Web communities of a given firm's Web site.

The experiment tasks were constructed based on this theme-based evaluation framework. Based on our previous work (Chen et al., 2001), we defined a theme as "a short phrase that summarizes a specific aspect of Web communities." Noun phrases like "consulting firms," "business intelligence," "java technology," "financial consulting" are examples of themes in the experiment tasks. The theme-based framework enables us to evaluate the effectiveness and efficiency in locating and analyzing the Web communities using different approaches. The performance measurements we used, including precision, recall, and F-measure, are further discussed in the next section.

DOI: 10.1002/asi

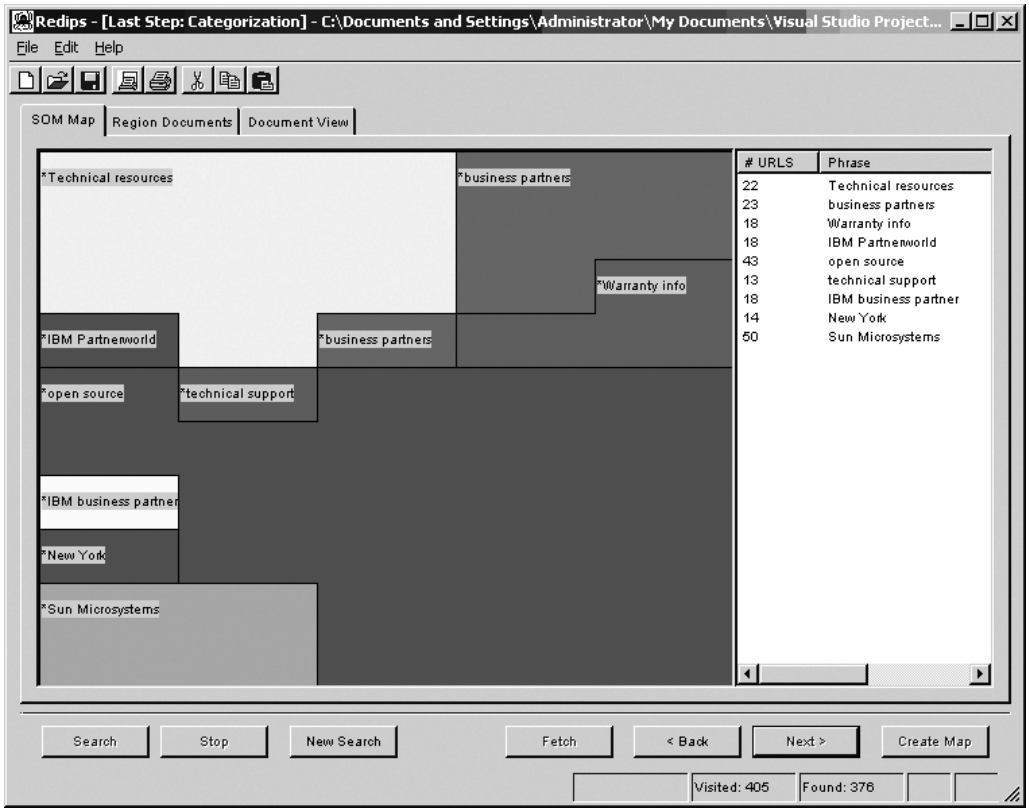

FIG. 6. Web pages categorized into Web communities.

In the experiment, each subject was asked to imagine themselves as part of a consultant group 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 given to the subject. The subject would then use the specified approach to search for the Web communities of the firm. Based on the findings, the subject was required to summarize the results as a number of themes, based on our definition above, which would give an overview of the Web communities of the firm. At the end of the experiment, each subject was asked to fill out a questionnaire about the user's experience with the different analysis approaches.

Our aim was to compare Redips with popular ways in which analysts conduct Web community analysis. Therefore, we chose two existing approaches to Web community analysis as our benchmarks for comparison. The first is the use of commercial search engines that support backlink search, such as Google and MSN Search. In this approach, users can get the Web sites that have a hyperlink pointing to the Web site of interest. This approach is the most natural existing method that an individual would undertake when asked to perform a task on Web community analysis. In addition, most users are familiar with Web search engines and Web searching so they would find it easy and comfortable to perform their tasks using this approach. The second benchmark is the use of manual browsing, where users freely explore the contents of a given Web site to find any Web communities using a Web browser.

We suggested that Redips would outperform the backlink search engine approach and the manual browsing approach in the business intelligence analysis of Web communities. The following hypotheses were posed in the experiment:

Hypothesis 1. Redips achieves higher effectiveness (measured by precision, recall, and *F*-measure) than backlink search engines for searching the Web communities of a firm.

Hypothesis 2. Redips achieves higher effectiveness (measured by precision, recall, and *F*-measure) than manual browsing for searching the Web communities of a firm.

Hypothesis 3. Redips requires less time than backlink search engines for searching the Web communities of a firm.

Hypothesis 4. Redips requires less time than manual browsing/searching for searching the Web communities of a firm.

Hypothesis 5. It is easier to search the Web communities of a firm using Redips than using backlink search engines.

Hypothesis 6. It is easier to search the Web communities of a firm using Redips than using manual browsing.

The hypotheses were tested using six business firms: IBM, Microstrategy, Sun Microsystems, Inc., Morgan Stanley, Eiffel Software, and Boston Consulting Group. These firms were chosen based on the consultation with a researcher with expertise in business intelligence analysis and also based on the data used in previous research (Chen et al., 2002; Reid,

DOI: 10.1002/asi

2003). To minimize the selection effects on the experiment results, the firms were selected as diversified as possible, ranging from an information technology (IT) firm like IBM to a consulting firm like Boston Consulting Group.

Two pilot studies were conducted for us to refine the experimental tasks and experiment design. During the real experiment, 30 subjects, mostly third-year students from the School of Business at the University of Hong Kong, were invited to participate; each subject was required to search and analyze the Web communities of three out of the six firms using the three different analysis approaches. The three analysis approaches are summarized as follows:

1. Redips: The subject would use all the search and analysis capabilities in Redips to perform the task.
2. Backlink search engines: The subject would choose to use one or more of the three search engines given (Google, Altavista, and MSN Search) and would be free to use all the search and analysis capabilities in these search engines to perform the task.
3. Manual browsing: The subject would use the Internet Explorer as the Web browser and would start browsing from the Web site of the firm and follow any links on the site to perform the task.

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 and a business school graduate majoring in information systems were invited as the expert judges for this experiment. The judges were given the definition of themes in our evaluation. Each judge then identified and browsed the Web sites of the Web communities of each firm using a combination of the three approaches, and individually summarized the results into a number of themes. The themes from the two judges were then combined into a single set. These themes formed the basis for evaluation and measurement of performance, which is discussed below.

Performance Measure

The experiment examined both quantitative and qualitative data. Our primary interests for quantitative data were in the performance and efficiency of the analysis approaches. Performance was evaluated by theme-based precision, recall, and *F*-measure (Chen et al., 2001), whereas efficiency was measured by the analysis time of the subjects. Qualitative data were drawn from user search logs and questionnaires results.

Precision rate was the proportion of identified themes that were actually relevant. Recall rate was the proportion of relevant themes retrieved. Precision rate (P) and recall rate (R) were calculated using the following formulas, respectively:

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

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

A tradeoff between precision and recall forms an inverse relationship between precision and recall. A combination of precision and recall in one measure is hence useful in the evaluation. The *F*-measure (Van Rijsbergen, 1979) is one widely used measure that serves this purpose. The *F*-measure was calculated and computed in the following formula.

$$F\text{-}measure = \frac{(b^2 + 1)PR}{b^2 P + R},$$

where b is the weighting between recall and precision. Our experiment used $b = 1$, the most popular value used signifying that recall and precision are equally weighted in the *F*-measure.

Analysis time was recorded as the total duration spent on search and analysis, including both the response time of the system and subjects' browsing and analysis time. The time the subjects took to write their answers on the answer sheet was included in the analysis time as well.

The experimenter also recorded user search logs for observations of user behaviors as well as the user's think-aloud disclosure during the experiments. After the experiment, each subject was asked to fill in a questionnaire, which was designed to evaluate and compare the usability of the three analysis approaches.

Experiment Results and Analysis

Performance

The quantitative results on theme-based precision, recall, and *F*-measure of the three approaches are shown in Table 1. Paired *t* tests were also conducted to investigate whether any statistically significant differences were found among the three approaches. The *t* tests results are summarized in Table 2.

TABLE 1. Experiment results.

	Redips		Backlink search engines		Manual browsing	
	M	Variance	*M*	Variance	*M*	Variance
Precision	0.598	0.057	0.468	0.095	0.422	0.050
Recall	0.390	0.025	0.237	0.018	0.262	0.016
F-measure	0.465	0.032	0.294	0.024	0.311	0.021

TABLE 2. *p*-Values of pairwise *t*-tests on effectiveness.

	Redips vs. backlink search engines	Redips vs. manual browsing	Backlink search engines vs. manual browsing
Precision	0.1080	0.0044**	0.4560
Recall	0.0005**	0.0008**	0.4460
F-measure	0.0008**	0.0004**	0.6410

**The difference is significant at the 1% level.

DOI: 10.1002/asi

According to Table 1, the precision (0.598), recall (0.390), and *F*-measure (0.465) of Redips were all better than that of the backlink search engines approach. The *t*-test values showed that Redips performed significantly better than the backlink search engines approach in recall ($p < .001$) and *F*-measure ($p < .001$), whereas the difference in precision is not significant ($p = 0.108$). From the statistical results Hypothesis 1 was supported in general.

In testing Hypothesis 2, we found that Redips also performed well compared with the manual browsing approach. The mean precision, recall, and *F*-measure of Redips were all significantly statistically higher than that of manual browsing ($p = .0044$, .0008, and .0004, respectively). Hypothesis 2 was supported. In other words, the experimental results supported that Redips demonstrated significantly better effectiveness than both the backlink search engines approach and the manual browsing approach in searching and analyzing the Web communities of a firm.

There are several reasons why Redips excelled in precision, recall, and the *F*-measure. First, Redips has the ability of meta-searching and provides additional features that help improve the quality of the search results. As described earlier, meta-backlink searching can leverage the capabilities of multiple backlink search engines and provide a simple, uniform user interface for users to perform searches more effectively. This greatly improved the precision and recall of the retrieved Web communities. Redips also allowed subjects to enter one or more keywords to be included in the returned Web pages to extract the backlinks that are more relevant. This feature helped increase the quality of the result set of Web pages, generating themes that were more related to the Web communities. The filtering feature was also useful in filtering Web pages that no longer exist or Web links on the same Web domain. This feature also served to increase the precision of the resulting Web pages. These additional features differentiated Redips from the other two approaches in which these features were not available.

Second, clustering techniques like Arizona noun phraser and SOM helped users narrow down the search scope and focus on the Web communities of interest. When viewing the analysis results from the noun phraser and the SOM, subjects could click on any Web communities to discover a subset of Web documents that focus on the Web communities of interest. This helped the subject to decide if the Web communities were of interest to the firm and improved search precision.

Third, the interactive user interface of Redips made it possible for the subjects to focus on the Web communities of interest. The SOM technique used in Redips played an important role: SOM was a summary to group the incoming Web links into the Web communities. The results helped the subject to focus on the most frequent topics in the backlinks that form the Web communities of the firm.

Finally, the two other methods suffered several shortcomings. By manually browsing the Web, the subjects could only browse the content that could be directly linked from the firms' Web site. This set of sites could be focused on internal links within the site or favorable sites such as the firm's partners. However, the sites of other groups such as the media, pressure groups, or customers would not be directly accessible and could be easily overlooked by the subjects. For subjects using backlink search engines, we suggest that they did not perform as well as Redips because they often could only rely on one search engine at a time. We observed that most of them only used one backlink search engine in their tasks, rather than using multiple search engines. In addition, as the search engines did not provide any analysis on the search results, the subjects had to categorize the search results manually. These factors have limited the quality of their search performance.

Efficiency

Another focused aspect was the efficiency of the different approaches. It was measured by the search and analysis time of the subjects. The results are summarized in Table 3. Pairwise *t* tests were also performed and the results are shown in Table 4.

The mean analysis time of Redips was 204 seconds, which was higher than that of both the backlink search engines (168 seconds) and manual browsing approaches (128 seconds). The result is contrary to Hypotheses 3 and 4, in which we suggested that Repids would require shorter time than the other two approaches. The analysis time of the manual browsing approach was significantly shorter than that of Redips ($p = .0017$) and the backlink search engines approach ($p = .0043$), but the difference between Redips and the backlink search engines approach was not significant ($p = .0563$).

The hypotheses on efficiency were rejected due to the fact that Redips used a lot of time in fetching the full text of the URLs returned by the underlying backlink search engines and performing postretrieval filtering and analysis. The subjects as well had comments on the analysis time used and recommended improvements in the speed issues. However,

TABLE 3. Analysis time.

	Redips		Backlink search engines		Manual browsing	
	M	Variance	*M*	Variance	*M*	Variance
Analysis time (seconds)	204	17644	168	6677	128	1860

TABLE 4. *p*-Values of pairwise *t*-tests on efficiency.

	Redips vs. backlink search engines	Redips vs. manual browsing	Backlink search engines vs. manual browsing
Analysis time	0.0563	0.0017**	0.0043**

**The difference is significant at the 1% level.

DOI: 10.1002/asi

we found that the most time was spent on the searching and actual fetching of the complete documents of Web pages. These documents made the advanced analysis like noun phrasing and SOM possible and subjects only need to browse the verified and summarized results instead of manually going through the whole process. The findings concur with the results in our previous analysis in which subjects were not able to save a lot of time when using a tool that fetched the full content of search results, but were able to obtain improved performance in search effectiveness (Chen et al., 2002).

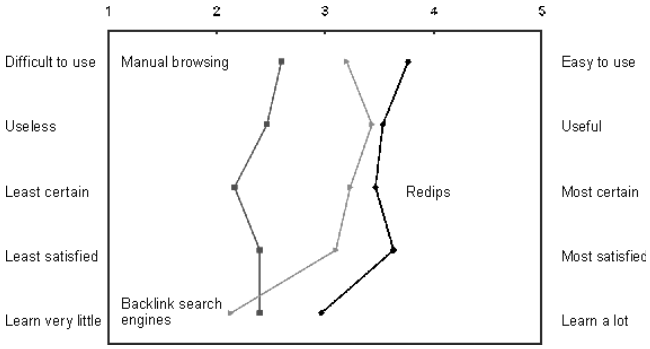

FIG. 7. Profile analysis of the three analysis approaches.

Questionnaire Results

The questionnaire was designed primarily to discover users' attitudes and subjective experience with the three analysis approaches. The questions were designed to evaluate and compare the analysis approaches on five dimensions: (a) user interface, (b) usefulness of the information retrieved in answering the analysis questions, (c) subjects' level of certainty about their answers, (d) user satisfaction of the analysis experience, and (e) the amount of knowledge obtained after the analysis.

The measurement scale used was a 5-point Likert scale, in which the numbers used to rank the objects also represent equal increments of the attribute being measured. The results of the questionnaire are summarized in Table 5 and *t*-test results are given in Table 6. The profile analysis of different approaches was plotted using the mean ratings for each approach on each scale and is depicted in Figure 7.

TABLE 5. Results from questionnaires.

	Redips		Backlink search engines		Manual browsing	
	M	Variance	*M*	Variance	*M*	Variance
Ease of use	3.77	0.530	2.60	1.42	3.20	1.06
Usefulness	3.53	0.533	2.47	1.02	3.43	1.22
Certainty	3.47	0.602	2.17	1.11	3.23	0.737
Satisfaction	3.63	0.585	2.40	1.08	3.10	0.714
Knowledge obtained	2.97	1.14	2.40	1.08	2.13	1.09

TABLE 6. *p*-Values of pairwise *t* tests on questionnaire results.

	Redips vs. backlink search engine	Redips vs. manual browsing	Backlink search engine vs. manual browsing
Ease of use	<0.0001**	0.0384*	0.0505
Usefulness	<0.0001**	0.6690	0.0048**
Certainty	<0.0001**	0.3050	0.0020**
Satisfaction	<0.0001**	0.0332*	0.0200*
Knowledge obtained	0.0325*	0.0052**	0.3860

*The difference is significant at the 5% level.
**The difference is significant at the 1% level.

The data confirmed that Redips scored higher in ease of use (3.77) than backlink search engines (2.60) and manual browsing (3.20). The differences between Redips and the two benchmark approaches were both statistically significant at the 1% level. Hypothesis 5 and Hypothesis 6 were confirmed, suggesting that it is easier to search Web communities of a firm using Redips than using only backlink search engines or manual browsing.

Not only did Redips score higher in ease of use, it also scored higher in all the four other dimensions: usefulness, certainty, satisfaction, and knowledge obtained. Most differences were statistically significant at the 1% or 5% level when compared with the other two approaches, with exceptions in usefulness and certainty when compared with manual browsing. The subjects generally ranked Redips as the best approach among the three.

The user interface and visualization feature of Redips were the dominant themes in the verbal analysis and questionnaire comments analysis. The subjects were especially interested in the SOM and noted that it is "helpful for searching the Web communities" of the firm. One subject stated, "SOM map is a very impressive searching analyzer." Another subject further suggested it would be even more useful if multilevel clustering was incorporated into the SOM in the tool. On the other hand, we also found that not all subjects were confident in using these advanced analysis tools. One subject said, "I'm paranoid about the program hiding information that I might actually find useful."

Limitations

One of the limitations of this research lies in the identification of themes that describe the Web sites of the Web communities of each firm. In our research, we invited two expert judges for the experiment and they were asked to identify the themes using the three methods being evaluated. However, one should note that Redips was the only method that provided textual analysis (like noun phrasing) to generate themes. In the other two methods, the judges had to utilize their own knowledge to generate themes. The reason that Redips performed better in the experiment may be solely

DOI: 10.1002/asi

due to such ability rather than the meta-searching component. Consequently, one should note that our analysis could be largely dependent on the experiment condition and the performance of the two judges.

Conclusions and Future Directions

Seeing a demand for a tool for strategic business intelligence application in Web community analysis, we designed a new tool called *Redips* that aimed to help analysts to work more efficiently. An experiment was conducted to evaluate the performance of the new tool and the experimental results were encouraging. We found that Redips achieved a higher precision, recall, and *F*-measure in searching and analyzing Web communities than general backlink search engines and manual browsing. Our results also showed that subjects found it easier to search the Web communities of a firm using Redips rather than the other two benchmark approaches.

We expect that Redips would be useful in the business intelligence analysis process. Redips can help analysts perform a comprehensive analysis of the environment and find information about the external environment of the firm on the Internet. Our study shows that Redips can use backlink search and analysis techniques to identify Web communities, which are difficult to obtain using traditional analysis methods. With additional knowledge about the environment of organizations and the firm's Web communities like suppliers, customers, competitors, regulators, and pressure groups, analysts can better formulate a firm's strategic planning process, which can, in turn, create added value for the firm.

One of the future research areas would be to study how Redips would perform when applied in real business situations. We are currently planning to conduct a case study with business managers and analysts to investigate in detail how this tool could be used in their business intelligence analysis process and whether other applications are possible. We also plan to enhance the analysis and filtering capabilities of the tool. For example, it may be possible to incorporate advanced link analysis techniques like PageRank or HITS into Redips. Furthermore, we also plan to perform keyword-based filtering by submitting a query that combines links and keywords to the search engines, instead of downloading the actual documents to the local client computers. Such analysis and filtering may further improve the efficiency of the tool in the analysis process.

Acknowledgments

The Redips project has been supported in part by the following grants: HKU Seed Funding for Basic Research, "Searching the World Wide Web Backwards for Business Intelligence Analysis," 10206086 (PI: M. Chau), January 2005–December 2006; HKU Seed Funding for Basic Research, "Using Content and Link Analysis in Developing Domain-specific Web Search Engines: A Machine Learning Approach," 10205294 (PI: M. Chau), February 2004–July 2005; NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473 (PI: H. Chen), April 1999–March 2002.

We would also like to thank the domain experts and the participants for their help in our experiment.

References

Anick, P.G., & Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. In the Proceedings of the 20th Annual International ACM SIGIR conference on research and development. New York: ACM.

Brill, E. (1993). A corpus-based approach to language learning. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Brin, S., & Page, L. (1998, May). The anatomy of a large-scale hypertextual web search engine. Paper presented at the 7th International World Wide Web Conference, Brisbane, Australia.

Chakrabarti, S., van den Berg, M., & Dom. B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. In Proceedings of the Eighth International World Wide Web Conference, Toronto, Canada. Elsevier.

Chau, M., Zeng, M., & Chen, H. (2001). Personalized spiders for Web search and analysis. In Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01) (pp. 79–87). New York: ACM.

Chau, M., Zeng, D., Chen, H., Huang, M., & Hendriawan, D. (2003). Design and evaluation of a multi-agent collaborative Web mining system. Decision Support Systems, 35(1), 167–183.

Chau, M., & Chen, H. (2003). Comparison of three vertical search spiders. IEEE Computer, 36(5), 56–62.

Chen, H., Chau, M., & Zeng, D. (2002). CI Spider: A tool for competitive intelligence on the web. Decision Support Systems, 34(1), 1–17.

Chen, H., Fan, H., Chau, M., & Zeng, D. (2001). MetaSpider: Meta-searching and categorization on the web. Journal of American Society for Information Science & Technology, 52, 1134–1147.

Chen, H., Fan, H., Chau, M., & Zeng, D. (2003). Testing a cancer meta spider. International Journal of Human–Computer Studies, 59, 755–776.

Chen, H., Schufels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. Journal of Visual Communication and Image Representation, 7, 88–102.

Chung, W., Chen, H., & Nunamaker, J.F. (2003, January). Business intelligence explorer: A knowledge map framework for discovering business intelligence on the web. Paper presented at the 36th Annual Hawaii International Conference on System Sciences (HICSS-36), Big Island, Hawaii.

Chung, W., Chen, H., & Nunamaker, J.F. (2005). A visual knowledge map framework for the discovery of business intelligence on the web. Journal of Management Information Systems, 21(4), 57–84.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in knowledge discovery and data mining (pp. 1–36). Cambridge, MA: MIT Press.

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95) (pp. 112–117). Menlo Park, CA: AAAI.

Gilad, B., & Gilad, T. (1988). The business intelligence system. New York: AMACOM.

Google. (2004). Google Web APIs—Home. Retrieved January 13, 2005, from http://www.google.com/apis/

Google. (2005). Google web search features. Retrieved January 6, 2005, from http://www.google.com/help/features.html

Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann.

Hearst, M.A. (1997, July). Text data mining: Issues, techniques, and the relationship to information access. Paper presented at the UW/MS Workshop on Data Mining, Berkeley, CA.

DOI: 10.1002/asi

Johnson, R.J. (1994). A cognitive approach to the representation of managerial competitive intelligence knowledge. Unpublished doctoral dissertation, The University of Arizona, Phoenix.

Keiser, B.E. (1987). Practical competitor intelligence. Planning Review, 8, 14–18.

Kohonen, T. (1995). Self-organizing maps. Berlin: Springer-Verlag.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1998, April). Trawling the web for emerging cyber-communities. Paper presented at the 8th International World Wide Web Conference, Brisbane, Australia.

Lin, X. (1997). Map displays for information retrieval. Journal of the American Society for Information Science, 48, 40–54.

Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'91) (pp. 262–269). New York: ACM Press.

Orwig, R., Chen, H., & Nunamaker, J.F. (1997). A graphical self-organizing approach to classifying electronic meeting output. Journal of the American Society for Information Science, 48(2), 157–170.

Prescott, J.E., & Smith, D.C. (1991). SCIP: Who we are, what we do. Competitive Intelligence Review, 2(1), 3–5.

Reid, E.O.F. (2003, January). Identifying a company's non-customer online communities: A proto-typology. Paper presented at the Hawaii International Conference on System Sciences, Big Island, Hawaii.

Salton, G. (1986). Another look at automatic text-retrieval systems. Communications of the ACM, 29, 648–656.

Schermerhorn, J.R. (2001). Management. New York: Wiley.

Selberg, E., & Etzioni, O. (1997). The MetaCrawler architecture for resource aggregation on the Web. IEEE Expert, 12(1), 11–14.

Sutcliffe, A.G., Ennis, M., & Hu, J. (2000). Evaluating the effectiveness of visual user interfaces for information retrieval. International Journal of Human–Computer Studies, 53, 741–763.

Tan, A.H. (1999, April). Text mining: The state of the art and the challenges. Paper presented at the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Beijing, China.

Tolle, K., & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science, 51, 352–370.

Van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths.

Vedder, R.G., Vanecek, M.T., Guynes, C.S., & Cappel, J.J. (1999). CEO and CIO perspectives on competitive intelligence. Communications of the ACM, 42(8), 109–116.

DOI: 10.1002/asi