EBizPort: Collecting and Analyzing Business Intelligence Information

Byron Marshall, Daniel McDonald, Hsinchun Chen, and Wingyan Chung
Artificial Intelligence Lab, Management Information Systems Department, University of Arizona, Tucson, AZ 85721. E-mail: {byronm, dmm, hchen, wchung}@eller.arizona.edu

To make good decisions, businesses try to gather good intelligence information. Yet managing and processing a large amount of unstructured information and data stand in the way of greater business knowledge. An effective business intelligence tool must be able to access quality information from a variety of sources in a variety of forms, and it must support people as they search for and analyze that information. The EBizPort system was designed to address information needs for the business/IT community. EBizPort's collection-building process is designed to acquire credible, timely, and relevant information. The user interface provides access to collected and metasearched resources using innovative tools for summarization, categorization, and visualization. The effectiveness, efficiency, usability, and information quality of the EBizPort system were measured. EBizPort significantly outperformed Brint, a business search portal, in search effectiveness, information quality, user satisfaction, and usability. Users particularly liked EBizPort's clean and user-friendly interface. Results from our evaluation study suggest that the visualization function added value to the search and analysis process, that the generalizable collection-building technique can be useful for domain-specific information searching on the Web, and that the search interface was important for Web search and browse support.

Introduction

Business intelligence (BI) has been defined as the acquisition, interpretation, collation, assessment, and exploitation of business-related information (Chung, Chen, & Nunamaker, 2003a).The primary objective of a BI system is to support sound decision making. However, gathering business intelligence from the vast set of available resources can be a challenge. Key issues include the rapid rate of change in the business and information technology (IT) environment, the often-referenced problems of information overload (Bowman, Danzig, Manber, & Schwartz, 1994), and the questionable quality of many resources available on the Web. Appropriately adapted collection building and metasearching techniques developed in the digital library community hold promise to help address the challenges of the BI domain. In this report, we review the characteristics of existing techniques from a business intelligence perspective, explore a method of collecting quality resources, and test an interface designed to support BI activities.

Once quality resources have been gathered, users need to interpret, assess, and exploit the knowledge embedded in those resources. While intelligent analysis tools and other forms of post-retrieval processing may support these processes, the human-computer interaction aspects need to be studied. Analysis is a complex human task but most search tools provide relatively simple search interfaces. This work seeks to identify the kind of user interface elements that can enhance human analysis of document lists returned from a query, and to assess the usefulness of a proposed document search interface in supporting Web information seeking in the business/IT domain.

Literature Review

Business organizations need to stay on top of current information and developments in their field. Top executives spend a large portion of their time scanning for strategic information and rely on external sources for the information they need (Vedder, Vanecek, Guynes, & Cappel, 1999). Interestingly, smaller organizations scan the Internet more frequently than larger ones. Increased frequency of Internet scanning increases the effectiveness of environmental scanning (Tan, Teo, Tan, & Wei, 1998). Even when frequent scanning is conducted, low recall rates, outdated indexes, and limited analysis tools have been identified as impediments to effective information search (Chau, Zeng, & Chen, 2001). The use of improved information tools and collections would increase the effectiveness of the time spent on these scanning activities.

Accepted November 7, 2003

 ● Published online 25 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20037

Fuld & Company prepare an annual review of competitive intelligence (CI) software (Fuld, Sawka, Carmichael, Kim, & Hynes, 2002). In their 2002 Intelligence Software Report, they evaluated 13 Competitive Intelligence tools using a range of criteria and came to several conclusions. Good CI tools should handle multiple information repositories stored in various formats. Tools should be able to take advantage of various information repositories through metasearch, where a single query is passed to multiple search engines in multiple languages on the Internet. Tools should support comprehensive search capabilities such as phrase searching and searching with date ranges. Other functional effectiveness criteria for CI tools include filtering out noisy or irrelevant content through controlled Web-spidering, automatic filtering, relevance ranking, and dynamic summarization of articles and documents. Fuld considers support for information analysis to be the most lacking functionality in currently available competitive intelligence tools. These criteria identify important features and characteristics appropriate for the gathering of competitive intelligence.

Digital library and information retrieval researchers have studied techniques that address many of the challenges identified for the business intelligence domain. The following section of this literature review focuses on the gathering of appropriate resources. Information Seeking for Business Intelligence provides background material on the information search process.

Gathering Appropriate Resources

Content quality is a major contributor to overall system quality in a business intelligence tool. In this section, we review vertical search engines, describe techniques used to collect relevant documents, and present background information related to the measurement of information quality. Several techniques have been developed to build searchable, domain-specific, or vertical collections. Some of these approaches are described and considered in Building Vertical Collections. One alternative to collection-building is metasearch. A metasearch tool accepts user queries and passes them to external information sources for processing. Metasearch reviews the metasearch approach to gathering quality resources. Vertical collection building and metasearch processes aim to gather relevant resources. Relevance evaluation is usually based on some measure of topicality. Resources are selected because they are about the topic of interest. However, other measures of information quality may also be employed to gather useful resources. Information Quality and Business/IT Information Providers expands the notion of information quality and surveys existing types of business/IT information providers.

Building vertical collections. Vertical search engines provide access to information on specific segments of the documents available on the Web. Vertical search engines provide customized functions and more precise results when compared to general search engines like Google or AltaVista. One research group estimates the number of special purpose search engines at a quarter of a million sites (Bergman, 2001). Numerous small collections have been created using human-selected documents or content from published works. Human selection of documents is an expensive and increasingly inadequate process as the number of pages found on the Internet is counted in billions. Many techniques have been developed to collect documents for vertical search engines.

Vertical collections can be created using a number of different automated approaches. Automated approaches use document collection mechanisms, often called spiders, to access pages. Spiders fetch documents from a list of starting or "seed" URLs. They identify additional page references in the document and often save the resources for further processing. This kind of "spidering" or Web-crawling process uses the structural information embedded in the Web to identify additional, potentially appropriate resources. Spiders follow links found in one page to additional pages and additional links. Sometimes, these spiders are guided by algorithms that assess the potential relevance of new links before fetching the pages.

Collection algorithms can exploit content and/or structural information found in the documents to identify appropriate pages. Examples of the use of structural information to evaluate resources include PageRank (Brin & Page 1998), which uses inlink analysis to rank results, and the HITS algorithm (Kleinberg, 1999), which identifies hubs and authorities. Although some use of structure is common in collection-building tools, the use of document content to recognize relevant documents is crucial to most vertical collection building approaches. A few examples of vertical collection building systems are listed here. The Deadliner system uses formatting and key word heuristics to recognize desirable research documents (Kruger et al., 2000). The MERCATOR system allows for various controlling functions to be implemented (Najork & Heydon, 2001). It was used to test a focused-crawling collection-building approach that simultaneously creates a set of collections by finding documents clustered around term and document vector centroids (Bergmark, 2002). Chau and Chen (2003) compare breadth-first search, Page Rank, and Hopfield Net spiders' effectiveness in building vertical collections in a medical domain. These and other collection techniques are intended to identify focused lists of documents related to a topic or domain of interest.

Metasearch. Metasearching identifies relevant documents through unified access to multiple search engines. A metasearch tool processes queries from users by formatting the requests and passing them on to other search engines. The results are collated and presented to a user. Metasearch tools are intended to increase search coverage and improve search scalability (Meng, Yu, & Liu, 2002). There are billions of documents available on the Internet. Because even the larg-

est of the general purpose search engines index only a fraction of the World Wide Web (Kobayashi & Takeda, 2000), many search engines have been created and are available on the Internet. Metasearching aims to link distributed collections of information. Meng et al. (2002) identify a number of challenges to be faced when developing metasearch systems, including integrating local systems that use different indexing techniques and query types, discovering knowledge about component search engines, developing effective merging methods, and deciding where to place the software components of a metasearch engine. Stated simply, metasearch tool developers have to address the difficulties of accessing and merging information from disparate sources.

In order to provide access to external information, a metasearch system developer must acquire an understanding of the functionality of the target information search engines. Search engine functions can be grouped into three categories (Huang, Ulrich, Hemmje, & Neuhold, 2001). Classification selection controls manage the sources or classes of documents to consider. Result display controls customize the format, size, or sorting of results. Query input controls accept terms, term modifiers, and logical operators. Metasearch tools must be flexible enough to match the variety of different search interfaces. One interesting system intended to help in establishing metasearch interfaces is the ACQUIRE system, which moves towards automatically extracting and maintaining a metasearch interface to search tools. It used the information it saves to generate its own user interface (Huang et al., 2001). The National Science Digital Library (NSDL) project aims to create widely accepted standards for interfacing with search portals (Lagoze et al., 2002). But for now, metasearch system developers are faced with a tremendous variety of search functions, input formats, and output formats.

The second challenge for a metasearch system, integration of results, has been addressed in the literature using source selection and result merging algorithms. Systems have been developed using several classifications of algorithms to select the target search sites for a query. Site selection approaches have included rough representative, statistical representative, and learning-based approaches. The information about available collections is used to direct a query to appropriate sources and to select the number of documents to return from each of those sources. Meng, Wu, Yu, & Li (2001) review a number of these approaches. As an alternative to merging metasearch results based on heuristics and metadata, the results lists from several sources can be merged by fetching, indexing, and ranking each document. One example of this alternative approach is seen in MetaSpider (Chen, Fan, Chau, & Zeng, 2001). Metaspider uses a metasearch, fetch, and categorize technique to present integrated metasearch results to users.

Merging results from various sources is a difficult task. Information about the local ranking of a document relative to a query for a particular database can be used but is not always available. When it is available, it is not in a consistent format from source to source. No metasearch result merging approach has emerged in the literature or in industry practice as a clear favorite. Identifying a superior merging algorithm depends on identifying appropriate resource ranking measures.

Information quality and business/IT information providers. Additional measures beyond traditional term relevance rankings can be considered when gathering results to present in response to a user query. For example, credibility is important when a system acts as a knowledge source or provides decision aids (Fogg & Tseng, 1999). Many factors may influence a user's credibility decision, but researchers generally agree that trustworthiness and expertise are key components. Trustworthiness relates to the degree to which a Web site is truthful and unbiased while expertise relates to the experience and competence of those publishing on the Web site. Credibility, as demonstrated in trustworthiness and expertise, is important in this application because one of the main objectives of BI systems is providing support for decision making.

While assessing a Web page's information quality is considered an inexact science, useful metrics have been used in past research to evaluate a site's information. Based on the framework of information quality (Wang & Strong, 1996), Katerattanakul and Siau (1999) define several useful characteristics of high-quality pages. Different information sources can be rated by their intrinsic quality and contextual quality. Intrinsic quality is interpreted to mean the accuracy of the contents and the accuracy of the hyperlinks contained in the Web page. This is measured by the lack of errors in the page content and the meaningful placement of working hyperlinks. The contextual quality of a Web site relates to the parties of the communication and the effect of the communication. Contextual quality improves when there is a clear identification of the author (Hlynka & Welsh, 1996). Contextual quality is also benefited when the author or Web site has some level of authority. These notions of information quality are seen in varying degrees in the products of existing business/IT information providers. Business/IT information providers can be loosely grouped into 4 categories: (1) content-focused sites, (2) news magazine portals, (3) information management tool builders, and (4) general search engines.

- *Content-focused sites:* These sites create and collect quality information resources and provide access to them, often for a fee. Factiva, a Reuters company, is one example of this category of provider. It provides an interface to Reuters' news. Much of the value of these sites comes from the special information it makes available to subscribers, information that is not available to the general public. These sites are popular sources of information in the business IT community. In Fogg and Tseng's (1999) terms, perceived credibility of information found on these sites comes from the expertise and special knowledge of the content creators.
- *News magazine portals:* Several IT news magazines have

publicly available, searchable Web portals, which provide access to the content published in their magazines. Computerworld and IDG are two examples of this category. News magazine sites provide search interfaces to content generated by a staff of reporters. Articles provided by these sites are generally attributed to an author and presented with a publication date. These characteristics contribute to the contextual quality of the resources. The trustworthiness of these sites is sometimes questioned by readers who believe that there is a bias towards certain vendors.

- *Information management tool builders:* These software companies provide tools to help users or companies deal with information search needs. Copernic (www.copernic.com), for instance, sells enterprise searching and indexing software solutions for companies as well as desktop software to support user searches. Copernic also provides a dynamic summarization tool intended to help users more quickly identify relevant documents. Some of the features described in Fuld et al. (2002) are available in this provider category.
- *General search engines:* Google, Yahoo, and AltaVista are representatives of this category. These well-recognized service providers build very large collections and provide a search interface. While they process millions of requests for users all over the world daily, they have several significant limitations in the business intelligence domain. They are susceptible to search engine persuasion and spamming (Bergmark, 2002). They cover only a fraction of the Web and can contain many low-quality documents (Meng et al., 2001). They also do not provide interface elements to support analysis of returned lists of documents. In spite of these limitations, these resources form an important tool for business/IT managers because they provide access to a larger selection of resources than those found in specialized sites.

Content gathering for BI. While the collection and integration of resources is a difficult task in any domain, the BI domain highlights some weaknesses in existing document-gathering techniques. High quality resources are maintained in a number of news magazine and content-focused portals on the Internet. However, because many of the content providers compete for readers or subscribers, they are not adequately cross-linked. The vertical collection building literature shows that most vertical collection algorithms rely on the structural information provided in embedded cross-links to arrive at new pages. Therefore the lack of cross-linking between BI sources impacts the effectiveness of most of the previously described spidering techniques. Also, we observe that IT/Business news magazine articles are not frequently returned when using general search engines. Perhaps this is because competing information resource providers rarely link to each other so that the structural information used by general search engines does not tend to favor these credible, high quality resources. The competitive environment also reduces the effectiveness of many metasearch algorithms. Commercial sites do not provide the kind of detailed collection statistics that are used by many result-merging algorithms.

Once appropriate methods have been constructed to identify and access quality information, people need to interact with the documents to perform the complex searching and analysis tasks that make the information useful for decision making. Thus, it is useful to review past research into human information seeking.

Information Seeking for Business Intelligence

Information seeking, as studied in previous research, typically adopts a process model. The process consists of various stages of problem identification, problem definition, problem resolution, and solution presentation (Wilson, 1999). Variations of the process model also can be found in the literature (Bates, 1989; Kuhlthau, 1991; Sutcliffe & Ennis, 1998). Bates' model for information search, called "berrypicking," captures the idea of an evolving multistep search process as opposed to a system that supports submitting single queries alone. Kuhlthau found that High School students began research assignments by conducting general browsing and performed more directed search as their understanding of the subject increased. Sutcliffe and Ennis succinctly summarized four main activities in their process model of information searching: problem identification, need articulation, query formulation, and results evaluation.

In addition to directed searching, browsing the Internet is another strategy that information seekers frequently employ. Marchionini and Shneiderman (1988) defined browsing as "an exploratory, information seeking strategy that depends upon serendipity." Spence (1999) defined "browse" as the registration of content into a human mental model. Having compared various definitions, Chung, Chen, and Nunamaker (2003a) defined "browsing" as an exploratory information seeking process characterized by the absence of planning, with a view to forming a mental model of the content being browsed.

The Internet has recently evolved into a major information-seeking platform and was found to be one of the top five information sources for business analysis (The Futures Group, 1995). Information seeking on the Web has been characterized by its different depths of analyses. Chung et al. (in press) conducted an experiment to study the use of meta-searching, summarization, and categorization to help business analysts obtain business intelligence. They found that summarization and categorization were helpful in searching and browsing business information, and that their system could significantly augment existing search engines. In an intensive two-week study of Web-use activities, Choo, Detlor, and Turnbull (2000) found that knowledge workers engaged in a range of complementary modes of information seeking in their daily work. The 34 study participants came from 7 companies and held jobs as IT technical specialists or analysts, marketing staff, consultants, etc. They primarily utilized the Web for business purposes. The study confirmed that knowledge workers performed such multiple analyses as browsing, differentiating, monitoring, and extracting in the business domain. Although the study suggests a behavioral framework for Web-based information

seeking, it did not provide a system-based solution to domain-specific information seeking on the Web. Both a searchable, high-quality, domain-specific collection and browsing supports such as summarization, categorization, and visualization are needed to effectively facilitate multiple analyses. Moreover, an advanced document search interface is needed to support more efficient information seeking on the Web. The next section reviews search interface literature to understand how interfaces can be adapted to provide the above mentioned supports.

User Interfaces for Information Seeking

In information-seeking environments today, "designers often fail to provide appropriate views of materials to give an overall sense of the structure and materials available" (Greene, Marchionini, Plaisant, & Shneiderman, 2000). Browsing tools should be provided along with query interfaces to help users navigate the information space and do so in a timelier manner. Many examples of browsing tools have been studied in the literature. Greene et al. suggest the use of previews and overviews. A preview acts as a surrogate for a single document of interest while an overview represents a number of documents. Overviews assist users in extracting the "gist more accurately and rapidly than traditional hit lists." An effective overview gives the information seeker an idea of the size and extent of the information represented by the overview, how documents in the overview relate, and what kinds of documents are missing.

Overviews. Browsing, or "overview," tools in information visualization aim to increase users' knowledge of their search space and increase the effectiveness and efficiency with which they search (Börner, Chen, & Boyack, 2003). Visualizing information is thought to improve the process of interacting with large volumes of data (Gershon, Eick, & Card, 1998; Card, Mackinlay, & Schneiderman, 1999; Chen, 1999; Spence, 2001). Documents (e.g., articles, Web pages, or patents) are the most common unit used in creating knowledge maps (Börner et al., 2003). Maps created from co-citation (Chen, 1999; Chen, Paul, et al. 2001), co-authorship (Mahlck & Persson, 2000), and co-word analysis (Bhattacharya & Basu, 1998) are also relatively common. Co-citation maps aid in inferring the intellectual structure of a field, co-authorship maps convey the social network of a discipline, and co-word analysis is used to understand the cognitive structure of a field (Börner et al., 2003). Document-level maps, however, have been used in tasks commonly associated with business intelligence, such as document retrieval, domain analysis (Small, 1999, 2000), and competitive intelligence analysis (Boyack, Wylie, & Davidson, 2002). Because document-level visualization is the most common in the area of information retrieval, we focus our review on visualization techniques that facilitate document visualization and analysis.

The Self-Organizing Map (SOM) classifies documents based on their text content using a neural network (Kohonen, 1995). The SOM algorithm automatically places documents into different regions on a map based on their similarity to one other. Adjacent regions are more similar than those regions far away on the map. The SOM is a document-based visualization tool that adds additional semantics by labeling the map. Lin, Soergel, and Marchionini (1991) first adopted the SOM for information visualization to document spaces. Their work visualized important concepts queried from a database. Chen, Houston, Sewell, & Schatz (1998) present a study using an SOM to visually categorize the Entertainment subcategory of Yahoo's directory. The usefulness of the SOM in browsing is compared to the usefulness of the human-generated Yahoo categories. The results show the tools performed comparably. The SOM has also been studied in general Internet search tasks and the SOM has been shown to work with very large collections (Chen, Schufels, et al., 1996).

Clustering techniques have also been used to reduce the dimensionality of the vector space model and group similar documents together. When producing the same number of clusters, Hearst (1999) showed that no one clustering algorithm outperformed the alternatives. Pirolli, Schank, Hearst, and Diehl (1996) use clustering to create groups of documents with similar topics. Automatic summaries are then generated at varying levels of granularity on the documents within the same cluster. The tree map approach (Shneiderman, 1992) has been used to determine the placement of clustered documents. Feng and Börner (2002) proposed semantic tree maps that rely on forced directed placement to calculate the placing of clusters.

Pathfinder networks have also been used to display the content similarities among documents. A pathfinder network uses pairwise similarity measures to place nodes in a graph. Distance measures can come from subjective computation or numerical analysis. Chen (1997) developed a generalized framework for visualizing information sources called *generalized similarity analysis* (GSA). GSA constructs pathfinder networks based on content similarity between pages. The time to generate a pathfinder network makes it impractical to use in a dynamic Web search setting.

Previews. Different from overviews, effective previews shown at the appropriate time give users enough information about a document to make an accurate relevance assessment (Greene et al., 2000). Usually, this information includes a list containing information about documents sorted by their relevance to the query. These lists usually contain the title of the document, followed by metadata, such as the publishing date of the article, and a short summary of the document. This type of information has been called a document surrogate (Witten, Nevill-Manning, McNab, & Cunningham, 1998). Document surrogates have been represented visually as well as in text. Hearst's (1995) tile bars visually represent a document's relation to the

query terms using rectangular bars. The rectangles are displayed in varying shades of gray to highlight the existence of query terms (Hearst, 1995). Other tools that visually represent document similarities to query terms include VIBE (Korfhage, 1991) and Lyberworld (Hemmje, Clemens, & Willett, 1994).

Document summarization tools create textual document previews. Indicative summaries are those intended to assist the user in making relevance decisions (Firmin & Chrzanowski, 1999). Query-biased summaries use the query-terms entered by the user to determine what piece of the document to use as a preview for the user (Sanderson, 1998). Such summaries, sometimes referred to as key-word-in-context (KWIC), can include full sentences around query terms or only sentence fragments. Query-biased summaries were found to be more useful as previews than the first two sentences of a document (Tombros & Sanderson, 1998). Different from query-biased summaries, generic summaries use document-specific heuristics, such as $tf \times idf$, and cue phrases to pick sentences that best represent the essence of the document (McDonald & Chen, 2002). Summaries are often combined with other metadata to provide document previews.

Research Questions

Our review of previous research and existing tools raises two important issues: (1) How can quality resources be identified and organized for retrieval? and (2) What kinds of tools can assist with the kind of searching and analysis done in this process?

The BI domain highlights some important issues related to vertical collection building and metasearching. Vertical collection-building methods generally rely on structural cross-link information to direct spiders to potentially relevant information. Effective metasearching, on the other hand, involves the merging of results from several sites. The information to do this merging can be difficult to obtain. Because cross-linking within the BI domain is limited and statistical collection information is largely unavailable, the effectiveness of traditional vertical collection techniques and metasearch methods is impaired.

Support for analysis functions has been identified as a weakness in existing BI tools (Fuld et al., 2002). One part of the BI process is scanning, which can be seen as a kind of browsing task. Overviews and previews have been identified as important browsing support tools. However, the particular usefulness of various kinds of search support interface modules should be evaluated in the BI context. The BI process would be more effective if there were a smaller gap between the complexities of human analysis and the inadequacies of Web document search interfaces.

To explore these issues we formed three research questions:

1. What kind of user interface elements can enhance human analysis of document lists returned in response to a query?
2. Can we develop an adaptable approach to building a high-quality domain-specific collection of Web documents?
3. To what extent is the document search interface and the high-quality, domain-specific collection useful to human beings who are engaged in Web information seeking in the business/IT domain?

Research Testbed

The EBizPort (Electronic Business Intelligence Portal) system was developed to provide a platform for exploring these research questions. EBizPort System Design and Implementation describes the design and implementation of the EBizPort system. EBizPort implements metasearching as a user interface and as a method of harvesting Web documents. It also integrates analysis, visualization, and summarization functionality, which may assist users as they seek for relevant knowledge that can be gleaned from available information. Subsequently, Brint.com is discussed as a benchmark that will be compared to EBizPort (also see Evaluation Methodology and Experimental Results and Discussion). Brint is a vertical search engine focused on the Business/IT domain, which provides unified access to a variety of information sources.

EBizPort System Design and Implementation

EBizPort was designed to support IT managers and other IT professionals. We reviewed available online portals and identified desirable information characteristics and search tool features from an IT manager's point of view. Among other tasks, IT managers negotiate with vendors and scan the environment for upcoming threats and opportunities. Based on the review of the tasks and existing tools (see Literature Review), we have identified the following requirements for our system:

- *Credible sources:* Our system attempts to direct users to credible sources. Information retrieved should be reliable and credible to support good decision making.
- *Timeliness:* Our system identifies the timeliness of the information retrieved. We want users to be able to control the currency of the information when searching.
- *Efficient browsing:* We want to save users time in browsing for information by providing only material relevant to the domain. Also, we want to provide tools to help users sift through documents, identifying those that are pertinent to the task at hand.
- *Integration of financial information:* We want to integrate some of the financial information available on the Internet. We want them to be able to easily connect companies in documents to financial and market information. Well-made IT investment decisions require a variety of types of information. Product reviews and industry trends are well covered in the magazines selected as primary sources in EBizPort. However, the financial strength of a potential vendor is often important. IT investments often amount to long-term part-

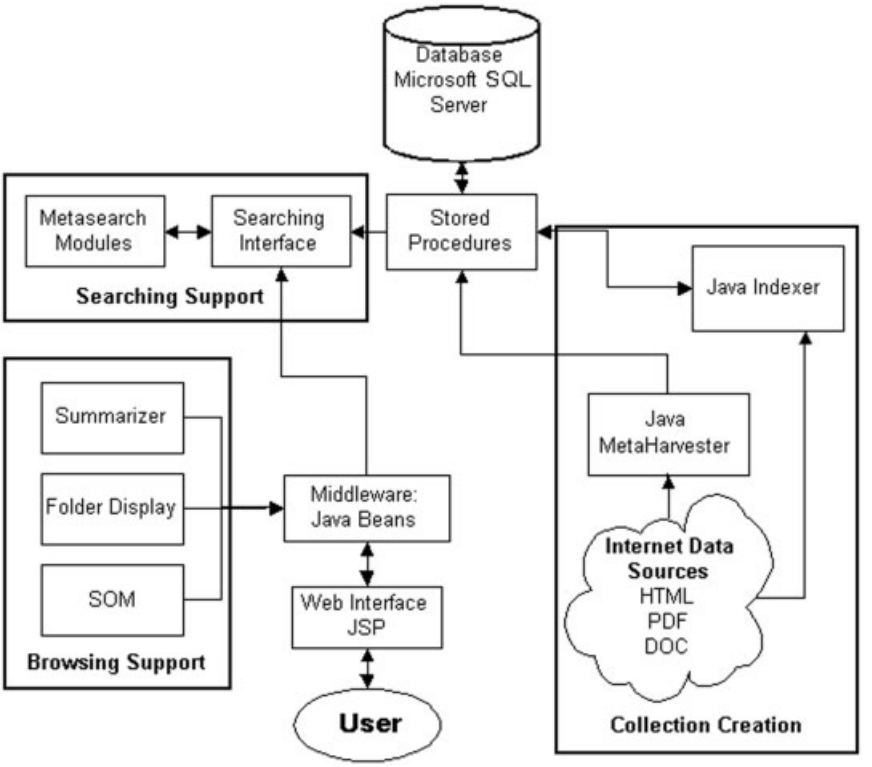

FIG. 1. EBizPort architecture.

nerships. Organizations want to choose partners who will be around in the future.

These requirements are focused on business IT users but we believe the general techniques identified will be applicable to other domains.

EBizPort architecture and components. With the constantly changing nature and relevance of business intelligence information, the architecture for a business intelligent system must be flexible and modular. These two ideals guided the design of the EBizPort architecture. The EBizPort site is built on a three-tier, client/server architecture as shown in Figure 1. The architecture is data centric with most modules interacting with the database. The three main functions provided in the architecture are collection creation, searching support, and browsing support.

The local collection portion of the EBizPort is created off-line. Content is collected by gathering URLs from site-specific search engines using a metaharvest program. The metaharvester written in Java connects to several Information Technology (IT) news sources and queries the collections using key words listed in a text file. A stored procedure stores the resulting URLs in the database. After harvesting the URLs, a Java indexer retrieves unindexed or out-of-date URLs from the database. The pages are fetched and parsed for phrases, key words, and publication date. The indexed data is saved in a text file, which is later added to the database using a stored procedure.

The user interacts with the search and browse interface elements. The search interface dispatches the query to all the search engines selected by the user. This includes both the local index and the metasearch modules. The middleware then combines the results and presents them to the user sorted by search engine and relevance. Browsing support is provided once search results have been returned. All browsing tools are dynamic and are generated when selected by the user. Browsing modules are called by the middleware allowing a clean interface for additional browsing modules in the future. The user can connect using any browser and submit searching and browsing requests to the system. Interactions with the database are conducted through stored procedures and utilize a database pool of reusable connections.

Collection creation. To address the limitations of existing collection building and metasearching techniques, we decided to implement a hybrid metasearch approach. The lack of structural links to help guide spiders to additional pages is addressed by passing queries to existing search portals. The items returned from these portals are then spidered and parsed to create a local index. The local index of many potentially relevant documents allows us to integrate results based on the term frequency characteristics seen in a large collection of documents. Access to a full index allows the use of traditional, effective ranking algorithms and heuristics. This kind of algorithm does not require the kind of

TABLE 1. EBizPort content sources.

Source	Description	Harvested URLs
Computerworld	Contains "IT technologies, trends, career topics and management issues."	71,730 pages
Industry Standard	Includes articles from Jan. 1998 through Sept. 2001.	38,881 pages
Wired	Delivers "analysis and resonant storytelling on high-tech and business topics."	40,682 pages
PC World	Searched through the Business2.0 interface	26,863 pages
InternetWeek	Includes content on supply chains, web development, security, and IT services. Includes TechWeb articles.	33,500 pages
Infoworld	Searched through the Business2.0 interface	30,696 pages
CNet	"Source for Computing and Technology"	26,775 pages
IDG	Offers content from over 300 technology magazines and newspapers.	37,955 pages
ITWorld	Provides news, educational content, and interactive media.	15,846 pages
CIO	Reports on major news and events that affect IT.	20,200 pages
Business 2.0	Contains "articles on the smartest, most innovative business practices."	23,276 pages
Informationweek	Searched through 1DG interface	20,200 pages
RedHerring	Analyzes the "driving forces affecting innovation, technology, and financial markets."	10,503 pages

statistical information that is needed by many metasearch result merging algorithms. That is important because such information is very difficult to collect at best. At the same time, this approach avoids some of the issues associated with other vertical collection building techniques. Notably, the lack of cross-linking between sites does not affect the collection process because the spiders do not follow any links found in the pages.

Based on conversation with IT professionals and business school professors and an extensive Web search, we identified 13 different sources for news and commentary in the IT domain, shown in Table 1. Contextual quality of these resources is enhanced because they provide author names and publication dates. Credibility is enhanced when the reporters and columnists are perceived to have substantial expertise. The inclusion of several different providers is intended to help address potential concerns about bias. While a user may perceive a particular magazine as biased, perhaps providing results from several sources will increase the chance that a user will encounter believable information. After initially including general search engines, we limited the content providers to commercial sites with name recognition in the IT industry. The content provided by these sites is usually managed using content management software and is not heavily linked to pages outside the site's domain. The small number of in-links to these pages means they do not rank high in the relevance rankings of search engines like Google. The high quality of these pages and their low representation in general search engines makes them good candidates for a domain-specific search engine.

To build our local collection, we arranged access to the 13 sites, generated a list of 635 queries, and implemented several document conversion modules. The queries contain words extracted from approximately 3 years of Computerworld back issues. A portion of the queries contain non-IT words in an attempt to expand our list of unique URLs. For example, certain key words that provide long lists of results in target systems are combined with the names of each of the 12 months. The 12 resulting queries return a larger set of unique URLs. Using these queries and 13 sources of IT news and commentary, we were able to gather and index over 400,000 URLs to provide broad coverage of business IT topics. Because business intelligence information is often in a format other than HTML on the Internet, our collection also includes Word, Excel, PDF, and text format files. Although our site does index the content, it does not actually serve up the articles. Instead, EBizPort directs the user to the commercial site to access the content, thereby not infringing on copyrights.

Web pages, including those from commercial content providers, are noisy. Advertisements, menu bars, and dynamic news content are mixed with the relevant business intelligence content on each page. In order to filter out the unwanted text, we used HTML parsing and sentence recognition routines to index only the content that appeared in a full sentence. This allowed us to ignore nearly all the irrelevant content without missing important information about a page. In addition, the collection index included the tag source of each term to record whether it appeared in a page's header, meta, title, or body tags. These tags are used to improve retrieval accuracy (Arasu, Cho, et al. 2001). Finally, we extracted the publish date of the article where that date was available in the meta tags. We also used date recognition techniques to extract the date from the page content itself when not available in the meta tags. Dates extracted were checked against the current date to avoid erroneous publication dates. In all, we extracted publication dates for nearly 50% of the pages contained in our local index.

Searching support. The EBizPort searching interface is shown in Figure 2. A search for the words "Dell" and "Expansion" is entered in the search box for 10 hits within the last year. The user can choose to search other sites as well as searching the local database. Figure 2 also shows the interface for selecting a few of the search sources. Results that are returned from the metasearch modules are ordered and listed by source, while the results from the local index have the advantage of having been merged and ranked together. As users metasearch the news sites, new URLs returned that are not in the local collection are added to the

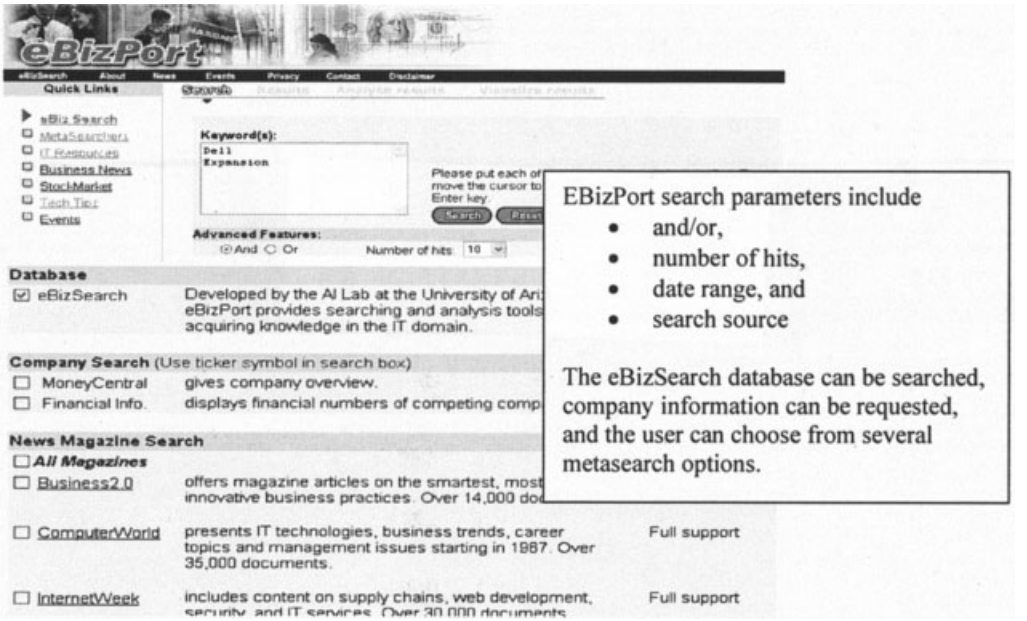

FIG. 2. Enter search parameters.

database to keep the site up-to-date. The searching interface includes several "advanced" options. All these options are available when searching the local index and are available in different combinations when metasearching. Users can search for phrases by putting several words on the same line and use an and/or radio button option to request Boolean processing of the phrases. To support IT managers' need for searching by date, nearly 50% of the indexed pages contain dates. These pages are searched when the user selects a date parameter and the entire collection is searched when the option "all documents" is chosen. The TF*IDF equation orders the list of results. Individual terms are weighted differently based on their position within different HTML tags. The summary or snippet provided by the information source is presented with each search result. The user can also elect to see a generic summary with a user-defined number of sentences. In addition to the searching features provided to the user, a human-generated collection of links was added to the left column of the search screen, which can be reviewed by IT managers. The links are placed in general categories and are representative of the relevant IT resources available on the Internet. The information categories represented are "IT Resources," "Business News," "StockMarket," "TechTips," and "Events."

In addition to the metasearch modules that gather IT news articles, there are also two that gather financial data (see Figure 3). As users scan the competitive landscape, they are able to obtain more specific company descriptions as well as the financial positions of companies relative to their main competitors by using the company-related meta-

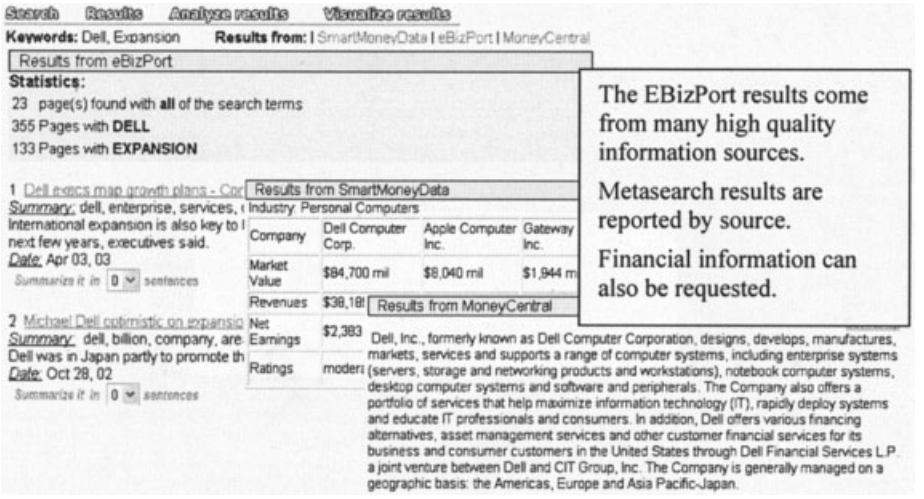

FIG. 3. EBizPort news article and financial data search results.

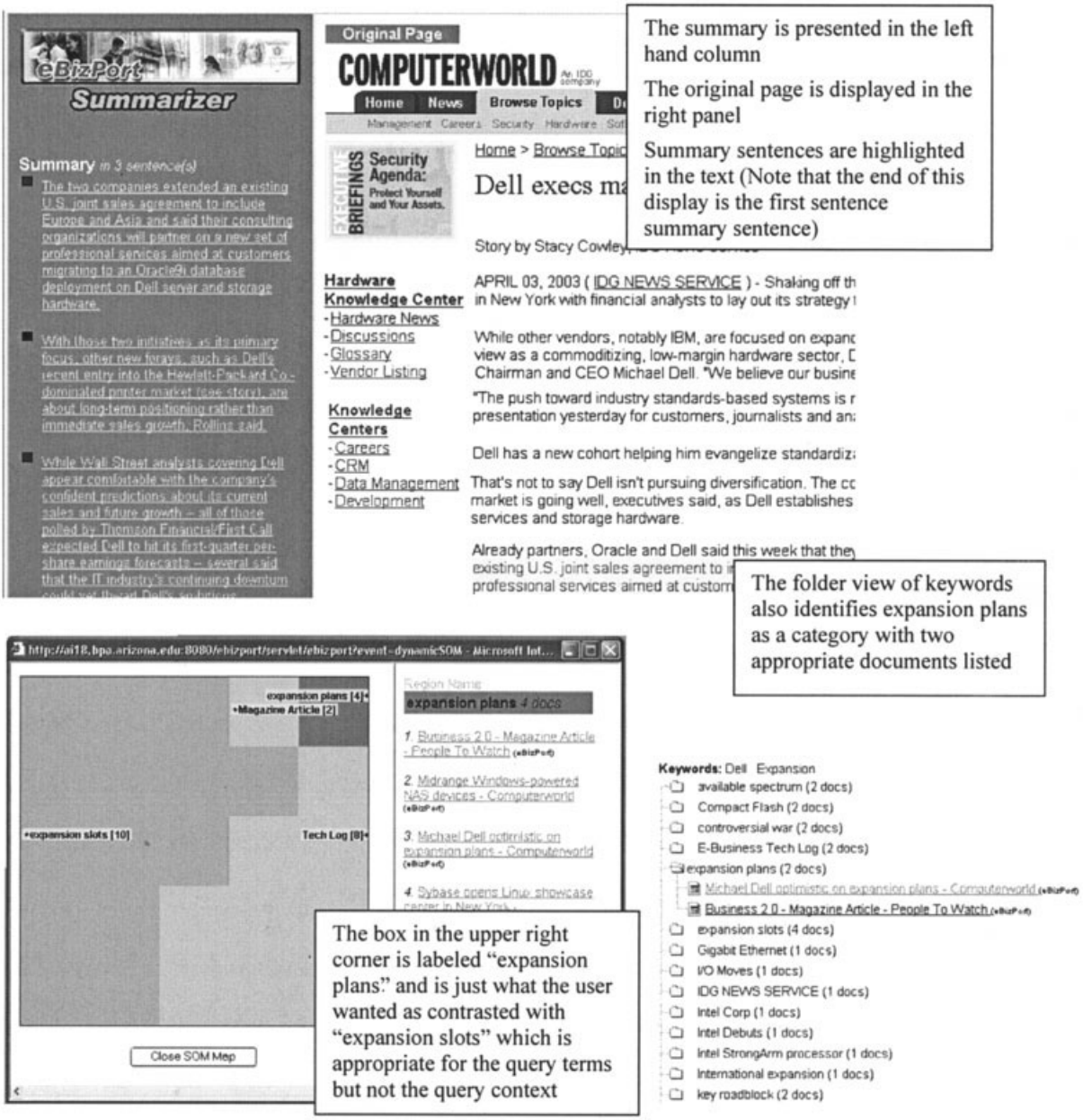

FIG. 4. Browse tools.

search engines. Modules access SmartMoney.com and MoneyCentral.com to obtain financial information. In the present implementation, the stock symbol is entered in the first line of the query box. Financial results are presented along with search results in the results page.

Browsing support. EBizPort implements three analysis tools to help users sift through available data. The categorizer and visualizer provide overviews while the summarizer provides previews. The categorizer provides a folder display based on document clusters. It leverages a common interface metaphor to help users find relevant documents and browse the results space. The visual organization produced by the SOM algorithm in the visualizer allows users to see the major clusters of documents found in a search result. The summarizer tool lets users see the main themes in a document while showing where those themes occur in the source text. These tools are shown in Figure 4.

The first tool is a summarizer that generates generic summaries of single documents found in the list of results (McDonald & Chen, 2002). Summary sentences are generated and given a color code on the left side of the screen. On the right side of the screen, the original document is loaded with the summary sentences highlighted according to the color scheme. By clicking on the summary sentence, users are taken to the location in the document that contains the sentence of interest. By including a summary along with the original, users can quickly go to the areas of the document

that seem the most relevant. This feature seems especially useful with longer documents. The second browsing tool built into the EBizPort is an analyzer that creates a folder view of the retrieved content. In this view, all the pages returned in a search, whether metasearch or local, are sent to the Arizona Noun Phraser (Tolle & Chen, 2000). Phrases that occur frequently in the results collection are selected as the folder labels. If a document contains the same phrase as one of the folder labels, it is included underneath that folder. Documents can appear in several folders if they share many important phrases with other documents. The third browsing tool we incorporated in the EBizPort is the Self-Organizing Map (SOM) (Chen, Houston, Sewell, & Schatz, 1998). The map classifies all the pages returned in the list of results onto a 2-dimensional display. Users can click on the different regions of the map and see the URLs under the selected region in the right-hand column of the display. A larger region means more documents are in that classification. Adjacent regions are more similar in content than nonadjacent regions.

EBizPort design summary. EBizPort creates value for users in two main ways. It provides access to a focused and credible set of resources and it implements a set of interesting analysis tools. In addition, the techniques used to create EBizPort are relatively scalable and modular with a built-for-maintenance paradigm. This combination of characteristics may be useful to others who want to create vertical search portals or digital libraries.

Because the business information landscape changes rapidly, the architecture and database implementation of the system support site maintainability. The ranking algorithm was altered slightly so that documents could be added without re-ranking the entire database. This was done by altering the standard tf*idf calculations to use a constant of 1 million in place of the total number of documents in the collection. This has little impact on results but eases the computational burden of adding incremental results. The clear separation of functions between the front and back ends of the system makes it easier to implement changes. The JSP and servlet model used to create the user interface allow services like the market and financial information lookup module to be added based on other resources on the Internet without disrupting the existing functionality. The collection-building technique also is designed for maintainability. Using the metaharvesting approach to collect URLs allows the collection to be easily "refreshed." We have refreshed our collection twice by rerunning the queries. Resulting URLs are checked for duplication as they are loaded into the database. This combined with the database-driven indexing procedure allows easy implementation of various heuristics to identify pages that need to be respidered on a frequent basis.

Metasearching for URLs and then creating a local index based on those documents is a promising method for building vertical search engines. The technique leverages the value created by other vertical collection builders. It does not replace, but leverages focused crawling and other collection-building techniques. Sending queries to general search engines and implementing unguided crawling produce noisy results. Competitive intelligence system developers may be able to use the metaharvesting approach described here to reduce the complexity of collection-building processes. Sending queries to selected sites allows the builder to address the perceived credibility of the documents in a collection although it requires site-specific adjustments to accomplish the harvesting of valuable documents. Once pages are collected, site-specific heuristics are required to do a creditable job of content date extraction and filtering of unwanted information from Web pages. This is not an easy process but it does create significant value in the usefulness of the index.

Brint.com

Brint.com, the BizTech Network, is a Web portal that provides information from many sources in various formats regarding general business, information technology, and knowledge management. It was founded by Dr. Yogesh Malhotra, a Management faculty at Syracuse University, with co-founder Mini Malhotra. Before its 1996 launch, Brint had existed as the site "A Business Researcher's Interest" dating back to 1994. Brint's free content consists of two main information access methods: "Topic Specific Portals" and the "Portal on Demand" search. "Topic Specific Portals" provide directory-style access to information and can be reached through the menu of Channels, Resources, and Community. "Portal on Demand" search, the signature service of Brint, aggregates and consolidates content from multiple back-end databases. "Portal on Demand" uses a MetaSearch engine that gathers information from some of the best business and technology publications on the Web. Content from external sources are merged with local resources and ranked according to query relevance in real time.

Brint.com is a good information portal benchmark for the EBizPort for three reasons. First, Brint is a popular site that is frequently visited. Brint currently has a 4:1 lead in link popularity statistics over prominent Web publications and Web sites on the topic of knowledge management (www.brint.com). Furthermore, Brint holds a 10:1 lead in link popularity over most of the top ten Web sites returned by search engines for a "knowledge management" search. In addition, Brint has approximately 100,000 registered users, even though registration is not required to use the site. Among the registered users are Fortune 500 and Global 2000 companies. Second, Brint offers a lot of information about the Information Technology industry, including current news headlines. InfoWorld recognized Brint as one of the "best Web sites for keeping up with hi-tech industry innovations. . . ." In March 1997, Computerworld's Annual Forecast declared Brint a "Best" site "especially for its relevance to the multifaceted needs and concerns of IS

professionals. " Brint's large volume of IT news and information makes its content comparable to that of the EBizPort. Finally, Brint's collection-building techniques of integrating a local index with metasearching are analogous to the strategies incorporated by EBizPort. The popularity of Brint along with its similarity to EBizPort in content and collection-building technique makes it a good benchmark for testing the additional functionality of the EBizPort.

Evaluation Methodology

In this section, we describe our methodology used to evaluate EBizPort. We present the objectives, experimental tasks, hypotheses, and design as follows.

Objectives and Experimental Tasks

Our evaluation objectives were threefold: (1) to evaluate the effectiveness, efficiency, and usability of EBizPort; (2) to evaluate the quality of the collection built using our proposed approach; and (3) to evaluate how the tools (summarizer, categorizer, visualizer, navigation side-bar) of EBizPort are used to help browsing for business information on the Web. In all the above evaluation objectives, BRINT.com (BRINT), a business search portal, was selected to be the benchmark for comparison.

To evaluate how the search portals assist business information seeking, scenario-based search tasks and browse tasks were designed. An example of a search task is "In December 2002, IBM announced the acquisition of software company for $2.1 billion. What is the name of this company? " An example of a browse task is "What is the expansion plan of Dell Inc. as reported within the most recent year? Summarize your answers into a number of distinct themes. " The theme identification method was used to evaluate performance in browse tasks (Chen et al., 2001).

To achieve objective (1), we compared the performance of EBizPort with that of BRINT in search and browse tasks. We also asked the subjects to provide subjective ratings on the usability and information quality of the two portals. To achieve objective (2), we compared the search performance between using EBizPort's domain-specific collection together with other metasearchers, and only using EBizPort's metasearchers without searching its domain-specific collection. To achieve objective (3), we analyzed how subjects used each tool in terms of the length of time spent on it and the number of themes obtained.

Hypotheses

Three groups of hypotheses were tested (see Table 2). To compare the effectiveness of the systems, we used accuracy for search tasks and precision and recall for browse tasks. A single measure called F value was used to combine recall and precision (Shaw, Burgin, & Howell, 1997). Efficiency was measured by the amount of time used to finish a task.

TABLE 2. Hypotheses tested in the experiment.

Code	Hypothesis
1. Effectiveness and efficiency of the portals	
H1.1	EBizPort is more effective than BRINT in search tasks.
H1.2	EBizPort is less efficient than BRINT in search tasks.
H1.3	EBizPort is more effective than BRINT in browse tasks.
H1.4	EBizPort is less efficient than BRINT in browse tasks.
H1.5	A combination of EBizPort and BRINT is more effective than either EBizPort or BRINT in browse tasks.
2. Users' subjective evaluation on usability and information quality	
H2.1	EBizPort achieves higher usability than BRINT.
H2.1a	In terms of users' subjective ratings on usefulness, EBizPort achieves higher usability than BRINT.
H2.1b	In terms of users' subjective ratings on ease of use, EBizPort achieves higher usability than BRINT.
H2.1c	In terms of users' subjective ratings on information display and interface design, EBizPort achieves higher usability than BRINT.
H2.2	EBizPort provides a higher information quality than BRINT.
H2.2a	In terms of presentation quality and clarity, EBizPort provides a higher information quality than BRINT.
H2.2b	In terms of coverage and reliability, EBizPort provides a higher information quality than BRINT.
H2.2c	In terms of usability and analysis quality, EBizPort provides a higher information quality than BRINT.
H2.3	EBizPort users achieve a higher overall satisfaction than BRINT users.
3. Quality of EBizPort's domain-specific collection	
H3.1	Using EBizPort's domain-specific collection together with EBizPort's metasearchers is more effective than only using EBizPort's metasearchers in search tasks.
H3.2	Using EBizPort's domain-specific collection together with EBizPort's metasearchers achieves comparable efficiency to only using EBizPort's metasearchers in search tasks.

The formulae used to calculate the above metrics are stated below.

$$Accuracy = \frac{Number\ of\ correctly\ answered\ parts}{Total\ number\ of\ parts}$$

$$Precision = \frac{Number\ of\ relevant\ results\ identified\ by\ the\ subject}{Number\ of\ all\ results\ identified\ by\ the\ subject}$$

$$Recall = \frac{Number\ of\ relevant\ results\ identified\ by\ the\ subject}{Number\ of\ relevant\ results\ identified\ by\ the\ expert}$$

$$F\ value = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

Hypotheses on effectiveness and efficiency. In H1.1–H1.4 (Table 2), we hypothesized that EBizPort would be more

effective than BRINT in search and browse tasks because EBizPort's high quality collection and domain-specific metasearchers provide better search results than BRINT's metasearching. Also, we believed that EBizPort's browse supports would be better than BRINT's navigation supports. However, because BRINT has professional support while EBizPort is just a prototype, users should be less familiar with EBizPort and would, therefore, spend more learning to use it. Thus, we hypothesized that EBizPort would be less efficient than BRINT. In H1.5 (Table 2), we believed that combining browse supports from both systems would achieve better effectiveness than only using the browse supports from either one of the systems. Since we expected to obtain significantly different results from the two systems, combining the results from them would significantly increase recall but create only a small change in precision. Through this arrangement, we tried to mimic a situation in which each subject was allowed to use EBizPort and BRINT together to solve the same problem.

Hypotheses on usability and information quality. In H2.1 (Table 2), we hypothesized that EBizPort would achieve a higher usability than BRINT because EBizPort provides a wide range of metasearchers to choose from, supports specialized functions like company profile search, and has a clean and user-friendly interface. In H2.2 (Table 2), we believed that EBizPort would have a better information quality because EBizPort has a high quality, domain-specific collection that gives good results for searching and browsing. Also, the different tools of EBizPort add values to searching and browsing. Based on the cited advantages of EBizPort, we therefore believed that EBizPort's users would achieve higher overall satisfaction ratings (H2.3, Table 2).

Hypotheses on quality of EBizPort's domain-specific collection. In H3.1 (Table 2), we believed that using EBizPort's domain-specific collection would increase effectiveness of EBizPort's metasearchers in performing search tasks because the collection was built from a wide range of high-quality information sources. In H3.2 (Table 2), we believed that using EBizPort's domain-specific collection would not significantly increase the time required for searching, thus efficiency would not be changed significantly.

Experimental Design

The experiment required each subject to perform 3 search tasks and 1 browse task using each system. A time limit of 3 minutes was imposed on each search task and 8 minutes on each browse task. Among the 3 search tasks, the first two tasks asked about recent activities of a company in the IT industry and the third task asked about the company's profile information. The fourth task was a browse task that asked subjects to find themes related to recent events of the company. When using EBizPort, subjects were instructed not to use EBizPort's domain-specific collection in task 1 but had to use EBizPort's collection in task 2 (but they could use EBizPort's metasearchers in either task), so that we could compare the effect of using EBizPort's collection. All tasks were randomly assigned to different questions to avoid bias due to task content. A pilot test involving three subjects was conducted to evaluate the appropriateness of the tasks before they were actually used in the experiment.

Thirty subjects, mainly business college students, were recruited and each of them received a fixed amount of money as an incentive for their voluntary participation. Many of the subjects also had prior industry work experience. During the experiment, a subject used each of the two systems to perform the tasks. The order in which the systems was used was randomly assigned to the subjects to avoid bias due to system sequence. As each subject was asked to perform similar tasks using the two systems, a one-factor repeated-measures design was used, because it gives greater precision than designs that employ only between-subjects factors. Verbal comments and observational notes were recorded and analyzed.

After finishing the tasks with a system, a subject rated the system on: (1) the usability of the system, (2) the information quality provided by the system, and (3) his/her overall satisfaction with the system. To measure usability, we identified three categories (perceived usefulness, perceived ease of use, and information display and interface design) of items from two usability questionnaires frequently used in the information system field (Davis, 1989; Lewis, 1995). To measure information quality, we modified the 16-dimension construct developed in Wang and Strong (1996) by dropping the dimension on "security," which is not relevant because the information provided by the systems is already public. In addition, because there are different levels of importance in the remaining 15 dimensions, we invited an expert to provide ratings on the relative importance of different dimensions. Such ratings were used to weigh the different dimensions of information quality for the business domain. His ratings as well as the definitions of the 15 dimensions categorized into three groups are shown in Table 3. The expert holds an MBA degree with extended studies in management information systems, marketing, and entrepreneurship and has 14 years of work experience in the IT industry.

Upon finishing with the two systems, the subject was asked to fill in a post-study questionnaire asking about their preferences on the systems and suggestions for improvements or other comments. Answers from search tasks were graded by their correctness while answers to browse tasks were graded according to the expert's judgment. To increase the quality of the expert's judgment, he was first required to provide a set of answers after using both EBizPort and BRINT, and to organize the answers into themes. After the data from all subjects had been collected, the expert read subjects' answers and modified his original answers if

TABLE 3. Definitions of 15 dimensions of information quality and expert ratings.

Dimension	Definition	Expert rating
Presentation quality and clarity		
Accessibility	The extent to which information is available, or easily and quickly retrievable	3
Concise representation	The extent to which information is compactly represented	1
Consistent representation	The extent to which information is presented in the same format	3
Ease of manipulation	The extent to which information is easy to manipulate and apply to different tasks	1
Interpretability	The extent to which information is in appropriate languages, symbols, and units, and the definitions are clear	3
Coverage and reliability		
Appropriate amount of information	The extent to which the volume of information is appropriate for the task at hand	2
Believability	The extent to which information is regarded as true and credible	2
Completeness	The extent to which information is not missing and is of sufficient breadth and depth for the task at hand	3
Free-of-error	The extent to which information is correct and reliable	3
Objectivity	The extent to which information is unbiased, unprejudiced, and impartial	3
Usability and analysis quality		
Relevancy	The extent to which information is applicable and helpful for the task at hand	2
Reputation	The extent to which information is highly regarded in terms of its source or content	3
Timeliness	The extent to which information is sufficiently up-to-date for the task at hand	2
Understandability	The extent to which information is easily comprehended	2
Value-Added	The extent to which information is beneficial and provides advantages from its use	1

Note. Expert rating: 3 = extremely important, 2 = very important, 1 = important.

needed. The final list of the expert's answers was obtained after this two-step process and was used to evaluate the performance of the systems.

Experimental Results and Discussion

In this section we describe and analyze the results of our user evaluation study. Table 4 summarizes the system effectiveness and efficiency in search and browse tasks. Table 5 shows the mean ratings on various dimensions, Table 6 shows the *p* values and results of testing various hypotheses, and Table 7 summarizes subjects' profiles.

TABLE 4. Searching and browsing performance of EBizPort and BRINT.

Portal	Task	Measure	Mean performance	Std. deviation
EBizPort	Search[a]	Accuracy	72.22%	29.14%
		Efficiency[b]	2.48	1.28
	Browse	Precision	81.67%	31.29%
		Recall	22.06%	11.58%
		F value	34.00%	16.27%
		Efficiency[b]	5.46	2.60
BRINT	Search[a]	Accuracy	47.22%	27.36%
		Efficiency[b]	2.61	1.32
	Browse	Precision	83.05%	23.63%
		Recall	25.67%	14.44%
		F value	37.85%	17.79%
		Efficiency[b]	5.56	3.19
Combination (EBizPort + BRINT)	Browse	Precision	87.54%	13.09%
		Recall	45.94%	18.24%
		F value	57.97%	17.66%

[a] The performances of the three search tasks were averaged.
[b] Efficiency was measured in minutes.

Effectiveness and Efficiency of the Portals

The result of testing hypothesis H1.1 shows that *EBizPort was significantly more effective than BRINT for search tasks*. We believe this could be attributed to EBizPort's high-quality domain-specific collection and company profile search function that provide precise and relevant search results. The insignificant difference in search efficiency (H1.2) could be attributed to the fact that both portals metasearched other information sources and hence took a significant amount of time to obtain results. The results of testing H1.3 and H1.4 show no significant difference in the effectiveness and efficiency for browse tasks. We believe that both portals provided useful browse supports that were complementary to each other, because EBizPort obtained results from major business information sources while BRINT obtained results from major search engines. The results of testing H1.5 show that a combination of the two portals outperformed both EBizPort and BRINT in browse tasks, thereby confirming our belief that *EBizPort could significantly augment BRINT with its browse supports*.

Users' Subjective Evaluation and Verbal Comments

The results from testing H2.1–H2.2 show that EBizPort achieved a significantly higher usability and information quality than BRINT along all the different dimensions. We believe that EBizPort's superior usability was attributed to its user-friendly interface and functionalities that support

TABLE 5. Results of users' subjective evaluations.

Dimension	EBizPort		BRINT	
	Mean rating	Std. deviation	Mean rating	Std. deviation
Information quality (overall)	5.27[a]	1.11	4.15[a]	1.46
Presentation quality and clarity	5.30[a]	1.24	4.19[a]	1.38
Coverage and reliability	5.29[a]	1.09	4.29[a]	1.73
Usability and analysis quality	5.17[a]	1.26	4.04[a]	1.66
Usability (overall)	2.70[b]	1.31	4.19[b]	1.47
Perceived usefulness	3.03[b]	1.55	4.48[b]	1.95
Perceived ease of use	2.39[b]	1.34	3.51[b]	1.49
Information display and interface design	2.67[b]	1.54	4.58[b]	1.53
Overall satisfaction	3.20[b]	1.88	4.90[b]	1.88

[a] The range of rating is from 1 to 7 with 7 being the best.
[b] The range of rating is from 1 to 7 with 1 being the best.

searching and browsing high quality business information. In particular, users' ratings on EBizPort's information display and interface design contributed most to EBizPort's superior usability, as reflected by the low p value (=0.000) in testing H2.1c. It showed that EBizPort has provided a highly usable document search interface that contributes to users' satisfaction in searching and browsing. As subject 22 commented on EBizPort: "excellent interface, gives all possible options, very flexible, very interactive, easy to learn and use. " In addition, we believe that EBizPort's significantly higher information quality was attributed to the use of our vertical collection-building approach that yielded a high-quality, domain-specific collection to facilitate credible, precise, and relevant searching. Subject 16 pointed out that EBizPort had "credible sources and high flexibility."

The result of testing H2.3 shows that users' overall satisfaction with EBizPort was significantly higher than that of BRINT. This is because EBizPort provided searching and browsing for high-quality business information and a variety of other credible sources, together with such useful browse supports as navigation links, summarizer, categorizer, and visualizer. As subject 12 said, "the visualization tool is helpful in terms of its classifications. The appearance of the Web site (EBizPort) is appealing and easy to understand. The meta search function allows for detailed search within each magazine of interest." In contrast, BRINT only provided metasearching and many hyperlinks, some of which were broken or irrelevant to the tasks at hand. Subjects complained about BRINT's cluttered user interface. As sub-

TABLE 6. Results of hypothesis testing.

		p value	Result
H1: Effectiveness and efficiency of the portals			
H1.1	Search task effectiveness: EBizPort > BRINT	0.001	Confirmed
H1.2	Search task efficiency: EBizPort < BRINT	0.616	Not confirmed
H1.3	Browse task effectiveness: EBizPort > BRINT	0.479	Not confirmed
H1.4	Browse task efficiency: EBizPort < BRINT	0.867	Not confirmed
H1.5a	Browse task effectiveness: Combined > EBizPort	0.000	Confirmed
H1.5b	Browse task effectiveness: Combined > BRINT	0.000	Confirmed
H2: Users' subjective evaluations			
H2.1	Usability: EBizPort > BRINT	0.000	Confirmed
H2.1a	Usability (usefulness): EBizPort > BRINT	0.003	Confirmed
H2.1b	Usability (ease of use): EBizPort > BRINT	0.007	Confirmed
H2.1c	Usability (information display and interface design): EBizPort > BRINT	0.000	Confirmed
H2.2	Information quality: EBizPort > BRINT	0.000	Confirmed
H2.2a	Information quality (presentation quality and clarity): EBizPort > BRINT	0.000	Confirmed
H2.2b	Information quality (coverage and reliability): EBizPort > BRINT	0.000	Confirmed
H2.2c	Information quality (usability and analysis quality): EBizPort > BRINT	0.001	Confirmed
H2.3	Satisfaction: EBizPort > BRINT	0.002	Confirmed
H3: Quality of EBizPort's domain-specific collection			
H3.1	Search effectiveness: (EBizPort collection + meta searchers) > EBizPort meta searchers	0.211	Not confirmed
H3.2	Search efficiency: (EBizPort collection + meta searchers) = EBizPort meta searchers	0.958	Confirmed

Note. Alpha error = 5%; For details of the hypotheses, please refer to Table 2.

TABLE 7. Subjects' profiles.

Attribute	Subjects' profile
Average number of years in industries	2.32
Degree of reliance on the Web to search for information in subjects' work experience	2.36 (range: 1–7, with 1 being "strongest reliance")
Time spent on using computer (hours per week)	5 subjects spent between 20–25 hours per week, 6 subjects spent between 25–30 hours per week, 6 subjects spent between 30–35 hours per week, 12 subjects spent more than 40 hours per week
Gender	27 subjects are male, 3 are female
Education level	1 subject is an undergraduate student, 8 subjects have earned a bachelor's degree, 18 subjects have earned a master's degree, 3 subjects did not provide education information
Age	12 subjects age between 18–25, 8 subjects age between 26–30, 6 subjects age between 31–35, 1 subject ages between 36–40, 3 subjects did not provide age information

TABLE 9. Subjects' preferences along different dimensions.

	Number of subjects who selected the following system to be their first preference	
Dimension	EBizPort	Brint
To search for specific company information	17	1
To search for information from various business information sources	20	2
To search for high-quality business information	18	3
To browse business-related Web resources	18	3
To categorize a large number of search results into meaningful groups	20	4
To visualize a large number of search results	23	2
To achieve effective integration of information from different sources	20	3
To discover meaningful patterns among search results	20	2
To analyze business information about certain issues	16	4

ject 7 commented that BRINT's "interface (was) painful to look at, (provided) out-of-date information and dead links." Subject 5 said that "information overload" was a major problem of BRINT and more than 10 subjects said that BRINT put too much information on the Web pages, thus making the interface too cluttered. Table 8 summarizes subjects' verbal comments and Table 9 lists subjects' preferences along different dimensions.

From the above results, we conclude that *EBizPort achieved significantly better usability, information quality, and user satisfaction than BRINT*.

EBizPort's Domain-Specific Collection and Browse Supports

The results of testing H3.1 and H3.2 show that the search effectiveness and efficiency of using EBizPort's collection plus metasearchers were not significantly different from those of only using metasearchers. We believe this could be attributed to the way that EBizPort presented results. EBizPort presented lists of results according to different sources. While users might expect to see one integrated long list, they actually got many lists of results put on the same page in a linear manner. Some users did not pay particular attention to the list of results from EBizPort's collection because that list was not always put at the top of the page. They might incorrectly think that those items put at the top of the page were more important than those put at the bottom.

To understand how EBizPort helped browsing for business information, we recorded subjects' actions in using different tools, namely, search function, navigation side bar, summarizer, categorizer, and visualizer. We also have defined an information-seeking episode as a subpart of the whole information seeking process that involves using a

TABLE 8. A summary of subjects' verbal comments.

Portal	Positive comments	Negative comments
EBizPort	- Interface is simple and easy to use	- The processing speed was sometimes slow
	- Provided useful tools to enhance the searching ability	- Too many results coming from different magazines might overwhelm users
	- Allowed summarization, categorization, and visualization that other search engines couldn't provide	- Users were not familiar with categorizer and hence did not find it useful
	- Provided searching of credible sources	- Irrelevant results were found
BRINT	- Users were more familiar with its interface, which is similar to that of Yahoo or Google	- The interface was too cluttered and created information overload
	- Provided links to reputable Web sites	- Many broken links were found
	- Provided searching of many information sources	- Not flexible in selecting information sources
		- Information provided was outdated and not from credible sources

TABLE 10. Usage of EBizPort's tools.

Metric	Search function	Summarizer	Categorizer	Visualizer	Navigation side bar
Proportion of themes obtained (%)	68.83	0.00	3.70	27.47	0.00
Average amount of time spent (in seconds)	111	3.43	27	55.1	1.33
Number of subjects who used this tool in					
First episode	28	0	0	0	2
Second episode	9	3	6	7	0
Third episode	11	2	1	4	0
Fourth episode	5	0	3	2	0
Fifth episode	5	2	1	1	0
Sixth episode	4	0	1	2	0

particular tool to achieve a subgoal of the process. By identifying information-seeking episodes of each subject, we can analyze how the tools were used and how they contributed to finding results. Table 10 summarizes the usage of the tools. Each subject on average went through 3.77 episodes in performing his browse task. As expected, the search function was the most frequently used tool and contributed most to finding themes. While the visualizer and categorizer were the next two most frequently used tools, the visualizer was found to contribute much more help to users in find themes, as shown in the proportion of themes obtained. We, therefore, conclude that *EBizPort's search function and visualizer were the two major tools that contributed to subjects' performance in browse tasks*.

Implications of the Results

The results of this user evaluation study have three implications. First, the encouraging results suggest that our vertical collection-building approach can benefit domain-specific information seeking on the Web. Since many domain-specific communities use the Web to share and disseminate information, our approach for building searchable collections is valuable to these communities. Second, Web document search interface was found to be very important for search and browse support as shown in subjects' favorable comments on EBizPort's interface. EBizPort provides an example of a clean, user-friendly document search interface that reduces information overload in Web information seeking. However, the confusing presentation of search results (see EBizPort's Domain-Specific Collection and Browse Supports) and the relatively small usage of certain interface components (e.g., navigation side bar) also suggest that further improvements of EBizPort's interface are needed. Third, visualization was found to add value to Web searching and browsing because it amplifies human cognition and reduces information overload. Further research in developing visualization tools that can benefit information seeking on the Web is thus needed.

Conclusions and Future Directions

To utilize the valuable information on the Internet, Business Intelligence tools must address the searching and analysis challenges that come with extracting useful intelligence information from large volumes of information and data. Such a portal must address the challenges of content gathering and the reported lack of analysis tools. EBizPort was created with these specific needs in mind. We incorporated metasearch with local indexing to leverage different information sources and yet address the challenge of merging results. We provided a search interface with preview and overview tools to enhance the users' ability to analyze information retrieved. To evaluate the effectiveness of our tools, we conducted a user study comparing the EBizPort to an existing and reputable knowledge management portal. Results showed that EBizPort was significantly more effective in search tasks and could augment a comparable site in browse tasks. Users also rated EBizPort as having a significantly higher information quality, usability, and overall satisfaction than the comparable site.

The analysis tools in the EBizPort are the first step to helping users dynamically visualize the news and rapidly changing conditions in the IT industry. With the large amount of time-sensitive material published on the Web, we are interested in analyzing the changes of IT topics over time. The publication date information extracted by EBizPort should help in this task. In addition, a tool kit containing some of the analysis tools we have described is under development. We plan to make this tool kit available to the research community.

Acknowledgments

This research is partly supported by NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999 to March 2002. We greatly appreciate the help of our expert, Mr. Cortez W. Smith, and our evaluation team members Alfonso A. Bonillas and Theodore Elhourani for their hard work in completing the user study.

References

Arasu, A., Cho, J., Garcia-Molina, H., & Raghavan, S. (2001). Searching the Web. ACM Transactions on Internet Technologies 1(1), 2–43.

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the on-line search interface. Online Review 13(5), 407–431.

Bergman, M.K. (2001). The deep Web: Surfacing hidden value. The Journal of Electronic Publishing 7(1).

Bergmark, D. (2002). Collection synthesis. In G. Marchionini & W. Hersh (Eds.), The Second ACM/IEEE-CS Joint Conference on Digital Libraries. Portland, OR: ACM Press.

Bhattacharya, S., & Basu, P.K. (1998). Mapping a research area at the micro level using co-word analysis. Scientometrics, 43, 359–372.

Börner, K., Chen, C., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology, 37, 179–255.

Bowman, C.M., Danzig, P.B., Manber, U., & Schwartz, F. (1994). Scalable Internet resource discovery: Research problems and approaches. Communications of the ACM, 37(8), 98–107.

Boyack, K.W., Wylie, B.N., & Davidson, G.S. (2002). Domain visualization using VxInsight for science and technology management. Journal of the American Society for Information Science and Technology, 53, 764–774.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertext Web search engine. In P.H. Enslow & A. Ellis (Eds.), Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia: Elsevier Science Publishers B.V.

Card, S., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). Readings in information retrieval: Using vision to think. San Francisco, CA: Morgan Kaufmann.

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

Chau, M., Zeng, D., & Chen, H. (2001). First ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 79–87). Roanoke, VA: ACM Press.

Chen, C. (1997). Structuring and visualizing the WWW with generalized similarity analysis. In M. Bernstein, K. Østerbye, & L. Carr (Eds.), Proceedings of the Eighth ACM Conference on Hypertext (pp. 177–186). Southampton, UK: ACM Press.

Chen, C. (1999). Informaton visualisation and virutual environments. London: Springer-Verlag.

Chen, C., Paul, R.J., & O'Keefe, B. (2001). Fitting the jigsaw of citation: Information visualization in domain analysis. Journal of the American Society for Information Science and Technology, 52, 315–330.

Chen, H., Schufels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. Journal of Visual Communication and Image Representation, 7(1), 88–102.

Chen, H., Houston, A.L., Sewell, R.R., & Schatz, B.R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. Journal of the American Society for Information Science, 49(7), 582–603.

Chen, H., Fan, H., Chau, M., & Zeng, D. (2001). MetaSpider: Meta-searching and categorization on the Web. Journal of the American Society for Information and Science and Technology, 52(13), 1134–1147.

Choo, C.W., Detlor, B., & Turnbull, D. (2000). Information seeking on the Web: An integrated model of browsing and searching. First Monday, 5(2).

Chung, W., Chen, H., & Nunamaker, J.F. (2003a). Business Intelligence Explorer: A knowledge map framework for discovering business intelligence on the Web. Proceedings of the 36th Hawaii International Conference on System Sciences, Hawaii, Big Island, USA, IEEE Computer Society.

Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T.H., & Chen, H. (in press). Internet searching and browsing in a multilingual world: An experiment on the Chinese Business Intelligence Portal (CBizPort). Journal of the American Society for Information and Science and Technology.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340.

Feng, Y., & Börner, K. (2002). Using semantic treemaps to categorize and visualize bookmark files. In R.F. Erbacher, P.C. Chen, M. Groehn, J.C. Roberts, & C.M. Wittenbrink (Eds.), Proceedings of the SPIE Conference on Visualization and Data Analysis. (p. 24). San Jose, CA: IS&T and SPIE.

Firmin, T., & Chrzanowski, M.J. (1999). An evaluation of automatic text summarization systems. In I. Mani, & M.T. Maybury (Eds.), Advances in automatic text summarization, I (pp. 325–336). Cambridge: The MIT Press.

Fogg, B.J., & Tseng, H. (1999). The elements of computer credibility. CHI 99 conference on Human factors in computing systems: The CHI is the limit. Pittsburgh: ACM Press.

Fuld, L., Sawka, K., Carmichael, J., Kim, J., & Hynes, K. (2002). Intelligence software report 2002. Cambridge, MA: Fuld & Company.

The Futures Group (1995). Ostriches and eagles: Competitive intelligence capabilities in U.S. companies. Glastonbury, CT: The Futures Group.

Gershon, N., Eick, S., & Card, S. (1998). Design: Information visualization. Interactions, 5(2), 9–15.

Greene, S., Marchionini, G., Plaisant, C., & Schneiderman, B. (2000). Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. Journal of the American Society for Information Science, 51(4), 380–393.

Hearst, M. (1995). TileBars: Visualization of term distribution information in full text information access. In I.R. Katz, R. Mack, L. Marks, M.B. Rosson, & J. Nielsen (Eds.), Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (pp. 29–66), Denver, CO.

Hearst, M. (1999). User interfaces and visualization. In B. Ribeiro-Neto (Ed.), Modern information retrieval (pp. 257–339). New York: ACM Press.

Hemmje, M., Clemens, K., & Willett, A. (1994). LyberWorld: A visualization user interface supporting fulltext retrieval (pp. 249–259). The 17th Annual International ACM/SIGIR Conference. Dublin, Ireland.

Hlynka, D., & Welsh, J. (1996). What makes an effective home page? A critical analysis. Retrieved from http://www.umanitoba.ca/faculties/education/cmns/aect.html.

Huang, L., Ulrich, T., Hemmje, M., & Neuhold, E.J. (2001). Adaptively constructing the query interface for meta-search engines. The Sixth International Conference on Intelligent User Interfaces. Santa Fe: ACM Press.

Katerattanakul, P., & Siau, K. (1999). Measuring information quality of Web sites: Development of an instrument. The 20th International Conference on IS. Charlotte, NC: Association for Information Systems.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), 604–632.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the Web. ACM Computing Surveys (CSUR), 32(2), 144–173.

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

Korfhage, R.R. (1991). To see or not to see: Is that the query? Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 134–141). Chicago, IL: ACM Press.

Kruger A., Giles, C.L., Coetzee, F.M., Glover, E., Flake, G.W., Lawrence, S., et al. (2000). DEADLINER: Building a new niche search engine. The Ninth International Conference on Information and Knowledge Management. McLean, VA: ACM Press.

Kuhlthau, C.C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science, 42(5), 361–371.

Lagoze, C., Arms, W., Gan, S., Hillman, D., Ingram, C., Krafft, D., et al. (2002). Core services in the architecture of the national science digital library (NSDL). The Second ACM/IEEE-CS Joint Conference on Digital Libraries. Portland, OR: ACM Press.

Lewis, J.R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7(1), 57–78.

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

Mahlck, P., & Persson, O. (2000). Socio-bibliometric mapping of intra-departmental networks. Scientometrics, 49, 81–91.

Marchionini, G., & Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. Computer, 21(1), 70–79.

McDonald, D., & Chen, H. (2002). Using sentence-selection heuristics to rank text segments in TXTRACTOR. Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 28–35). Portland, OR: ACM Press.

Meng, W., Yu, C., & Liu, K.-L. (2002). Building efficient and effective metasearch engines. ACM Computing Surveys (CSUR), 34(1), 48–89.

Meng, W., Wu, Z., Yu, C., & Li, Z. (2001). A highly scalable and effective method for metasearch. ACM Transactions on Information Systems (TOIS), 19(3), 310–335.

Najork, M., & Heydon, A. (2001). High-performance Web crawling (SRC Research Report No. 173). Retrieved October 10, 2002, from http://gatekeeper.research.compaq.com/pub/DEC/SRC/research-reports/abstracts/scr-rr-173.html

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In M.J. Tauber (Ed.), Proceedings of the Conference on Human Factors in Computing Systems (pp. 213–220). Vancouver, British Columbia, Canada: ACM Press.

Sanderson, M. (1998). Accurate user directed summarization from existing tools. Conference on Information and Knowledge Management, Bethesda, MD.

Shaw, W.M.J., Burgin, R., & Howell, P. (1997). Performance standards and evaluations in information retrieval test collections: Cluster-based retrieval models. Information Processing and Management, 33(1), 1–14.

Shneiderman, B. (1992). Tree visualization with tree-maps: A 2-D space filling approach. ACM Transactions on Graphics, 11, 92–99.

Small, H. (1999). A passage through science: Crossing disciplinary boundaries. Library Trends, 48, 72–108.

Small, H. (2000). Charting pathways through science: Exploring Garfield's vision of a unified index to science. In H.B. Atkins (Ed.), The web of knowledge: A Festschrift in honor of Eugene Garfield (pp. 449–473). Medford, NJ: Information Today, Inc.

Spence, B. (2001). Information visualization. Reading, MA: Addison-Wesley.

Spence, R. (1999). A framework for navigation. International Journal of Human-Computer Studies, 51(5), 919–945.

Sutcliffe, A.G., & Ennis, M. (1998). Towards a cognitive theory of information retrieval [Special issue]. Interacting with Computers, 10, 321–351.

Tan, S.S.L., Teo, H.-H., Tan, B.C.Y., & Wei, W.-K. (1998). Environmental scanning on the Internet. Proceedings of the International Conference on Information Systems. Helsinki, Finland: Association for Information Systems.

Tolle, K.M., & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science, 51(4), 352–370.

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. Melbourne, Australia: ACM SIGIR.

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), 108–116.

Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(14), 5–34.

Wilson, T.D. (1999). Models of information behavior research. Journal of Documentation, 55(3), 249–270.

Witten, I.H., Nevill-Manning, C., McNab, R., & Cunningham, S.J. (1998). A public library based on full-text retrieval. Communications of the ACM, 41(4), 71–75.