Available online at www.sciencedirect.com
SCIENCE DIRECT®
Decision Support Systems xx (2006) xxx–xxx
Decision Support Systems
www.elsevier.com/locate/dsw

Building a scientific knowledge web portal: The NanoPort experience

Michael Chau [a,*], Zan Huang [b], Jialun Qin [c], Yilu Zhou [c], Hsinchun Chen [c]

[a] School of Business, Faculty of Business and Economics, The University of Hong Kong, Pokfulam, Hong Kong
[b] Department of Supply Chain and Information Systems, Smeal College of Business, The Pennsylvania State University, PA 16802, USA
[c] Department of Management Information Systems, Eller College of Management, The University of Arizona, Tucson, Arizona 85721, USA

Received 30 January 2004; received in revised form 14 January 2006; accepted 25 January 2006

Abstract

There has been a tremendous growth in the amount of information and resources on the World Wide Web that are useful to researchers and practitioners in science domains. While the Web has made the communication and sharing of research ideas and results among scientists easier and faster than ever, its dynamic and unstructured nature also makes the scientists faced with such problems as information overload, vocabulary difference, and lack of analysis tools. To address these problems, it is highly desirable to have an integrated, "one-stop shopping" Web portal to support effective information searching and analysis as well as to enhance communication and collaboration among researchers in various scientific fields. In this paper, we review existing information retrieval techniques and related literature, and propose a framework for developing integrated Web portals that support information searching and analysis for scientific knowledge. Our framework incorporates collection building, meta-searching, keyword suggestion, and various content analysis techniques such as document summarization, document clustering, and topic map visualization. Patent analysis techniques such as citation analysis and content map analysis are also incorporated. To demonstrate the feasibility of our approach, we developed based on our architecture a knowledge portal, called NanoPort, in the field of nanoscale science and engineering. We report our experience and explore the various issues of relevance to developing a Web portal for scientific domains. The system was compared to other search systems in the field and several design issues were identified. An evaluation study was conducted and the results showed that subjects were more satisfied with the NanoPort system than with Scirus, a leading search engine for scientific articles. Through our prototype system, we demonstrated the feasibility of using such an integrated approach and the study brought insight into applying the proposed domain-independent architecture to different areas of science and engineering in the future.

Keywords: Web portals; Design; Web spiders; Meta-search; Document summarization; Document clustering; Self-organizing maps; Patent analysis; Visualization; Nanotechnology

1. Introduction

The growth of the Web has made the communication and sharing of research ideas and results among scientists easier and faster than ever. The Web has made available a large amount of useful information and

* Corresponding author.
E-mail addresses: mchau@business.hku.hk (M. Chau), zanhuang@psu.edu (Z. Huang), qin@u.arizona.edu (J. Qin), yilu@u.arizona.edu (Y. Zhou), hchen@eller.arizona.edu (H. Chen).

0167-9236/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2006.01.004

resources that can be useful in various scientific research areas, such as papers reporting research results and patents describing industrial innovation that are critical to the success of researchers and scientists. Knowledge of what others may be doing helps them to avoid duplication of efforts and to improve research techniques. It can also help researchers solve dilemmas and provide insights into future research.

Various scientific fields, such as bioinformatics and nanotechnology, have experienced tremendous growth over the past several years and are facing increasingly more complex and challenging research issues. While the Web provides a convenient way for information searching, researchers often find themselves facing the information overload problem [2], a problem in which a search employing a general-purpose search engine such as Google (www.google.com) can result in thousands of hits. As a discipline often encompasses a diversity of research perspectives and application areas, terminology and vocabulary differences have also emerged, when researchers in different disciplines use different terminologies to describe their findings [7,15]. As a result, user search terms may be different from the indexing terms used in databases.

Since the speed of knowledge creation and information generation is faster than ever, the problem known as the fluidity of concepts further complicates the retrieval issue [14]. A concept may be perceived differently by different researchers and may convey different meanings at different times. Subcategories also are changing much faster than in longer-established fields and new categories evolve frequently. Information sources and quality on the Web are equally diverse [30]. They may include scientific papers, journals, technical reports, patents, and Web pages of widely varied quality. Researchers often have to go to multiple information sources (Web sites, search engines, online academic databases, etc.) in order to identify quality, time-critical information. Moreover, as most current search engines do not provide any content analysis capabilities, users have to perform manual analysis on the documents retrieved from different data sources.

A tool for automatic analysis of different types of documents is highly desired. This is especially important for structured, interlinked documents such as patents, which can reveal industrial trends as well as the latest development in a particular field. Patent citation analysis and content analysis can often review important trends or technology breakthroughs in the industry, but such functionalities usually are not available to scientists and researchers. Consequently, they often find it difficult to catch up with the latest industrial and technological development.

To address the above problems, it is highly desirable to have an integrated, "one-stop shopping" Web portal to support effective information searching and analysis as well as to enhance communication and collaboration among researchers in various scientific fields. Building a successful Web portal to provide such an environment, especially for young and evolving fields, is a necessary and challenging task. In the current project, we aim to explore the various issues of relevance to developing a Web portal for scientific domains. The proposed customized search capabilities aim to help researchers search more effectively and efficiently for relevant information and learn more about what is going on in the field. The project also aims to demonstrate the feasibility of such a Web portal approach which can be applied to different areas of science and engineering.

In this paper, we report our experience in the implementation of a Web portal in the domain of nanoscale science and engineering (NSE). The Web portal and the complementary intelligent search and analysis engine were designed to support the information needs of researchers in the NSE community. The rest of the paper is outlined as follows. Section 2 reviews related research in information searching, analysis, and visualization. In Section 3, we discuss the research questions investigated in this study. Section 4 describes the system architecture and main components of our proposed approach. In Section 5, we present a case study by discussing the NanoPort system—an implementation of our approach in the NSE domain. In Section 6, we present several sample user sessions to demonstrate how the system can be used to help users with their information needs in the NSE domain. In Section 7, we report the experiments conducted to evaluate the system. In Section 8, we discuss the general experience and the lessons learned in this project. We conclude the paper in Section 9 by summarizing our contributions and suggesting future research directions.

2. Literature review

Many approaches to document retrieval, analysis, and visualization have been adopted in academia and industries. In the following, we review these techniques and their strengths and weaknesses. Section 2.1 reviews various techniques that have been used for searching the Web and analyzing Web contents, including general search engines, vertical search engines, meta-searching, text indexing, and text clustering. Section 2.2 reviews academic and industrial research aimed at mining information from patents. In particular, our review focuses on patent citation analysis.

2.1. Web search and analysis

2.1.1. General-purpose search engines and vertical search engines

Many different search engines are available on the Internet. Each has its own characteristics and employs its preferred algorithm in indexing, ranking and visualizing Web documents. For example, AltaVista (www.altavista.com) and Google (www.google.com) allow users to submit queries and retrieve Web pages in a ranked order, while Yahoo (www.yahoo.com) groups Web sites into categories, creating a hierarchical directory of a subset of the Internet. Most prevailing search engines, such as Google, are keyword-based [1]. Although their search speeds are fast, their results are often overwhelming and imprecise. Low precision and low recall rates make it difficult to obtain specialized, domain-specific information from these search engines.

Vertical search engines, or domain-specific search engines, have been built to facilitate more efficient searching in various domains. These search engines alleviate the problem to some extent, by providing more precise results and more customized features. LawCrawler (www.lawcrawler.com), BuildingOnline (www.buildingonline.com), Scirus (www.scirus.com), BioView.com (www.bioview.com), and NanoSpot (www.nano-spot.org) are some examples. A good vertical search engine should contain as many relevant, high-quality pages and as few irrelevant, low-quality pages as possible. The search engine needs to locate the URLs that point to relevant Web pages. To improve efficiency, it is necessary for the spider to predict which URL is most likely to point to relevant material and thus should be fetched first.

Search engines usually use Internet spiders (also referred to as Web robots or crawlers) to retrieve pages from the Web by recursively following URL links in pages using standard HTTP protocols [4,13]. These are programs that collect Internet pages and explore outgoing links in each page to continue the process. Different techniques have been used to guide these "focused spiders" such that more relevant pages can be collected efficiently. While these methods have different levels of performance, these techniques have two major problems. First, most focused spiders use "local" graph search algorithms, meaning that the spiders can only visit pages which are directly or indirectly linked from the starting Web pages; documents that are not linked to these starting pages will be missed. Second, most spiders are not able to fetch documents that are behind the "Hidden Web"—documents hidden behind search forms or other interfaces which cannot be accessed by following hyperlinks.

2.1.2. Meta-search engines

Empirical studies show that every search service returns a different set of documents for the same query [30]. It has been suggested that when relying solely on one search engine, users could miss over 77% of the references they might find most relevant because no single search engine is likely to return more than 45% of relevant results [41]. A study by NEC Research Institute drew some similar conclusions, revealing an alarming fact about Internet search engines: they cannot keep up with the net's dynamic growth, and each search engine covers only about 16% of the total Web sites [30].

The emergence of meta-search engines provides a credible resolution of divergence by triangulating output from several engines to arrive at relevant results. Several server- and client-based meta-search engines such as Copernic (www.copernic.com), MetaSpider [5,10], MegaSpider (www.megaspider.com), MetaCrawler (www.metacrawler.com) "search the search engines" [41]. The results from other search engines are combined and presented to users. Although the information returned from meta-search engines is comprehensive, the problem of information overload worsens if no post-retrieval analysis is provided.

2.1.3. Search result analysis

In most search engines, search results are returned as a ranked list of Web pages. Such a list, however, does not provide a user with extra information about the set of returned documents. The user has to browse through the list of documents to locate relevant Web pages. Some search engines attempt to alleviate this problem by performing post-retrieval analysis and classification of documents returned. Automatic indexing algorithms have been used widely to extract key concepts from textual data, and it has been shown that automatic indexing is as effective as human indexing [39]. Many proven techniques have been developed. For example, linguistics approaches such as noun phrasing have been applied to perform indexing for phrases rather than just words [42]. These techniques are useful in extracting meaningful terms from text documents not only for document retrieval but also for further analysis.

Another type of analysis is text classification and clustering. Text classification is the classification of documents into predefined categories, while text clustering group documents into categories dynamically defined based on their similarities. Machine learning is the basis of most text classification and clustering applications. Text classification has been extensively reported at SIGIR conferences and evaluated on standard testbeds. Neural network programs also have

been applied to text classification, usually employing the feedforward/backpropagation neural network model [28,44]. Term frequencies or TF*IDF scores (term frequency multiplied by inverse document frequency) of the terms are used to form a vector [40] which can be used as the input to the network as training examples. Another new technique used in text classification is the support vector machine (SVM). Joachims first applied SVM to text classification [24]. It has been shown that SVM achieved the best performance on the Reuters-21578 data set for document classification [45].

Similarly to text classification, text clustering tries to assign documents into different categories based on their similarities. However, in text clustering, there are no predefined categories; all categories are dynamically defined. There are two types of clustering algorithms, namely hierarchical clustering and non-hierarchical clustering. The nearest neighbor method and Ward's algorithm [43] are the most widely used hierarchical clustering methods. For non-hierarchical clustering, one of the most common approaches is the K-means algorithm [38]. The Single-Pass method [22] is also widely used. However, its performance depends on the order of the input vectors and it tends to produce large clusters [37]. Suffix Tree Clustering, a linear time clustering algorithm that identifies phrases common to groups of documents, is another incremental clustering technique [47]. In addition, neural network approach also has been applied. For example, Kohonen's self-organizing map (SOM) [26], a type of neural network that produces a two-dimensional grid representation for *n*-dimensional features, has been widely applied in information retrieval applications [12,27,31]. The self-organizing map can be either multi-layered or single-layered.

2.2. Patent analysis

Patent is a special type of technology document, which contain rich content regarding technology innovations and is accessible by the general public. Patent documents are strictly structured, providing standardized fields like patent citation, issue date, assignee (the institution to which the patent is assigned to), inventors, technology field classification, and country and city of the assignee and inventors, etc. All these special features of patent documents make them a valuable source of knowledge.

Patent analysis has been widely used by academic researchers and industrial technology analysts. Patent analysis is important for scientific domains because they can provide useful information and insights about industrial trends and technology development. There is a substantial academic literature and many industrial practices of using patent analysis for such purposes [25,36]. Patent analysis allows researchers and scientists to keep up-to-date with the latest development in the field.

Patent analysis can be categorized into three types, namely patenting activity analysis, patent content analysis, and patent citation analysis. In patenting activity analysis, the main focus is basic patent indicators, such as number of patents. Patenting activity analysis is relatively simple and does not provide much information about new technologies or the relationships among the patents. The second type is patent content analysis, which has been mainly focused on indexing and classification of the patent documents to support efficient retrieval [29]. It is also quite desirable and valuable to summarize the major technology topics of a large collection of patent documents based on their content and to provide technology topic landscape of the field. Some studies have applied bibliometric maps using the co-word analysis to visualize the cognitive structure of technology knowledge bases and their interrelations [19,20].

The third type of patent analysis is patent citation analysis, which studies the relationships among the patents based on citations. Citations are required for patents to reveal relevant prior arts in other patents and scientific literature, and such citations contain rich information and have been the focus of patent analysis. Patent citation analysis originated in Eugene Garfield's work on science citation analysis [17,18]. Based on previous reviews of patent citation analysis, patent citation analysis can be further classified into four categories: technology performance evaluation, tracing the transfer of knowledge, identifying key earlier patents, and miscellaneous applications [25,36].

3. Research questions

The information overload issue and the lack of analysis tools for scientists and researchers remain to be two urgent problems to be resolved. Although general searching and analysis techniques exist, they are often not integrated effectively nor are they customized for a particular scientific domain. We pose the following research questions: (1) How can we collect and filter a set of documents from the Web that are relevant to a particular scientific domain? (2) Can an integrated Web search portal help scientists and researchers search for Web-based information? (3) Can we apply existing text analysis and visualization techniques to identify trends and interesting patterns in Web documents and patents that may reflect the historical and current development of a scientific domain?

4. Proposed system architecture

4.1. Generic portal design

Two main steps are often involved in building a Web portal: (1) creating a domain-specific collection of Web content and (2) supporting searching of the documents and analysis of search results. In the first step, it is common to use a domain-specific, *Vertical Spider* to collect relevant documents on the Web. A simple *Document Indexer* can be used to create a searchable index of the documents. These two steps are often performed in an offline batch process and the portal database is updated periodically (e.g., once a month). Fig. 1 shows a diagram of the process.

After a domain-specific collection is created, a Web portal needs to provide searching capabilities to users such that they can access the collection in the portal easily. To better support users' tasks, it is also highly desirable to have content analysis functionalities such as document clustering. A generic design incorporating these functionalities is shown in Fig. 2.

The first thing that an information seeker can do with the Web portal is to enter a search query to the system. A *Meta-searcher* component should retrieve documents related to the query in the portal database as well as from other data sources using meta-searching techniques [4,5,10,41]. The Meta-searcher, however, will not update the content in the portal database. Meta-searching is important for domain-specific Web portals. Whenever a search query is submitted to the portal, the query can be forwarded to different databases and search engines, and the search results can be compiled and presented to the users. This allows users to have a comprehensive set of different types of search results, such as Web pages, journal article abstracts, and patents. The search results will then be combined and presented to the user as a ranked list, just like most other search engines. Additionally, a *Keyword Suggester* component should return a set of relevant keywords such that the user can expand or refine the original search query. The keyword suggestion can be based on metrics such as pre-calculated word association or popularity of search keywords.

A *Content Analysis* component should be responsible for analyzing the documents retrieved from the different data sources. Many different information retrieval and text mining techniques, such as document summarization, document clustering, document visualization, or link/citation analysis, can be incorporated in this component. Such functionalities can allow the user to have a better overview of the retrieved documents and perform further analysis. For example, after a set of documents are retrieved, the system can extract the key words or phrases from these documents such that users can see a list of the important concepts covered in the retrieved documents [5,10,11]. Document clustering also can be used to provide a richer representation of the retrieved documents by grouping documents together based on their textual similarity. The presentation of these clustered documents provides a context for users

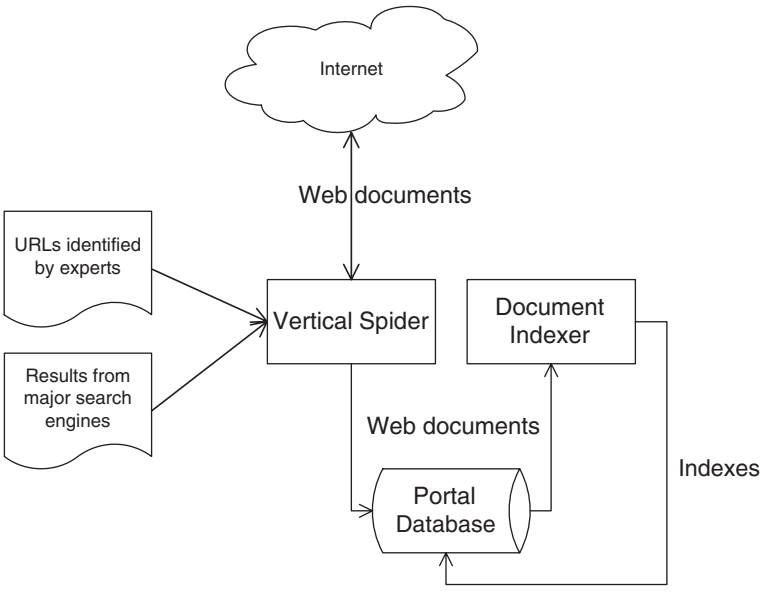

Fig. 1. Creating domain-specific content.

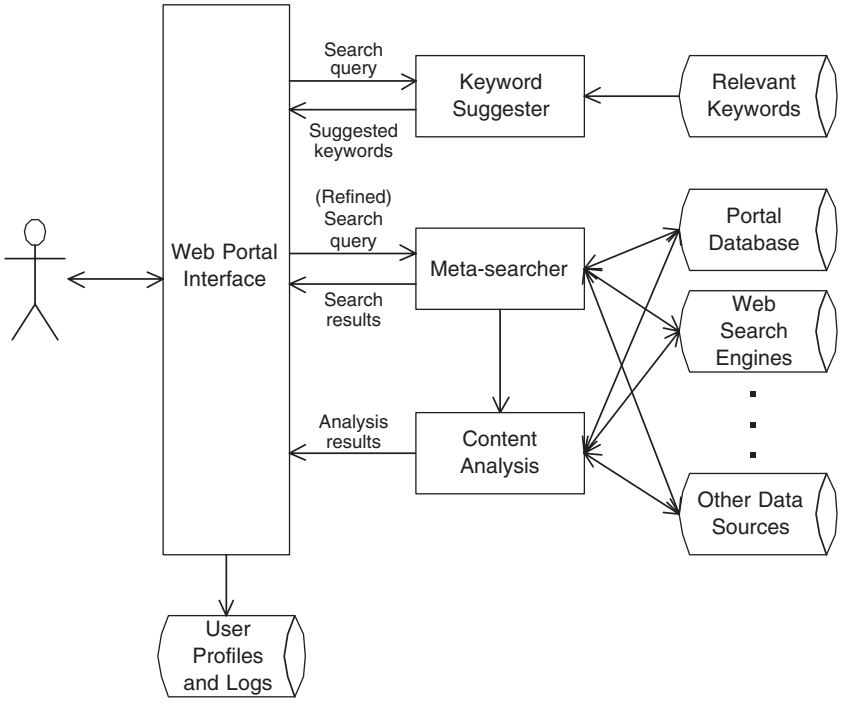

Fig. 2. Generic Web portal design.

to understand the relationships among retrieved documents and to identify documents of interest more easily [5,10,47]). It is also important to support search result visualization such that users can get a quick overview of the retrieved documents.

Content analysis also can be performed on individual documents. For example, single document summarization techniques can be used to extract the key sentences from a long Web page such that users can quickly decide whether a document is of interest without reading the document in detail [6,33]. Finally, user logs and profiles can be stored in the Web portal for system improvement or personalization features.

Our design is based on a server-side approach—most computational components reside on the system's server rather than user's own personal computer. This approach allows users to access the search service more easily through the Web site and has been used by most popular search engines. Any user with a Web browser can use the search tool conveniently from any platform, such as Windows, Linux, or Macintosh, and no software download or installation is needed. The Web browser interface, through which the server-side search tool is accessed, is also more familiar to users. On the other hand, while some Web search tools have been implemented using a client-side approach [6,11], such tools require users to spend time and effort to download the installation file from the Web and install the software on their own computers.

4.2. Proposed architecture

Based on the generic design discussed above, we propose a detailed architecture that incorporates our specific techniques. Our architecture integrates retrieval, analysis, categorization, and visualization of Internet-based information. The architecture is domain-independent and can be applied to different scientific domains. In addition, we also propose applying text analysis and visualization techniques to patents relating to scientific research.

Similar to the generic design, our architecture consists of two major modules: (1) content creation and (2) search and analysis support. The details of each module are discussed below.

4.2.1. Content creation

In the content creation process, as discussed earlier, a Vertical Spider is responsible for collecting domain-specific Web documents, which are then indexed by the Document Indexer. This process in our architecture is the same as that in the generic portal design (see Fig. 1). The specific techniques used in our architecture are discussed below.

4.2.1.1. Vertical Spider. To build a collection with a set of comprehensive and high-quality Web documents, we proposed to use a meta-search enhanced global spidering approach to address the problems of traditional focused spiders. In our approach, instead of using only focused spiders to fetch pages, we added meta-spiders to the crawling process to extend the ability of focused spidering. The focused spiders are developed based on our previous research [3] and start with a set of URLs defined by domain experts. Outgoing hyperlinks are extracted from the pages collected, sorted according to their relevance, and put into a URL queue. At the same time, the meta-spiders keep sending domain-specific queries to multiple search engines and adding the top URLs returned into the URL queue. The combined top results from multiple search engines are usually of high-quality and much diversity, thus adding them to the URL queue can effectively prevent the focused crawler being limited within particular hyperlinked communities. Hidden contents are also fetched by the meta-spiders and added to the Web page collection, which can greatly improve the quality of the collection.

4.2.1.2. Document Indexer. The Document Indexer is responsible for tokenizing each document into words, and recording the relationships between the words and the documents (i.e., indexing). Various information retrieval techniques, such as stemming and stop-word removal, can be applied as necessary. The resulting searchable index is then stored into a database for document retrieval and further analysis.

4.2.2. Search and analysis support

After a collection is created, the second module of the architecture is needed to support searching and analysis of the documents (see Fig. 3). This module forms the main part of the portal that interacts with the users. It consists of a Web-based user interface and six functional components, namely the *Keyword Suggester*, the *Meta-searcher*, the *Document Summarizer*, the *Document Clusterer*, the *Topic Map*, and the *Patent Analyzer*. The Keyword Suggester is designed to help users formulate and refine their search queries by suggesting synonyms or other relevant keywords based on co-occurrence analysis. The Meta-searcher is responsible for parsing users' search queries, forwarding them to various databases and search engines, and combining the search results. The Document Summarizer is used to summarize a given Web page into a few sentences such that users can grasp a quick overview of the page. Such overview can help users decide quickly whether a document is interesting or not. The Document Clusterer extracts key phrases from search results and organizes the Web pages into folders based on their topics. The Topic Map further analyzes the results by clustering them into different regions on a two-dimensional map based on the self-organizing map algorithm. The last component, the Patent Analyzer, specializes in patent content and citation analysis. We describe below the functionalities of each component in detail.

4.2.2.1. Keyword Suggester. The Keyword Suggester helps users expand or refine their search queries by suggesting related keywords. Two approaches are used for keyword suggestion in our architecture. The first approach is a *Concept Space* approach [12,8]. A concept space is created by computing the co-occurrence frequencies of each pair of phrases and words found in a set of documents. In the second approach, we utilize the keyword suggestion functionality available in the Scirus search engine (www.scirus.com), which provides a list of relevant keywords for users whenever they perform a search. Scirus, supported by Elsevier Science and the FAST search engine, is a leading science-specific search engine which covers over 150 million Web pages in the science domain. It also has indexed more than one million articles and abstracts from various journals and magazines.

4.2.2.2. Meta-searcher. The Meta-searcher collects documents from different data sources. It submits search request to each data source through HTTP protocol and extracts the search results from the returned pages. Data sources can be added to or removed from our system. We identified a set of relevant data sources which can be categorized into four categories: (1) scientific and domain-specific Web search engines, (2) online databases, (3) patent databases, and (4) academic abstract and journal databases.

4.2.2.3. Document Summarizer. The Document Summarizer is a tailored version of a summarizer called the Arizona Txtractor, developed and applied in our previous research [6,33]. A Web interface was developed on top of the original Arizona Txtractor. We also customized the Arizona Txtractor for Web documents, which are more unstructured. When a user chooses to summarize a Web document in our system, the system first parses the Web document to remove all content-irrelevant HTML tags such as hyperlinks and scripts and replaces all the HTML special characters with normal characters. The parsed Web document is then passed to the Arizona Txtractor. The Arizona Txtractor utilizes

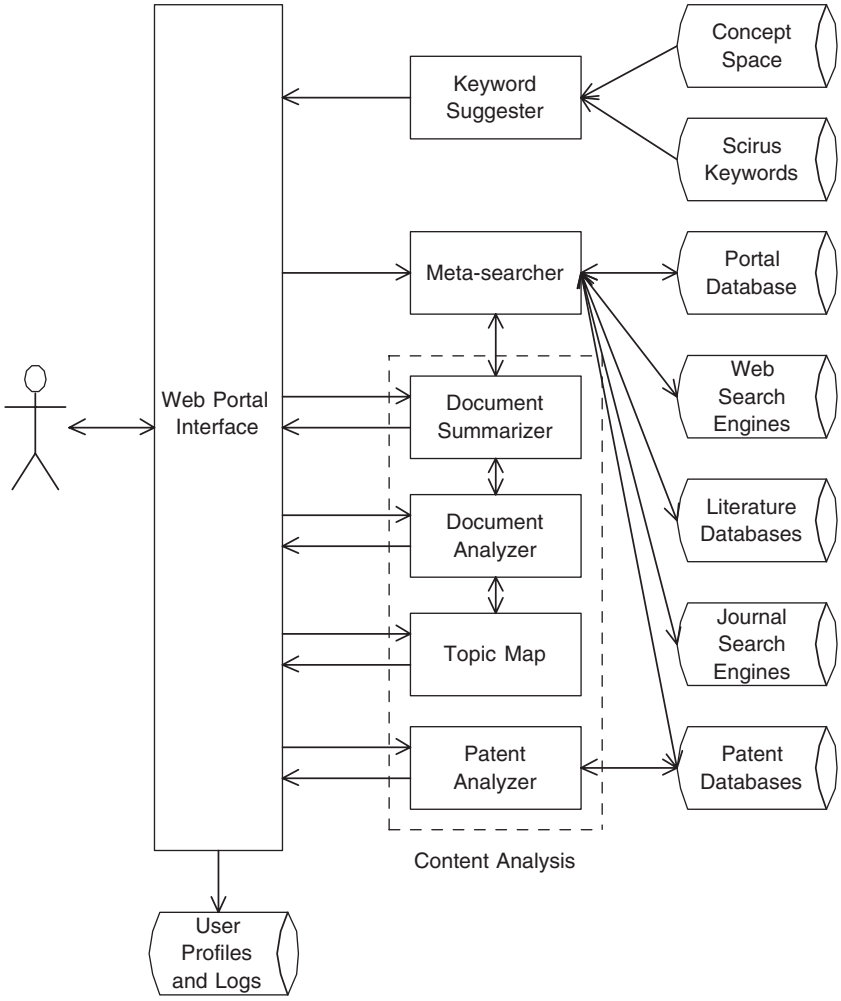

Fig. 3. Proposed Web portal system architecture.

both segmentation and summarization methods to identify representative sentences from the parsed Web document. The Arizona Txtractor has three main steps in its process: (1) sentence evaluation (2) segmentation/topic boundary identification and (3) segment ranking. The first step, sentence evaluation, involves ranking of all the sentences in the parsed document based on the existence of cue phrases in the sentence, the position of a sentence in a paragraph, the frequency of terms in a sentence relative to the document, and the existence of proper nouns. In the second step, the parsed document is processed by a segmentation algorithm similar to TextTiling [21]. The TextTiling algorithm analyzes a document and determines where the topic boundaries (places where the author of the document changes subjects or themes) are located. The text is divided into blocks of equal size and these blocks are compared with each other using a similarity function such as the Jaccard's score. A low similarity score between two adjacent blocks indicates a segment or topic boundary. The last step of the summarization process involves ranking segments based on the evaluation of the sentences in each segment. The program then produces a summary of the parsed Web document by extracting the highest-ranking sentences one by one from the document until the required summary length is reached. Once the summary sentences are extracted, the system locates and highlights these sentences in the original Web document.

4.2.2.4. Document Clusterer. The Document Clusterer processes all Web pages that are retrieved by the Meta-searcher. The Arizona Noun Phraser, developed in our previous research, is used to analyze the Web pages. It extracts and indexes all the noun phrases from each document based on part-of-speech tagging and

linguistic rules [42]. The Arizona Noun Phraser has three components. The *tokenizer* takes Web pages as text input and creates output that conforms to the UPenn Treebank word tokenization rules by separating all punctuation and symbols from text without interfering with textual content. The *tagger* module assigns part-of-speech to every word in the document. The module, called the *phrase generation* module converts the words and associated part-of-speech tags into noun phrases by matching tag patterns to noun phrase pattern given by linguistic rules. For example, the phrase *new research findings* will be considered a valid noun phrase because it matches the rule that an adjective–noun–noun pattern forms a noun phrase. Stop phrases, i.e., phrases that frequently appear in Web pages but do not relate to the chosen scientific domain (such as *home page* or *contact information*) are filtered. The frequency of every phrase is then recorded and the top 20 phrases with the highest frequencies are chosen as the key topics. The key topics are then displayed to the users as folders. The user can view the documents in each folder, i.e., the documents containing the phrase and related to the topic.

4.2.2.5. Topic Map. In addition to the folder display, we also propose to support search result visualization in our architecture using a two-dimensional topic map. The Topic Map component is based on Kohonen's self-organizing map (SOM), a neural network that has been extensively used in document clustering [9,26,31]. The SOM algorithm can automatically cluster documents into different regions on a two-dimensional map. Each document is represented as an input vector of keywords and a two-dimensional grid of output nodes are created. The distance between the input and each output node is then computed and the node with the minimum distance is selected. After the network is trained through repeated presentation of all inputs the documents are submitted to the trained network to form regions of documents with similar contents. Each region is labeled by the phrase which is the key concept that most represents the cluster of documents in that region. More important concepts occupy larger regions, and similar concepts are grouped in a neighborhood [32]. The map is displayed through the User Interface and the user can view the documents in each region.

4.2.2.6. Patent Analyzer. As an important source of scientific and technological knowledge, the well-structured patent documents provide a unique data set for text/document analysis techniques to automatically identify trends and patterns of the development in a scientific domain. In our proposed knowledge portal framework, the Patent Analyzer is an important component that provides such knowledge regarding the development of the domain, supported by integrated text mining, bibliometrics, and visualization techniques.

Being different from other search and analysis components, the Patent Analyzer operates only on the well-structured patent documents. As the major patent databases (United States, European, Japanese, etc.) all provide Web-based access, the Patent Analyzer employs the similar Web page fetching mechanism as the previously described meta-searching components. A specialized meta-searching component is designed as part of the Patent Analyzer to fetch the well-formatted HTML patent documents from the patent databases searching a list of keywords that define the particular scientific and engineering domain. Based on the standard structure of the HTML patent documents provided by the individual patent databases, a specialized patent parser is designed for each database. These patent parsers take as input the raw HTML patent documents and extract fields such as title, abstract, claims, specification, issue date, patent classification fields, assignee country, patent citations, etc., which provides input for the patent analysis framework described below. All parsed results are stored in a local database specifically designed following the information provided by specific patent databases. Most patent databases support time-based search, which enables the Patent Analyzer to update its local patent database on weekly or even daily basis by running the meta-searching and parsing component on newly issued patents.

The proposed Patent Analyzer supports three types of analysis: basic analysis, patent content map analysis and patent citation network analysis [23]. *Basic analysis* refers to the traditional patent analyses that have been widely applied in technology development analysis research and practice. Such analysis is based on basic indicators, such as number of patents, and various indicators derived from the patent citation. Patent content map analysis consists of two types of patent content maps: the overall map and time-series maps. They were generated using the hierarchical multi-level self-organization map algorithm [9,35]. These patent content maps can be used to identify major technology topics and their evolution over time. *Patent citation networks* among different analytical units: countries, institutions and technology fields were visualized using existing network drawing algorithms. We currently generate such networks using an open source graph drawing software, Graphviz, provided by AT&T Labs (available at: http://www.research.att.com/sw/tools/graphviz/) [16]. Such networks visually

present knowledge transfer patterns among countries, institutions and technology fields. All three types of analysis can be restricted to patents of a specified time period. Such time-based analysis capability are important for demonstrating the evolution of performance measures of individual analytical units, knowledge transfer patterns, and the overall trend of scientific and engineering development.

5. Case study—the NanoPort system

To demonstrate the feasibility of applying our approach in building scientific Web portals, we implement a prototype system in the domain of nanoscale science and engineering (NSE). NSE is one of the fastest-growing fields in science and is relatively young. The field has been recognized to be critical to a country's future science and technology competence and has recently attracted global research and development interests. Both long-term basic research and short-term development related to NSE are being actively explored across many scientific fields and industrial applications. The speed and scope of NSE development make it critical for researchers to be aware of progress in the field across different laboratories, companies, industries, and countries. It also encompasses a diversity of research perspectives and application areas such as nanoscale physics, nanoscale medicine, and nanoscale electronics. As such, researchers and scientists in the field have been facing the problems of information overload and vocabulary difference. This makes NSE an ideal domain for testing our proposed architecture. Based on the architecture, we implemented NanoPort, a system designed to be a comprehensive Web portal to serve the researchers, scientists, and practitioners in the NSE domain. In this section, we describe the implementation details and the domain-specific features of NanoPort.

5.1. Content creation

We started our development process by soliciting the user requirements for the NanoPort system. With the help of the National Science Foundation (NSF) and NSE researchers at our university, we conducted interviews with about 20 researchers who have been actively involved in NSE-related research. During the interviews, we identified their research interests, their information flow processes, their information needs, and the problems they faced. We also gathered a set of Web sites, search engines, and databases that are relevant to NSE research.

After the user requirement study, our next step was to collect a database of Web documents that are specific to the NSE domain. Following the proposed architecture, we used a Vertical Spider and a Document Indexer in this process. The Vertical Spider consists of focused spiders and meta-spiders. The focused spiders started with 110 authoritative NSE Web sites selected by the experts as the seed URLs. The meta-spiders connected with three chosen search engines: AltaVista (www.altavista.com), Scirus (www.scirus.com), and NanoSpot (www.nanospot.org), and seven online literature and patent databases: MedLine (www.ncbi.nlm.nih.gov/entrez/query.fcgi), U.S. Patent and Trademark Office (www.uspto.gov), Molecular Expression (micro.magnet.fsu.edu), Science (www.sciencemag.org), MIT Technology Review (www.technologyreview.com), the Proceedings of the National Academy of Sciences (www.pnas.org), and ScienceDirect (www.sciencedirect.com). In each round, a term was randomly chosen from a set of 590 NSE-related terms identified by domain experts and sent to the various search services. The URLs returned by these search services were then added to the URL queue and visited by the spiders. As a result, a collection of around 1,013,000 documents was built, including 580,000 documents from the original 110 starting URLs and 433,000 documents from the starting URLs collected by the meta-spiders. These pages were then stored in our NanoPort Database on a Microsoft SQL Server. Afterwards, each document in our collection was tokenized into words and indexed by the Document Indexer. The searchable index was stored into the NanoPort Database.

5.2. Search and analysis support

To address the needs for information searching and analysis of NSE users, two components in our architecture had to be tailored for the NSE domain—the Keyword Suggester and the Meta-searcher. In this section, we discuss how we customized these two components for NSE-specific contents. The other components, namely the Document Summarizer, Document Clusterer, Topic Map, and Patent Analyzer are domain-independent and do not have to be modified.

5.2.1. Keyword Suggester

In order to provide NSE-related search terms rather than general keywords for users to refine their search queries, we performed co-occurrence analysis on a set of around 1,833,000 documents relevant to NSE research. This set of documents consisted of about 1,461,000

medical abstracts from MedLine, 164,000 documents from INSPEC, and 208,000 documents from BIOSYS. The combined collection contained a total of 1,999,201 terms which were processed by the concept space program. Whenever a user asks for keyword suggestion, the phrases that co-occur most frequently with the search query terms will be suggested to the user. As discussed earlier, we also incorporated the keywords suggested by Scirus. Search queries are forwarded by the Keyword Suggester to the Scirus system to obtain its suggested keywords. Although these keywords are not specific to the NSE domain, they allow users to obtain keywords in some relevant disciplines. Finally, the top 10 keywords obtained from the concept space are displayed to users followed by the top 10 keywords obtained Scirus system. Users can use these terms for query expansion or refinement. We give a higher priority to the concept space terms because the concept space was trained from nanotechnology related corpora and are often more relevant to user's search query, while Scirus was for general purpose and the suggested keywords were less specific.

One should note that the keywords suggested by our system using both approaches are based solely on the search query supplied by the user but not the search results obtained by the system. This is different from some other systems that extract suggested terms from the retrieved document set, and could result in a lower precision in matching the suggested terms with the retrieved documents (but possibly a higher recall in retrieving other documents in the collection).

5.2.2. Meta-searcher

A diverse set of NSE-related information sources were incorporated in our portal. After our preliminary user requirement studies with NSE researchers and scientists, we identified a set of relevant data sources, categorized into the four categories discussed earlier: (1) scientific and NSE-specific Web search engines, (2) online databases, (3) patent databases, and (4) academic abstract and journal databases.

Scientific and NSE-specific search engines include NanoPort Database (our own portal database), NanoSpot (www.nanospot.org), and Scirus (www.scirus.com). NanoPort Database is an NSE specific search engine created by our vertical search engine component as described in Section 5.1. A searchable interface was created and linked to the Meta-searcher. NanoSpot (not to be confused with our system NanoPort despite the name similarity) is a search engine specialized in the NSE domain. It has indexes to the contents of over 25,000 selected Web sites. Scirus covers more than 69 million science-related Web pages including access-controlled sites. It also has indexed millions of articles and abstracts from other database or magazines.

Online databases incorporated in the Meta-searcher include MedLine (www.ncbi.nlm.nih.gov/entrez/query.fcgi), MatWeb (www.matweb.com), Molecular Expression (micro.magnet.fsu.edu), ScienceDirect (www.sciencedirect.com), and Radius (radius.rand.org). MedLine is the National Library of Medicine's premier bibliographic database of international biomedical literature. It includes links to many sites with full text articles and provides access to over 11 million MEDLINE citations. MatWeb is a material information database with data on more than 26,000 materials including metals, plastics, ceramics, and composites. The Molecular Expressions Website provides a searchable collection of color photographs taken through an optical microscope. ScienceDirect searches over 30 million abstracts from scientific articles representing over 1200 peer reviewed academic journals. RaDiUS is a comprehensive database on research and development projects funded by the Federal Government in the United States. Most of these databases provide coverage of general interest articles that may be of relevance to NSE.

The third class of data sources is patent databases. In the NanoPort system, we include the databases of the U.S. Patent and Trademark Office (www.uspto.gov), the European Patent Office (www.european-patent-office.org), the Japanese Patent Office (www.jpo.go.jp), and the World Intellectual Property Organization (www.wipo.org).

The last set of data sources is academic journal and abstract databases that are of relevance to NSE. Currently, we cover Science (www.sciencemag.org), MIT Technology Review (www.technologyreview.com), and the Proceedings of the National Academy of Sciences (www.pnas.org). These journals often cover articles that are relevant to NSE research. More journal and abstract databases will be added in the future.

While most meta-search engines merge the search results from the various data sources into a single list and re-rank the results using their own heuristics, we decided to leave the results in separate categories (still in a single HTML page shown to the users) but we did not merge them into a re-ranked list. The reason is our meta-search results come from a diverse set of data sources and contain many different document types (not only Web pages but also other documents such as journal article abstracts, material data, patent documents, etc.). Merging these together may make it more difficult for users to locate the information they need.

6. Using the NanoPort system

In this section, we present some example user sessions with the NanoPort system. The examples demonstrate how users can utilize the Web searching, Web analysis, and patent analysis functionalities of the system.

6.1. Web searching and search result analysis

After connecting to the Web portal, a user can perform a search by entering the search keyword(s) in the space provided (see Fig. 4). In this example, the user wants to search using the word *nanotube*. The user can choose the preferred search engines, databases, and journals from the lists provided. NanoPort Database, NanoSpot, MedLine, MatWeb, Science, MIT Technology Review, and all patent databases are chosen in our example. After choosing the data sources, the user can also specified various search options, such as the number of search results to be retrieved from each data source, the type of Boolean operators used, and the freshness of the documents retrieved (whenever the option is supported by the selected data sources). The user also can save the search preferences by clicking on the *Save Settings* button, such that the same settings can be loaded when the user logs on to the system in the future.

After specifying all the information, the user can perform a search by clicking the *Find Results* button. The search query is then forwarded to the specified data

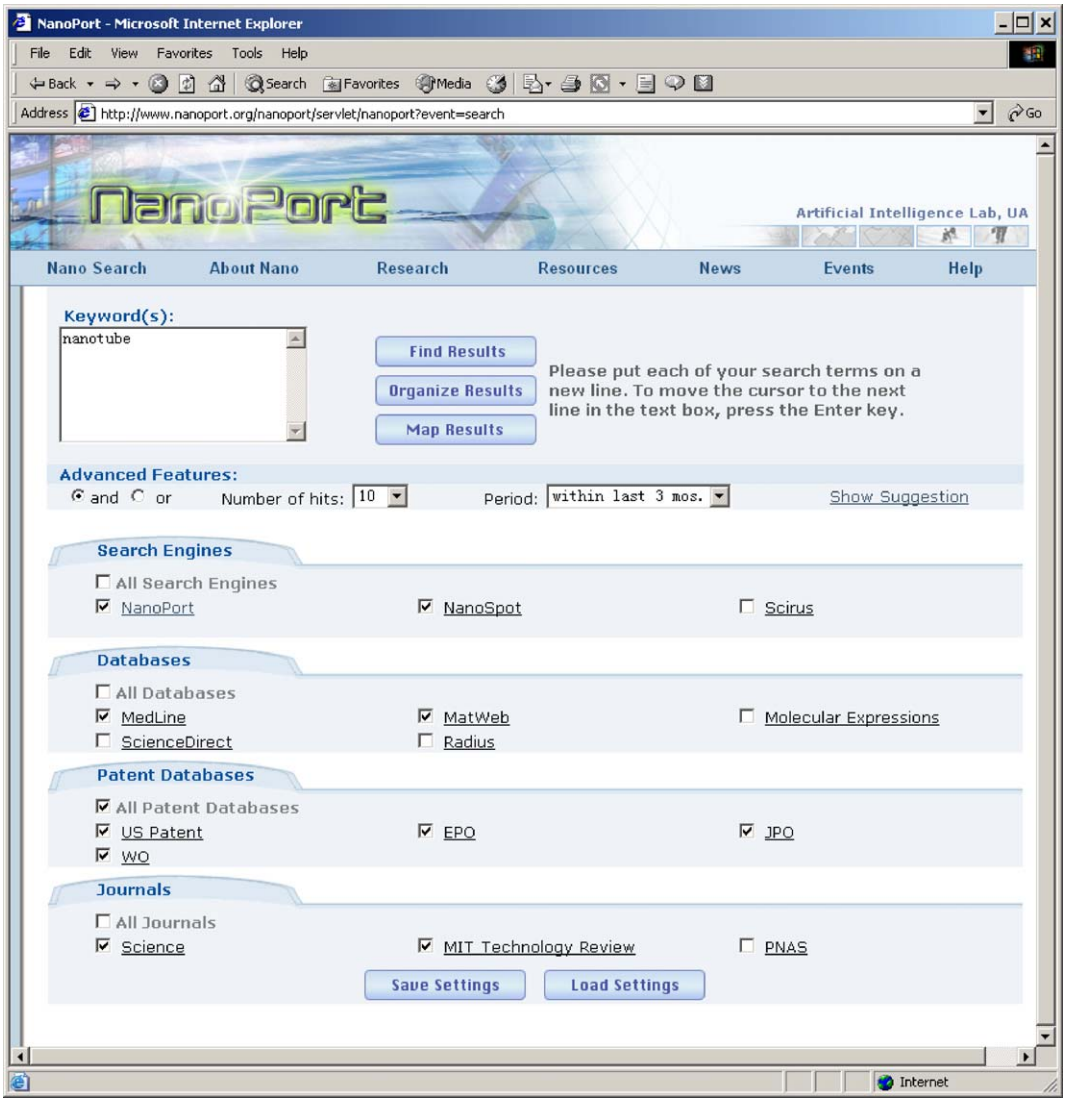

Fig. 4. Main page of the NanoPort system.

source(s) and the search results are displayed to the user (see Fig. 5). Currently, no re-ranking is performed on the search results; the search results are displayed according to the data sources selected. The title and summary of each search result are obtained directly from each data source and displayed to the user. The user can scroll through the search results or click on the name of a data source and directly jump to the search results retrieved from that data source.

A set of relevant keywords (shown on the right hand-side in Fig. 5) also are suggested to the user for expanding or refining the search query. If the user is not satisfied with the current search results, he/she can add any of the suggested terms to the search query, or click on one of these terms to perform a refined search.

The user can also invoke the Document Summarizer to get a dynamic summarization any Web document in the search result list. The user can choose to summarize a page in either three or five sentences by clicking on the respective number. An example of Web document summary is shown in Fig. 6. In the figure, the right-hand side displays the original Web page while the left-hand side shows the three-sentence summary automatically generated by the Document Summarizer.

In addition to the search and the summarization capabilities, the Document Clusterer can be used to provide post-retrieval analysis for the user on the fly. If the user clicks on the *Organize Results* button in the search results screen, the Document Clusterer will provide the user with a list of topics displayed as folders

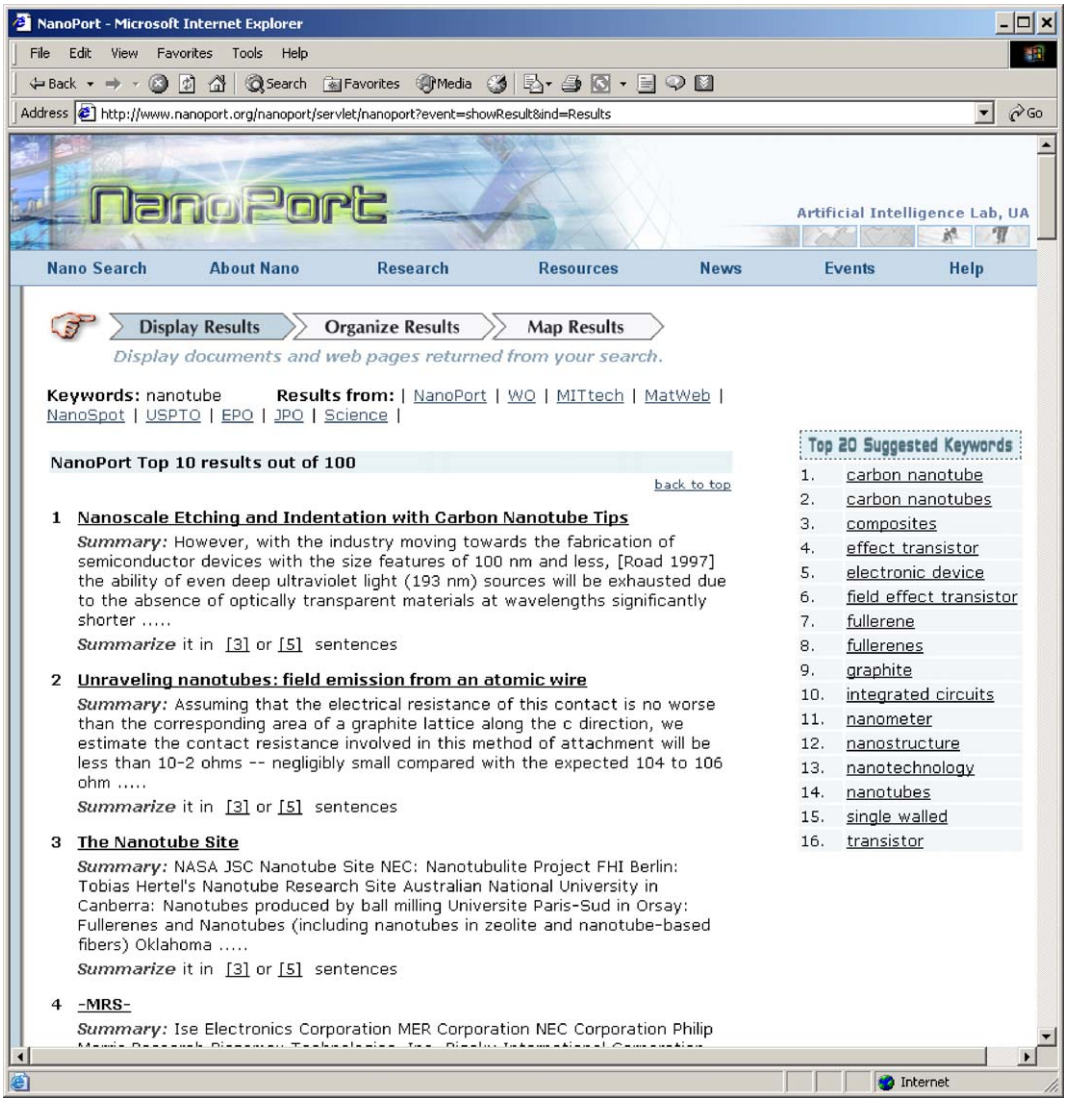

Fig. 5. Search results and suggested keywords.

(see Fig. 7). These topics are decided according to the noun phrases that appear most frequently in the set of documents retrieved. The user can view the number of documents in each topic and is provided links to the documents containing that phrase.

If the user wants to perform further analysis, he or she can click on the *Map Results* button to generate a two-dimensional map produced by the Topic Map component. The map is generated based on the self-organizing map algorithm discussed earlier and it provides the user with an overview of the set of documents retrieved as shown in Fig. 8. In our example, the documents retrieved were clustered into four regions, namely *nanotube computer*, *piezomax technologies*, *electron source*, and *application of nanotubes*. It is especially helpful when the number of documents is large [9]. The user can click on any of the regions to go to the list of documents that contain the corresponding phrases.

6.2. Patent analysis and visualization

Users can perform three types of analyses using the NanoPort Patent Analyzer: basic patent analysis, patent content map analysis, and patent citation network analysis. Currently, these analyses are built on a pre-collected NSE-related patent data set. It was collected from the U.S. Patent and Trademark Office's patent database. Currently we use a keyword-based approach to define NSE-related patents. The data set contains 77,605 U.S. patents issued between 1976 and 2002. The data set involves 69,927 assignees, 123,752 different inventors, 228 different countries, and covers 418 of 462 first-level United States Patent Classification categories.

6.2.1. Basic patent analysis

We have adopted six key indicators of technology development performance from the literature and industrial practice: number of patents, cites per patent, current impact index, technology independency, technology cycle time, and science linkage [34]. Patent-based technology indicators were computed for different analytical units for different time periods to evaluate the overall technology performances of analytical units and their evolution. Such analysis identifies major players in the field. For example, the United States, Japan and France were identified as the most active countries in NSE development, IBM and Xerox had the largest

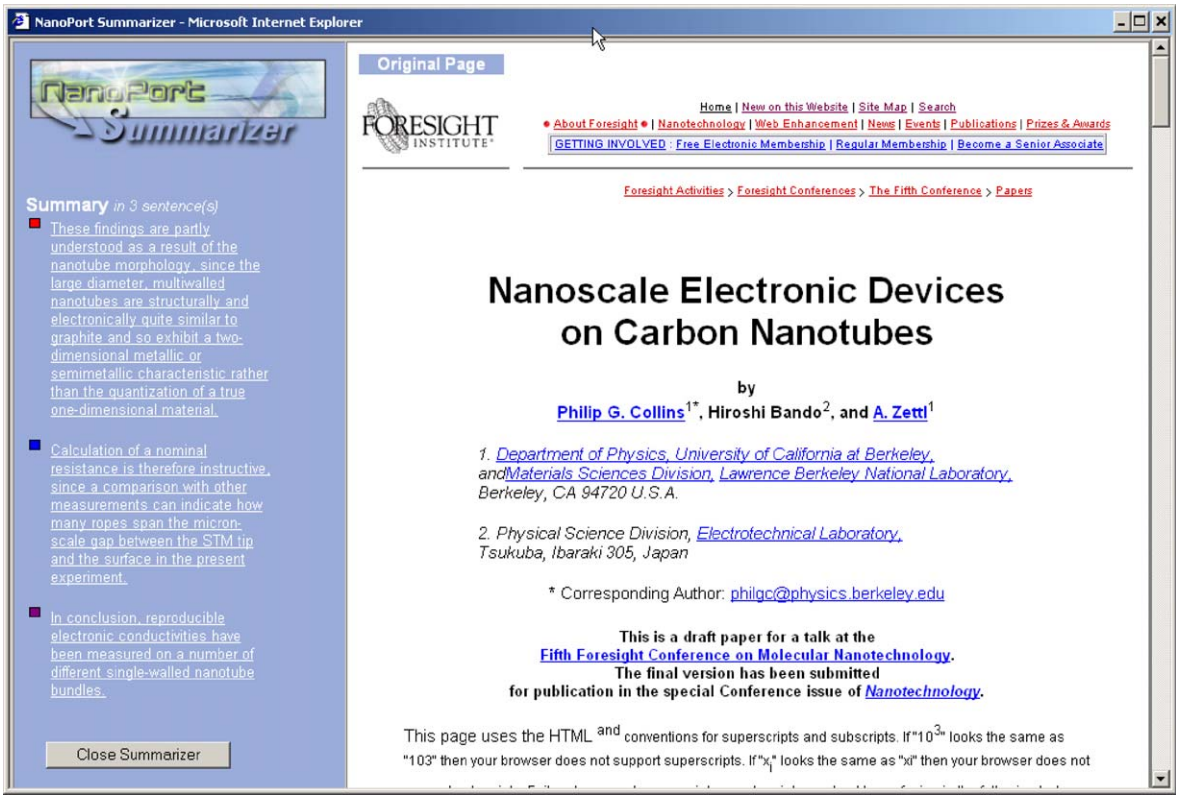

Fig. 6. A Web document and its three-sentence summary.

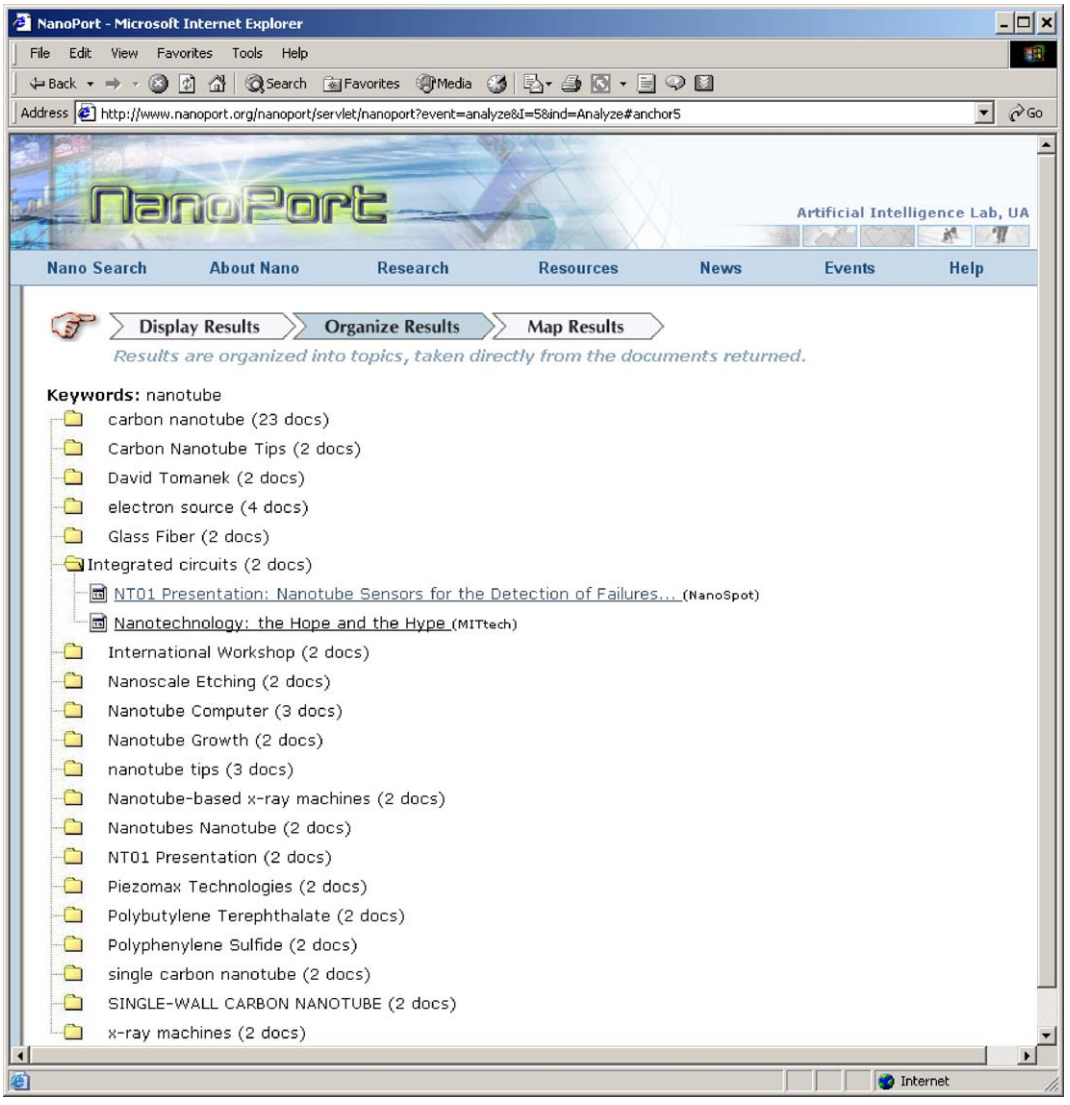

Fig. 7. Document Clusterer showing the topics that appear most frequently.

numbers of NSE-related patents, and 3M had the most influential patents.

6.2.2. Patent content map analysis

The patent content maps provide an alternative organization of the patent documents under hierarchies of dominating technology topics. Two types of patent content maps are provided: an overall content map and a set of time-series maps. Users can use the overall content map to identify major technology topics within the NSE domain, and use time-series content maps to understand the evolution of major technology topics.

The patent content map interface is shown in Fig. 9. The interface contains two components, a folder tree display on the left-hand side and a hierarchical content map in the right-hand side. The patent documents are organized under technology topics that are represented as nodes in the folder tree and colored regions in the content map. These topics were labeled by representative noun phrases that were identified by the heretical self-organizing-map algorithm. Numbers of patent documents that were assigned to the first-level topics are presented in parentheses after the topic labels. Users can either click the fold tree nodes or the content map regions to browse the lower-level topics under a high-level topic. The layers of the colored regions represent the levels of the hierarchies inside the specific regions. The right-hand-side content map display shows all topic regions in the same level under a particular higher-level technology topic region.

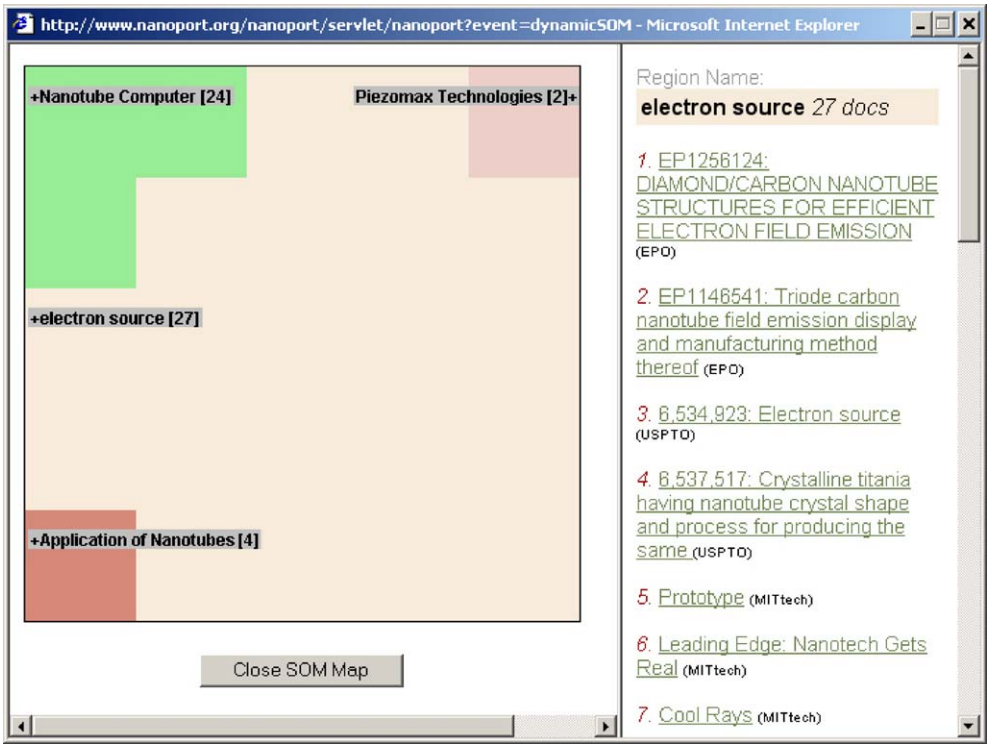

Fig. 8. A two-dimensional topic map.

In each level of such technology maps, conceptually closer technology topics were positioned closer geographically. Conceptual closeness was derived from the co-occurrence patterns of the technology topics in patent titles and abstracts. The sizes of the topic regions also generally correspond to the number of patent documents assigned to the topics [32]. Top-level technology topics of NSE-related patents are shown in Fig. 9. We can observe that closely related technology topics were positioned in neighborhoods (e.g., "ultraviolet radiations," "coating compositions," "electromagnetic radiation," and "optical systems" in the center of the map).

In order to reveal the evolution of major technology topics in the NSE field, we generated content maps for several time periods. Specifically, users can obtain 6 content maps for the time periods of 1976–1980 (3244 patents), 1981–1985 (4601 patents), 1986–1990 (8153 patents), 1991–1995 (10,447 patents), 1996–2000 (27,891 patents), and 2001–2002 (15,524 patents). By comparing the dominating regions in the top-level content maps in different time periods, we can observe some general trends in NSE development.

6.2.3. Patent citation network analysis

A large amount of valuable information is embedded in patent citations. The Patent Analyzer computes and summarizes the citation information for different analytical units: countries, institutions and technology fields.

An example of patent citation networks for countries is shown in Fig. 10. In these networks, arrow direction of the links represents the direction of the citation. For example, a link with the form, "Country A→Country B" means that country A's patents cited country B's patents, and the number besides the link represents the total number of these citations.

Based the citation network shown in Fig. 10, users can observe the following interesting country-level knowledge transferring patterns in the NSE field:

- The United States (US) dominated most of the citations and the U.S. patents intensively interacted with patents of most other countries;
- Japan (JP) was the second largest patent citation center following the United States.
- Other patent citation centers included France (FR), Great Britain (GB and GB2), and Switzerland (CH). There were large amounts of citation activities among the patents of the United States and these countries;
- Patents of Austria (AT), Netherlands Antilles (AN), Germany (DT), Norway (NO), and Singapore (SG) only interacted with the patents of the United States, but not other citation centers;

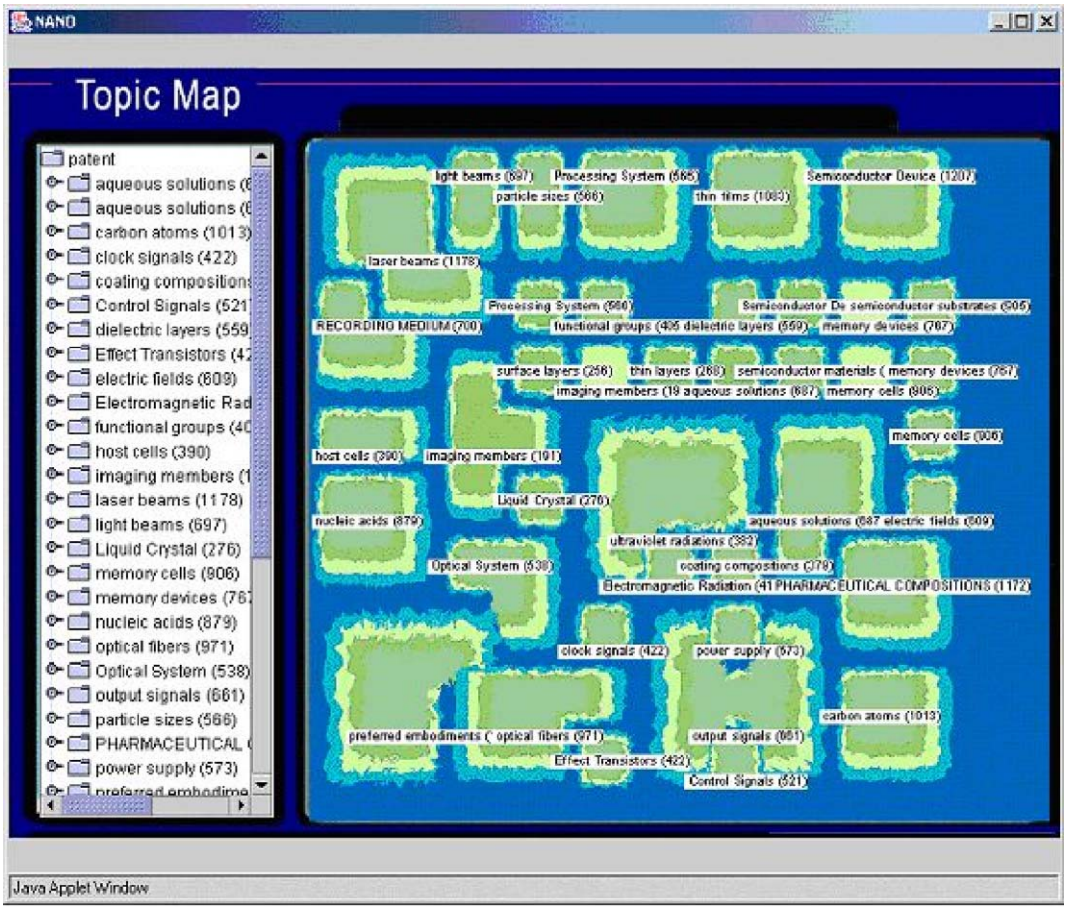

Fig. 9. Top-level patent content map.

- Several local country citation networks can be observed. Groups of the countries that had formed such local networks are: (1) United Kingdom (GB) and England (GB2); (2) France, Sweden (SE), Italy (IT), and Netherlands (NL); and (3) China (Taiwan) (TW) and Korea (KR).

7. System evaluation

In this section, we describe the evaluation studies that we conducted on the NanoPort system. The first study conducted was the system performance study which aimed to evaluate the computational efficiency of the system. In the second study, we conducted a user experiment to evaluate user satisfaction of using the system.

7.1. System performance evaluation

Fast performance is a very important issue for an online portal. If a single search were to last for more than a few minutes, or if the search results are outdated, a user would probably lose interest. In this section, we discuss the static versus dynamic nature of the content in our system, and the time efficiency of several computational expensive components in our system, namely Concept Space, Document Clusterer, and SOM Topic Map. We also present some preliminary performance evaluation results we obtained from testing the components in the prototype NanoPort system.

There are two types of content in our portal: spidered Web pages and results from meta-search engines. Spidered Web pages can be considered static to some extent, because they were collected in advance and their indexing information was stored in database for retrieval. Therefore, spidering has to be conducted frequently to ensure the freshness of the collection. Results from meta-search have a dynamic nature, because they are generated in real time by sending queries to other search engines. However, the results from meta-search can also be outdated because it the meta-search process relies on other data sources which may also be outdated.

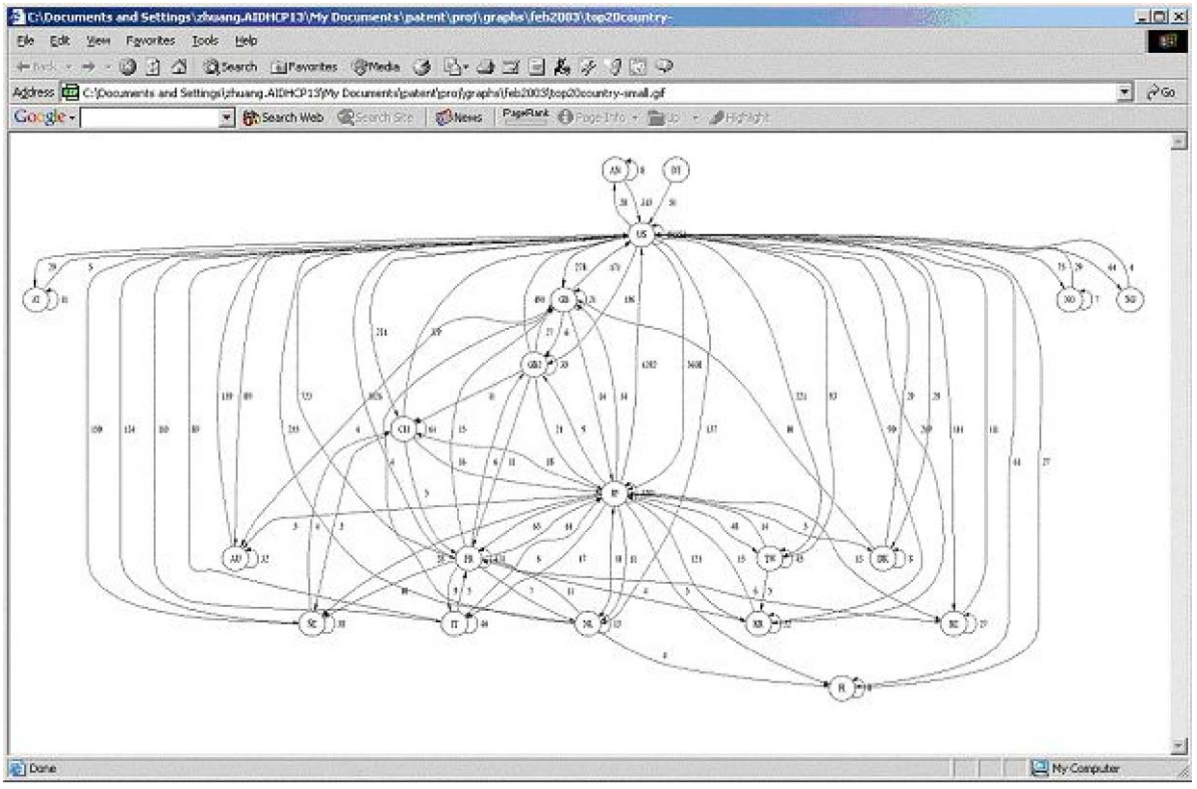

Fig. 10. Patent citation network—countries.

Computationally expensive programs, such as Concept Space, were run beforehand on static content (Web pages collected) to reduce system response time. When new documents were added to the collection, these programs were rerun to keep the information up-to-date. On the other hand, programs used in post-retrieval analysis, such as document cluster and SOM programs, need to be invoked in real time on dynamic results. Their processing time need to be controlled within seconds.

Concept Space is a computationally expensive process with a time complexity of $O(m^2 n)$ where m is the total number of phrases or words and n is the total number of documents. This can be very time-consuming especially when m and n are large. However, since Concept Space is an off-line process and is run at the portal building stage, it will not affect the run-time performance of the portal.

In the Document Clusterer, the Arizona Noun Phraser is the most time-consuming component. Our previous study showed that the Arizona Noun Phraser took approximately 0.08 s to process a journal abstract [42]. In practice, we could limit the response time of Document Clusterer within a few seconds by limiting the number of documents to be clustered (e.g. only the top 100 result documents) or limiting the size of the documents used in the clustering process (e.g.. only index the titles and summaries of the result documents). To evaluate the efficiency of the Document Clusterer in our application, 10 NSE-related queries were selected and tested in the prototype NanoPort system. The average response time of the Document Clusterer is 3.26 s.

The SOM Topic Map is another computational expensive procedure. Our previous study showed that, running on a DEC Alpha 3000/600 workstation with 200 MHz CPU and 128 MBs RAM, the SOM took around 50 s to generate a topic map of 202 short articles [32]. Today's computers are much faster and can generate topic maps more quickly. In our system, we also reduce the response time of the SOM by limiting the number and size of documents to be processed by the SOM. We tested the response time of the SOM Topic Map in the prototype NanoPort system using the same 10 queries selected. The average response time is only 5.87 s. We believe that the delay times of both components are acceptable given the useful functionalities of our portal.

7.2. User study

In order to study how users are satisfied with the system when using it for information search tasks, a

user study was conducted. In this study, we compared NanoPort with Scirus, a leading Web search engine for scientific Web documents. Four search tasks were designed for the experiment based on suggestions given by two NSE experts we consulted. In each search task, each subject was given a topic in nanotechnology and asked to search for relevant Web pages related to the topic. The subject was asked to summarize the findings as a number of themes [5,10,11]. Each subject was required to perform two of the search tasks using NanoPort and the other two using Scirus. Rotation was applied such that the order of search methods and search tasks would not bias our results. Each subject was asked to fill in a post-test questionnaire after performing all the four tasks. Questions related to overall system usability and user satisfaction were included in the questionnaire for both systems, and each subject was asked to give a rating on a Likert scale from 1 (strongly disagree) to 7 (strongly agree).

Fifteen undergraduate participants, who possessed basic knowledge in physics, chemistry and nanotechnology, were recruited as the subjects for the study. Pairwise *t*-tests were also performed to see whether there is any statistical difference between the values obtained by the two systems. Part of the evaluation results is shown in Table 1.

Overall, the NanoPort system was rated higher than Scirus in most aspects of system usability evaluation. Most subjects liked the interface and the analysis components provided by the NanoPort system. However, some of them also commented that the system was quite complicated as there were too many different components.

Table 1
Results of user study

Questionnaire item (1: strongly disagree; 7: strongly agree)	NanoPort	Scirus	*p*-value
I liked the system.	4.80	3.87	0.079*
Overall, I am satisfied with how easy it is to use this system.	4.80	4.33	0.290
I can effectively complete my work using this system.	4.67	3.73	0.074*
I feel comfortable using this system.	4.93	4.27	0.207
It was easy to learn to use this system.	5.13	4.47	0.096*
The organization of information on the system screens is clear.	5.13	4.13	0.092*
The interface of this system is pleasant.	5.33	4.47	0.103
This system has all the functions and capabilities I expect it to have	5.13	3.87	0.020**

* The difference is statistically significant at the 10% level.
** The difference is statistically significant at the 5% level.

8. Discussions

We have presented the architecture design of a scientific Web portal, a prototype system in the NSE domain called NanoPort, and a sample user session of the system. In this section, we compare our prototype with another existing system in the NSE domain, discuss the strengths and weaknesses of our approach, and discuss some design issues.

8.1. Comparison with existing systems

In order to show the effectiveness of our architecture, we compare our prototype system NanoPort with other existing search systems in the NSE domain. The NanoSpot search engine discussed earlier is the only NSE-specific search engine that we have identified. Because of the comprehensive designed used, our architecture has several advantages over the NanoSpot system:

- Query refinement support: NanoPort supports keyword suggestion to assist users in search query formulation and refinement.
- More post-retrieval analysis functionalities: NanoPort allows users to perform document summarization, document clustering, and topic map analysis on the retrieved documents, while NanoSpot does not supports any post-retrieval analysis.
- More comprehensive collection: NanoPort has covered over 1 million Web page in its own database as well as various other resources such as literature databases. NanoSpot only searches for its own database which has a limited (not disclosed) number of Web pages.
- Customized analysis features: Our system provides customized analysis for patents in scientific domains. Such analysis can help identify latest trend in the industry but is not available in most other science search engines.

On the other hand, our architecture also suffers from several weaknesses:

- Dependence on other data sources: Although our system has its own Portal Database, many data sources are maintained by other parties and we have no control over the quality and response time of these sources. We can, however, choose the best data sources to be included and remove those that are low-quality or unstable.
- High computational requirement: Because many of our components rely on artificial intelligence or

machine learning algorithms (such as noun phrasing and self-organizing maps), the computational requirement is much higher than that of other simple search systems. This would limit the scalability of our system and require a powerful server when the number of users increases.

8.2. System design issues

The layered architecture allows us great flexibility in maintaining and modifying our design. It is relatively easy to add new components to the framework, modify any components, or apply the framework to other domains. For example, we can add new a content analysis component to the system easily without having to change the overall architecture. Other databases also can be plugged into the system easily. In addition, as the presentation layer is separated from the application logic and the analysis components, it also does not require much effort to change the user interface of the system. Recently, we have successfully applied the architecture in the medicine and digital government domains [46,48].

Because the system components are mostly written in Java, it is possible for us to run the system on different platforms and operating systems. As some analysis components have high computational requirements, we can run these components on different servers in order to speed up the system's processing time.

By using a server-side Web portal approach, it is easier for developers to maintain and update server-side search engines when compared with client-side tools. Using the server-side approach, any changes to the algorithm or interface can be incorporated into the system, in a way transparent to the users. Users can go to the same Web site and use the system without even knowing that the system has been upgraded or new functionalities have been added. In contrast, a client-side tool is comparatively more difficult to update, because whenever a change is made to the system, each user needs to download and install the software or a patch again. This makes it difficult to upgrade and maintain the system. For example, whenever a data source (e.g., AltaVista) changes its query or result display format so dramatically that the Meta-searcher cannot handle it without being modified, every user must download and install a new component.

A server-side Web portal also has the possibilities of drawing people from various disciplines and countries to come to the same portal to form their online communities. For example, latest news, recent research findings, and conference/event information can be disseminated to users through the Web portal. User search histories and preferences with the Web portal also can be collected over time and be used to identify communities among researchers, educators, and practitioners with similar interests.

9. Conclusion and future directions

In this paper, we report our design of a Web portal architecture for information retrieval, analysis and visualization in scientific domains. To validate our approach, we implemented a prototype Web portal system in the NSE domain called NanoPort. Our framework provides an integrated approach to building Web-based information retrieval and analysis systems that incorporate various techniques and functionalities including collection building, meta-searching, keyword suggestion, and content analysis techniques such as document summarization, document clustering, topic map visualization, and patent analysis. Our layered architecture design also allows developers to maintain the system, revise user interface, or add new components easily. We also demonstrated the feasibility of using such an integrated approach to address the information needs of NSE researchers, thus achieving our goal of helping them search more effectively and efficiently for relevant information using various techniques through our Web portal. Because the techniques employed are not domain-specific, only minimal effort would be required to apply the architecture to different areas of science and engineering, such as bioinformatics, micro-electronic devices, future-generation wireless communications, etc.

Our future work will be carried out in several directions. First, more data sources, such as relevant journal databases, will be added to the NanoPort system to enhance its meta-searching capability. We also plan to perform a large-scale evaluation on the system to study how the system can address the information needs of users in the NSE domain. First, we will obtain a set of sample queries from the NSE community. The evaluation will be based mainly on precision and recall measures and the search portal will be compared against current state-of-the-art search engines like Google and domain-specific search engines like NanoSpot. Second, a qualitative study will also be conducted and access to NanoPort from the Web will be granted to participating researchers and students, as well as the general public. The users will be asked to switch between their favorite search tool and our portal and to use the same query for both tools should the results from one of them be unsatisfactory. Subjects will be asked to fill out

questionnaires covering general items related to the various components, the search functionalities, the contents, and the overall usability of the portal. We will also explore the broader social and educational issues relating to a Web portal for a young and growing discipline like NSE.

On the technical side, we are studying how to improve the individual components in the Web portal architecture. First, we are currently investigating how different optimization and graph search techniques, such as the genetic algorithms, can be applied to improve the performance of the vertical spider in the content creation process. Second, we plan to integrate the Patent Analyzer component into the user interface in the near future, which will allow users to select a subset of patents of interest to perform all three types of patent analyses. Third, we are planning to investigate how to utilize the metadata that are available in some documents for better analysis (e.g., keyword suggestion and clustering). With the growing popularity of the Semantic Web and XML documents, such enhanced capabilities will be very helpful. Finally, we are also exploring new analysis and visualization techniques for identifying new trends and summarizing the technology evolution history.

Acknowledgements

This research has been supported in part by the following grants:

- NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999–March 2002.
- NSF/NSE/SGER, "NanoPort: Intelligent Web Searching for Nanoscale Science and Engineering," CTS-0204375, February 2002–November 2002.
- NSF/IIS/SGER, "Intelligent Patent Analysis and Visualization," IIS-0311628, May 2003–April 2004.
- HKU CRCG, "Using Content and Link Analysis in Developing Domain-specific Web Search Engines: A Machine Learning Approach," HKU Seed Funding for Basic Research (10205294), February 2004–July 2005.

We would like to thank Dr. Mihail Roco and Dr. Steve Goldstein of NSF for their support and advices throughout the project. We would also like to thank Art Purcell of the U.S. Patent and Trademark Office for helping us access the U.S. patent database. We also appreciate the help from the professors and students with expertise in NSE from various departments at the University of Arizona, especially Dr. Paul Calvert, Dr. David Galbraith, Dr. Poul Jessen, Yi Yang, Wei Gao, and Wayne Huang. We thank Boby Shiu of the University of Hong Kong for his help in conducting the user study. Last but not the least, we thank all members of the Artificial Intelligence Lab at the University of Arizona who have contributed to the project, in particular Alan Yip, Yongchi Chen, Chunju Tseng, Ann Lally, Wai-Ki Sung, Yi Qin, Fei Guo, Xiaoyun Sun, Hui Liu, Daniel McDonald, Pei He, Gavin Ng, Benjamin Smith, Zhi-kai Chen, and Matthew Landon.

References

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, S. Raghavan, Searching the web, ACM Transactions on Internet Technology 1 (1) (2001) 2–43.
[2] C.M. Bowman, P.B. Danzig, U. Manber, F. Schwartz, Scalable internet resource discovery: research problems and approaches, Communications of the ACM 37 (8) (August 1994) 98–107.
[3] M. Chau, H. Chen, Comparison of three vertical search spiders, IEEE Computer 36 (5) (2003) 56–62.
[4] M. Chau, H. Chen, Personalized and focused web spiders, in: N. Zhong, J. Liu, Y. Yao (Eds.), Web Intelligence, Springer-Verlag, February 2003, pp. 197–217.
[5] M. Chau, H. Chen, D. Zeng, Personalized spiders for web search and analysis, Proceedings of the 1st Joint Conference on Digital Libraries, Roanoke, Virginia, June, 2001, pp. 79–87.
[6] M. Chau, H. Chen, J. Qin, Y. Zhou, Y. Qin, W.K. Sung, D. McDonald, Comparison of two approaches to building a vertical search tool: a case study in the Nanotechnology Domain, Proceedings of the 2nd Joint Conference on Digital Libraries, Portland, Oregon, July, 2002, pp. 135–144.
[7] H. Chen, Collaborative systems: solving the vocabulary problem, IEEE Computer, Special Issue on Computer-Supported Cooperative Work (CSCW), vol. 27(5), May 1994, pp. 58–66.
[8] H. Chen, K.J. Lynch, Automatic construction of networks of concepts characterizing document databases, IEEE Transactions on Systems, Man, and Cybernetics 22 (5) (1992) 885–902.
[9] H. Chen, C. Schuffels, R. Orwig, Internet categorization and search: a machine learning approach, Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries, vol. 7(1), 1996, pp. 88–102.
[10] H. Chen, H. Fan, M. Chau, D. Zeng, MetaSpider: meta-searching and categorization on the web, Journal of the American Society for Information Science and Technology 52 (13) (2001) 1134–1147.
[11] H. Chen, M. Chau, D. Zeng, CI Spider: a tool for competitive intelligence on the web, Decision Support Systems 34 (1) (2002) 1–17.
[12] H. Chen, A.M. Lally, B. Zhu, M. Chau, HelpfulMed: intelligent searching for medical information over the internet, Journal of the American Society for Information Science and Technology 54 (7) (2003) 683–694.
[13] F.C. Cheong, Internet Agents: Spiders, Wanderers, Brokers, and Bots, New Riders Publishing, Indianapolis, Indiana, USA, 1996.
[14] J. Courteau, Genome databases, Science 254 (October 1991) 201–207.

[15] G.W. Furnas, T.K. Landauer, L.M. Gomez, S.T. Dumais, The vocabulary problem in human-system communication, Communications of the ACM 30 (11) (November 1987) 964–971.
[16] E. Gansner, S. North, An open graph visualization system and its applications to software engineering, Software, Practice and Experience 30 (11) (2000) 1203–1233.
[17] E. Garfield, Citation indexes for science: a new dimension in documentation through association of ideas, Science 122 (1955) 108–111.
[18] E. Garfield, Citation Indexing: Its Theory and Application in Science, Technology and Humanities, John Wiley, New York, 1979.
[19] H. Grupp, U. Schomch, Perception of scientification of innovation as measured by referencing between patents and papers, in: H. Grupp (Ed.), Dynamics of Science-based Innovation, Spring Verlag, Heidelberg, 1992.
[20] E. Hassan, Simultaneous mapping of interactions between scientific and technological knowledge bases: the case of space communications, Journal of the American Society for Information Science and Technology 54 (5) (2003) 462–468.
[21] M.A. Hearst, TextTiling: segmenting text into multi-paragraph subtopics passages, Computational Linguistics 23 (1) (1997) 33–64.
[22] D.R. Hill, A vector clustering technique, in: Samuelson (Ed.), Mechanized Information Storage, Retrieval and Dissemination, North-Holland, Amsterdam, 1968.
[23] Z. Huang, H. Chen, A. Yip, G. Ng, F. Guo, Z.-K. Chen, M.C. Roco, Longitudinal patent analysis for nanoscale science and engineering: country, institution and technology field, Journal of Nanoparticle Research 5 (2003) 333–363.
[24] T. Joachims, Text categorization with support vector machines: learning with many relevant features, Proceedings of the European Conference on Machine Learning, Berlin, 1998, pp. 137–142.
[25] M.M. Karki, Patent citation analysis: a policy analysis tool, World Patent Information 19 (1997) 269–272.
[26] T. Kohonen, Self-organizing Maps, Springer-Verlag, Berlin, 1995.
[27] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, vol. 11(3), May 2000, pp. 574–585.
[28] S.L.Y Lam, D.L. Lee, Feature reduction for neural network based text categorization, Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA '99), Hsinchu, Taiwan, Apr, 1999.
[29] L.S. Larkey, A patent search and classification system, Proceedings of the Fourth ACM Conference on Digital Libraries, 1999, pp. 79–87.
[30] S. Lawrence, C.L. Giles, Accessibility of information on the web, Nature 400 (1999) 107–109.
[31] X. Lin, D. Soergel, G. Marchionini, A self-organizing semantic map for information retrieval, Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIRIR'91), 1991, pp. 262–269.
[32] C. Lin, H. Chen, J. Nunamaker, Verifying the proximity and size hypothesis for self-organizing maps, Journal of Management Information Systems 16 (3) (2000) 61–73.
[33] D. McDonald, H. Chen, Using sentence-selection heuristics to rank text segments in TXTRACTOR, Proceedings of the 2nd Joint Conference on Digital Libraries, Portland, Oregon, July, 2002, pp. 28–35.
[34] F. Narin, Tech-line Background Paper, CHI Research, Inc., 2000. Available at: http://www.chiresearch.com/techline/tlbp.pdf.
[35] T.-H. Ong, H. Chen, W.-K. Sung, B. Zhu, Newsmap: a knowledge map for online news, Decision Support Systems, Special Issue on Collaborative Work and Knowledge Management in Electronic Business, vol. 39(4), 2003, pp. 583–597.
[36] C. Oppenheim, Do patent citations count? in: B. Cronin, H.B. Atkins (Eds.), The Web of Knowledge, Information Today, Inc., Medford, 2000.
[37] E. Rasmussen, Clustering Algorithms, Prentice Hall, Englewood Cliffs, NJ, 1992.
[38] J.J. Rocchio, Document Retrieval Systems—Optimization and Evaluation. Ph.D. Thesis, Harvard University, 1966.
[39] G. Salton, Another look at automatic text-retrieval systems, Communications of the ACM 29 (7) (1986) 648–656.
[40] G. Salton, Automatic Text Processing, Addison-Wesley, Reading, MA, 1989.
[41] E. Selberg, O. Etzioni, Multi-service search and comparison using the metaCrawler, Proceedings of the 4th World Wide Web Conference (WWW4), 1995.
[42] K.M. Tolle, H. Chen, Comparing noun phrasing techniques for use with medical digital library tools, Journal of the American Society for Information Science 51 (4) (2000) 352–370.
[43] J. Ward, Hierarchical grouping to optimize an objection function, Journal of the American Statistical Association 58 (1963) 236–244.
[44] E. Wiener, J.O. Pedersen, A.S. Weigend, A neural network approach to topic spotting, Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.
[45] Y. Yang, X. Liu, A re-examination of text categorization methods, Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'99), 1999, pp. 42–49.
[46] C.Q. Yin, L.D. Nickels, C.Z. Chen, T.G. Ng, H. Chen, DGPort: a web portal for digital government, Proceedings of the National Conference on Digital Government Research, Boston, Massachusetts, USA, May, 2003.
[47] O. Zamir, O. Etzioni, Web document clustering: a feasibility demonstration, Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'98), Melbourne, Australia, Aug, 1998, pp. 46–54.
[48] Y. Zhou, J. Qin, H. Chen, Z. Huang, Y. Zhang, W. Chung, G. Wang, CMedPort: a cross-regional Chinese Medical Portal, Proceedings of the 3rd Joint Conference on Digital Libraries, Houston, Texas, USA, May, 2003, p. 379.

Michael Chau is a Research Assistant Professor in the School of Business at the University of Hong Kong. He received his Ph.D. degree in management information systems from the University of Arizona and a bachelor degree in computer science and information systems from the University of Hong Kong. His current research interests include information retrieval, Web mining, data mining, knowledge management, electronic commerce, security informatics, and intelligence agents. He has published more than 40 research articles in leading journals and conferences, including *IEEE Computer, Journal of the America Society for Information Science and Technology, Decision Support Systems*, and *Communications of the ACM*. More information can be found at http://www.business.hku.hk/~mchau/.

Zan Huang is an Assistant Professor of the Department of Supply Chain and Information Systems at the Pennsylvania State University. His research interests include recommender systems, data mining and text mining for bioinformatics and financial applications, knowledge management technologies, and experimental economics-related research for electronic markets. His articles have appeared in *ACM Transactions on Information Systems, IEEE Intelligent Systems, Journal of the American Society for Information Science & Technology, Decision Support Systems, Journal of Nanoparticle Research* and other publications. He received the B.Eng. degree in Management Information Systems from Tsinghua University, Beijing, China and the Ph.D. degree in Management Information Systems from the University of Arizona.

Jialun Qin is a Ph.D. Candidate in the Department of Management Information Systems at the University of Arizona. His research interests include knowledge management, data and Web mining, digital libraries, and human computer interaction. His publications have appeared in Decision Support Systems, Journal of the American Society for Information Science and Technology, and IEEE Intelligent Systems.

Yilu Zhou is a doctoral candidate in the Department of Management Information Systems at the University of Arizona, where she is also a research associate of the Artificial Intelligence Lab. Her current research interests include multilingual knowledge discovery, Web mining and human computer interaction. She received a B.S. in Computer Science from Shanghai Jiaotong University. Contact her at yilu@u.arizona.edu.

Hsinchun Chen is McClelland Professor of Management Information Systems at the University of Arizona and Andersen Consulting Professor of the Year (1999). He received the B.S. degree from the National Chiao-Tung University in Taiwan, the MBA degree from SUNY Buffalo, and the Ph.D. degree in Information Systems from the New York University. He is author/editor of 10 books and more than 130 SCI journal articles covering intelligence analysis, biomedical informatics, data/text/web mining, digital library, knowledge management, and Web computing. He serves on the editorial board of *ACM Transactions on Information Systems, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Systems, Man, and Cybernetics, Journal of the American Society for Information Science and Technology, and Decision Support Systems*.

Dr. Chen is a Scientific Counselor/Advisor of the National Library of Medicine (USA), Academia Sinica (Taiwan), and National Library of China (China), and has served as an advisor for major NSF, DOJ, NLM, and other international research programs in digital library, digital government, medical informatics, and national security research. Dr. Chen is the founding director of Artificial Intelligence Lab and Hoffman E-Commerce Lab. Dr. Chen is the conference co-chair of ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2004 and (founding) conference co-chair of the IEEE International Conferences on Intelligence and Security Informatics (ISI). Dr. Chen has also received numerous awards in information technology and knowledge management education and research including: AT&T Foundation Award, SAP Award, the Andersen Consulting Professor of the Year Award, the University of Arizona Technology Innovation Award, and the National Chaio-Tung University Distinguished Alumnus Award. Further information can be found at http://ai.arizona.edu/hchen/.