

CMedPort: An integrated approach to facilitating Chinese medical information seeking

Yilu Zhou ^{*}, Jialun Qin, Hsinchun Chen

Department of Management Information Systems, The University of Arizona, 1130 E Helen St. 430W, Tucson, AZ 85721, USA

Received 10 March 2004; received in revised form 17 November 2005; accepted 18 November 2005

Available online 18 January 2006

Abstract

As the number of non-English resources available on the Web is increasing rapidly, developing information retrieval techniques for non-English languages is becoming an urgent and challenging issue. In this research to facilitate information seeking in a multilingual world, we focused on discovering how search-engine techniques developed for English could be generalized for use with other languages. We proposed a general framework incorporating a focused collection-building technique, a generic language processing ability, an integration of information resources, and a post-retrieval analysis module. Based on this approach, we developed CMedPort, a Chinese Web portal in the medical domain that not only allows users to search for Web pages from local collections and meta-search engines but also provides encoding conversion between simplified and traditional Chinese to support cross-regional search and document summarization and categorization. User studies were conducted to compare the effectiveness and efficiency of CMedPort with those of three major Chinese search engines. Results indicate that CMedPort achieved similar accuracy for search tasks, but exhibited significantly higher recall than each of the three search engines as well as higher precision than two of the search engines for browse tasks. There were no significant differences among the efficiency measures for CMedPort and benchmarks systems. A post-questionnaire regarding system usability indicated that CMedPort achieved significantly higher user satisfaction than any of the three benchmark systems. The subjects especially liked CMedPort's categorizer, commenting that it helped improve understanding of search results. These encouraging outcomes suggest a promising future for applying our approach to Internet searching and browsing in a multilingual world.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Information retrieval; Internet searching and browsing; Search engine; Cross-regional search; Meta-search; Summarization; Categorization

1. Introduction

Rapid growth of the World Wide Web has produced a wealth of information in almost every major language, making the Internet a truly multilingual world. A report published by Internet World Stats ([http://](http://www.internetworldstats.com/stats7.htm)

www.internetworldstats.com/stats7.htm) in September of 2005 showed that the majority of total Internet users are from non-English speaking areas (68.3%). Furthermore, the population of non-English-speaking Internet users is growing much faster than that of English-speaking users. As a result, study of Internet searching and browsing support has become an interesting and challenging research problem in a multilingual world.

While the Web provides convenient information searching, users often face an information overload

^{*} Corresponding author.

E-mail addresses: yilu@u.arizona.edu (Y. Zhou), qin@u.arizona.edu (J. Qin), hchen@eller.arizona.edu (H. Chen).

problem [3]; a search employing a general-purpose search engine such as Google can result in thousands of hits. This problem is even more serious for non-English-speaking users because most Internet searching and browsing techniques are developed for speakers of English. A generic approach to facilitating Internet searching and browsing in any language is sorely needed.

Other than information overload, there are also some additional problems specifically faced by non-English speakers. As the second most often used language on the Web, Chinese provides a good example. Chinese users in mainland China, Hong Kong, and Taiwan make up 12.2% of the total world online population. Although Internet users in these 3 regions share enormous common information needs, Chinese search engine developers usually find it difficult to provide information from all regions because of problems such as the diversity of data sources and encoding differences. This greatly hinders information sharing among users from different regions. A generic Internet searching and browsing approach is needed to address these problems.

We proposed an integrated approach to facilitating Internet searching and browsing in non-English languages. We developed an experimental Chinese medical Web portal, called *CMedPort*, based on our proposed approach. We also conducted user studies to evaluate the performance of the proposed approach in assisting human information seeking behavior.

The remainder of the paper is structured as follows. Section 2 reviews related research, including studies on information seeking behavior, technologies that support searching and browsing, and search engines and medical portals in English and Chinese. In Section 3 we present our research questions. Section 4 describes our research testbed. In Section 5, we propose a generic framework to support information seeking and present the architectural design and major components of a prototype system: *CMedPort*. Section 6 describes our experimental design. In Section 7 we discuss experimental results and lessons learned. Finally, in Section 8, we conclude our study and suggest some future directions.

2. Related work

In this section, we review various issues related to Internet searching and browsing in a multilingual world. These include different information-seeking behaviors on the Internet, different Internet searching techniques, and some special issues in supporting Chinese Internet searching and browsing.

2.1. Information seeking behaviors on the internet: searching and browsing

Information seeking can be viewed as a “process in which humans purposefully engage in order to change their state of knowledge” [28]. A significant amount of research contributes to more recent information-seeking models on the World Wide Web. Ellis [17] proposed a general model of information seeking behaviors with six categories: starting, chaining, browsing, differentiating, monitoring, and extracting. Kuhlthau [22] provided a model that focuses on the information search process from the user’s perspective. Belkin’s model [2] addresses the idea of cognitive and situational aspects as the reason for information seeking. More recently, Meho and Tibbo [33] further expanded Ellis’ model and identified four additional information seeking behaviors: extracting, verifying, networking, and information management. These studies show that users’ particular patterns of information seeking enter into users’ mental models of search. Among all the information seeking behaviors identified, two typical types of Internet behavior have received the most attention from researchers: searching and browsing [4,9,29].

Internet searching is a process in which an information seeker uses a query to describe a request for information and the system must locate information that matches or satisfies that request [8]. Through searching, individuals seek to retrieve either specific information on a given topic or a specific piece of information. In other words, he or she has a specific object or target in mind. Internet browsing has been defined by Marchionini and Shneiderman as “an exploratory, information seeking strategy that depends upon serendipity” and is “especially appropriate for ill-defined problems and for exploring new task domains” [29]. Through browsing, individuals explore an information space to gain familiarity with it or to locate something of interest to them [8]. Browsing reflects a mental model in which the target is comparatively unfocused.

2.2. Techniques facilitating information seeking

Understanding of the information seeking process helps researchers develop tools and techniques to augment the process. The wealth of information available on the Internet has led to much research directed toward developing techniques to support Internet searching and browsing. In this section, we briefly review some of their strengths and weaknesses.

2.2.1. Web search engines

2.2.1.1. General-purpose search engines. General-purpose search engines are the most popular tools to help users locate information on the Web. A search engine usually consists of the following components: (1) Spiders (a.k.a. Crawlers) to retrieve Web pages by recursively following URL links, (2) an Indexer to tokenize Web pages into words or phrases, (3) a Query and Ranking Engine to retrieve search results, and (4) a User Interface [6,7].

Currently, many general-purpose search engines are available on the Web, each having its own characteristics and preferred algorithm for indexing, ranking, and presenting Web documents. For example, AltaVista and Google allow users to submit queries and retrieve Web pages in a ranked order, while Yahoo 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 overwhelmingly numerous and imprecise. It is often difficult to obtain specialized, domain-specific information from them.

2.2.1.2. Domain-specific search engines. Many domain-specific search engines (or vertical search engines) seek to support more effective searching by providing precise search results in particular domains and extra functionalities that are not possible with general search engines [30]. For example, LawCrawler (<http://www.lawcrawler.com>) allows users to search for legal information. CampSearch (<http://www.campsearch.com>) searches for summer camps for children and adults and Excite NewsTracker (<http://nt.excite.com>) searches for news articles.

A good domain-specific search engine should contain as many relevant, high-quality pages and as few irrelevant, low-quality pages as possible. To address this need, domain-specific search engines collect Web pages by using intelligent spiders that can predict whether a URL is likely to point to relevant material and thus should be fetched first. Some examples of algorithms that have been developed to guide spiders to locate Web pages relevant to desired domains include HITS [5], PageRank [14], Hopfield Net [6], and Reinforcement Learning [32].

2.2.1.3. Meta-search engines. Selberg and Etzioni [40] suggested that by relying on a single search engine, users could miss over 77% of the references they would find most relevant. A study by NEC

Research Institute drew some similar conclusions, revealing that Internet search engines cannot keep up with the Internet's growth and that each search engine studied covered only about 16% of total available Web sites [24]. However, meta-search engines leverage the capabilities of multiple Web search engines and other types of information sources, providing a simple, uniform user interface and relieving the user of having to deal with different search engines and information overload [9,35,41]. For instance, SavvySearch (<http://www.savvysearch.com>) supports up to 100 engines and allows users to customize a selection of search engines. MetaCrawler (<http://www.metacrawler.com>) searches the Internet's top search engines, including Google, Yahoo, AltaVista, and Ask Jeeves. Copernic Agent (<http://www.copernic.com>) collates results from more than 1000 search engines and provides filtering, summarization, and analysis on search results. MedTextus [26] and HelpfulMed [10] are two medical search engines that meta-search Web pages as well as online medical databases and journals.

2.2.2. Post-retrieval analysis

In most current search engine systems, returned results are presented as lists of ranked URLs. Post-retrieval analysis can be performed on a result list to help users quickly locate the information needed. Document categorization and summarization are the two major techniques used in post-retrieval analysis.

2.2.2.1. Summarization—document preview. Summarization is a post-retrieval analysis technique that provides previews of documents [18]. It can reduce the size and complexity of Web document lists by offering concise representations of the documents returned [31,32]. In a browsing scenario, summarization provides an efficient way to allow users to judge the relevance of a document and let the users decide whether or not the full text is worth viewing [39]. Two major approaches to text summarization are text abstraction and text extraction. Text abstraction, which generates grammatical sentences that summarize a document, involves a great degree of document processing and computation. Text extraction utilizes sentences from an original document to form a summary. It usually involves assigning importance scores to sentences, based on term frequency and other characteristics of the document, and top-scoring sentences are selected as a summary. Recent research in text summarization has focused on the text extraction approach [21,31].

2.2.2.2. Categorization—document overview. Users are often frustrated by long-list results returned by search engines. In a browsing scenario, it is highly desirable for a search system to provide an overview of the retrieved document set so that a user can explore a specific topic and gain a general view of a particular area of interest. Categorization has been shown to be a powerful post-retrieval document processing tool that can cluster similar documents and present the resulting clusters to the user in an intuitive and sensible way [9]. Hearst and Pedersen [20] and Zamir and Etzioni [45] demonstrated that document categorization has the potential to improve performance in document retrieval. Document categorization is based on the Cluster Hypothesis: “closely associated documents tend to be relevant to the same requests” [38]. There are two approaches to categorization [45]. It can be based on individual document attributes, such as query term frequency, size, source, topic, or author for each document. NorthernLight (<http://www.northernlight.com>) is an example of this approach. Categorization can also be based on inter-document similarities. This approach usually includes some machine learning techniques. For example, the self-organizing map (SOM), which uses a neural network algorithm to cluster documents, has been incorporated in several information retrieval systems [8]. In both approaches, documents need to be systematically segmented and indexed so that key phrases can be identified. Examples of key phrase extraction techniques include AZ Noun Phraser [43] and Mutual Information [37].

2.2.3. Chinese text processing

Textual information collected from the Web must first be processed by document indexing techniques. Indexing techniques have been widely studied for English documents, but these techniques do not apply to languages such as Chinese in which there are no explicit separators to indicate word boundaries [13]. Kwok [23] investigated three indexing techniques for Chinese that can be applied to virtually any languages without explicit boundaries: character-based (or 1-gram), bi-gram, and lexicon-based indexing. His results showed that character-based indexing is good and efficient while bi-gram and simple word-based indexing achieved higher precision. However, bi-gram indexing led to a large indexing term space and was not efficient. Lexicon-based indexing usually matched text to an existing word lexicon, but many valuable words could be missed if they were not included in the matching lexicon.

2.3. Information seeking in a multilingual world: research gaps

The Web provides convenient information searching and browsing. As more and more users on the Internet are from non-English-speaking countries, major search engines have attempted to provide search support in non-English languages. However, search engines originally designed for English speakers usually cover a very limited amount of non-English content and cannot serve the information need of the fast growing non-English-speaking online population. Although some English online search engines have moved toward offering multilingual support, there still exists a technology gap between systems in English and those in other languages which contributes to an information gap in various areas. This is mainly due to the fact that non-English contents bring many challenges that existing English-based information retrieval techniques do not address. For example, there are no explicit word boundaries in Chinese, making existing English indexing and searching techniques not directly applicable in Chinese information retrieval systems.

2.3.1. Comparison between general-purpose search engines

A report from Nielsen/NetRatings [36] in January 2003 rated Google (<http://www.google.com>), Yahoo (<http://www.yahoo.com>), and MSN (<http://www.msn.com>) as the most popular search engines in the United States [46]. In addition to providing easy access to more than 3 billion Web pages, Google has many special features such as cached links, site search, link search, Web page translation, stock quotes, and more. It supports more than 100 languages and its Chinese version of search engine has become one of the most popular search engine in China recently. Yahoo, the first human-compiled directory-based search engine, offers search results with directory-category links that have been reviewed by human experts. Microsoft's MSN Search provides a blend of human-powered directory information and crawler coverage different from any of the other top choices listed. It uses a Looksmart-powered directory, with secondary results from Inktomi. Other popular search engines in English include AltaVista (<http://www.altavista.com>), AOL (<http://search.aol.com/>), Ask Jeeves (<http://www.askjeeves.com>), HotBot (<http://www.hotbot.com>), etc. A number of these, such as Google, AltaVista, and Yahoo, have gradually expanded their services to non-English speakers.

Chinese is spoken by most people in mainland China, Hong Kong, and Taiwan, and most Chinese search engines have been developed to serve one of these regions. The most popular are Baidu (<http://www.baidu.com>) and Sina (<http://www.sina.com>) in mainland China, Yahoo Hong Kong (<http://hk.yahoo.com>) in Hong Kong, and Yam (<http://www.yam.com.tw>) and Openfind (<http://www.openfind.com.tw>) in Taiwan. Baidu has indexed 200 million Chinese Web pages and is the backend engine for over half of the major Chinese portals. In addition to basic Boolean search, it provides cached links and Chinese character encoding conversion. Sina, the most popular Web Portal in mainland China, offers directory-based searching with over 10,000 subcategories and recently adopted Google's search technology. However, advanced features such as encoding conversion were not provided in Sina. Yahoo Hong Kong, another directory-based search engine, returns results in both simplified and traditional Chinese. Openfind provides cached links and term suggestion functions, while Yam integrates encoding conversion to support cross-regional search. Compared with English search engines, Chinese search engines face more challenging issues, have fewer indexed pages, and provide fewer features. Contents focus on their own regions and contain much local information related to people's daily life.

2.3.2. Comparison between medical domain search engines

We further compared vertical search engines in the medical domain in two languages, because medical information is among the most popular resources on the Web. Supporting medical information searching on the Web also has attracted much attention from researchers and search engine builders, so a comparison of tools and search engines developed to support English and Chinese medical information seeking on the Web can help further illustrate the research gaps between them.

Numerous sites have been built to provide access to medical information over the Internet in English. For example, the National Library of Medicine's Gateway (<http://gateway.nlm.nih.gov/gw/Command>) and CliniWeb (<http://www.ohsu.edu/clinweb/>) provide access to Web pages from reputable organizations and institutions. In Gateway, Web pages are indexed according to the UMLS Metathesaurus, and in CliniWeb, the MeSH tree hierarchy. MDConsult (<http://www.medconsult.com>) and Medscape (<http://www.medscape.com>) aggregate journals, books, news, clinical symposia, and continuing medical education resources. MedTextus ([\[ai.bpa.arizona.edu/go/medical/MedTextus.html\]\(http://ai.bpa.arizona.edu/go/medical/MedTextus.html\)\) \[26\] and HelpfulMed \(<http://ai.bpa.arizona.edu/helpfulmed>\) \[10\] are two systems that search Web pages as well as medical databases and provide automatic thesaurus and document clustering functions to lower the requirement for human intervention.](http://</p>
</div>
<div data-bbox=)

For Chinese, the major Chinese medical portals include 999 (<http://www.999.com.cn>), MedCyber (<http://www.medcyber.com>), and WSJK (<http://www.wsjk.com.cn>) from mainland China, and TrustMed (<http://www.trustmed.com.tw>) from Taiwan. These portals have quite diverse content, ranging from general health to drugs, industry, research conferences, etc. However, surprisingly few of them incorporate a search function. Most medical portals serve as a medical content provider and their contents are manually updated. Only 999 provides a basic search function for Chinese medical information on the Internet. MedCyber and TrustMed provide a search function only within their own sites, while WSJK has no search ability. Most of these portals maintain a small collection of fewer than 10,000 pages and provide only the Chinese character version for their own region. Although there is a lot of medical information available on the Internet, few medical domain search engines have been developed. Most researchers need to rely on medical search engines in English which contain little Chinese information or they have to use general-purpose Chinese search engines.

Our comparison shows that in the medical domain, the gap between English and Chinese search engines is even greater than that of general-purpose search engines. English medical domain search engines have incorporated plenty of advanced features such as meta-search, medical thesaurus, and document clustering, while there exist few medical domain search engines in Chinese.

3. Research questions

Many Internet searching and browsing support techniques have been shown to be effective for English search engines, including meta-search, document categorization, and summarization. However, technologies for non-English languages are not as well developed. There is a desire to study how to generalize these techniques to non-English languages to support human information seeking.

Based on our review, we believe developing a generic framework with various searching and browsing support techniques promises to narrow the information gap between the English and non-English languages. In

this study, we aimed to address the following research questions:

1. How can we develop a generic approach to facilitating Internet searching and browsing by integrating selected search engine and post-retrieval analysis techniques for non-English content?
2. Would this integrated approach be more effective and efficient in facilitating information searching and browsing than other existing Internet systems for non-English content?
3. What is the users' level of satisfaction toward this integrated approach in comparison with existing systems for non-English content?

The remainder of the paper presents our work in studying these three questions.

4. A research testbed in the Chinese medical domain

Because of the importance of the Chinese language on the Web and the popularity of medical information, we selected the Chinese medical domain as our research testbed to investigate various issues in supporting Internet searching for non-English content.

As the largest non-English-speaking Internet population, the Chinese-speaking users in Mainland China, Hong Kong, and Taiwan make up 12.2% of the world online population (Global Internet Statistics. <http://www.gtreach.com/globstats/>). A recent report from China Internet Network Information Center [16] shows that Internet population in mainland China grew at a rate of 68% half-yearly, from 45.8 million in late 2002 to 68 million in early 2003, while Hong Kong and Taiwan are among the few regions in the world that have the highest Internet penetration rates. With the rapid growth of the Chinese online population, the need for information service is increasing dramatically.

Medical Web sites are among the most frequently visited Web sites on the Internet [42]. A tremendous number of Chinese medical information resources have been created on the Web, ranging from scientific papers and journals to general health topics and clinical symposia. These medical resources are of widely varied quality. However, Internet users are often frustrated when they try to look for Chinese health information online. The sheer volume of results returned by general Chinese search engines often overwhelms the users and there are few medical domain-specific search engines built for Chinese users. Compared to the wide availability of English medical information services such as

MEDLINE, CANCERLIT, and HelpfulMed [10], Chinese medical information services are under-developed to meet the growing medical information needs of the Chinese users.

Various factors contribute to the difficulties of supporting Chinese information-seeking in the medical area. One important problem is the regional differences between mainland China, Hong Kong, and Taiwan. Although the populations of all three regions speak Chinese, they use different Chinese characters. People from mainland China, where simplified Chinese is used, usually find it difficult to read traditional Chinese that is used in Hong Kong and Taiwan, while people from the latter two areas also have similar problems in reading simplified Chinese. Moreover, simplified Chinese and traditional Chinese are encoded differently in computer systems. Simplified Chinese is usually encoded using the GB2312 system and traditional Chinese is encoded using the Big5 system. When searching in a system encoded one way, users usually cannot get information encoded in the other.

Furthermore, Chinese medical search engines in mainland China, Hong Kong, and Taiwan usually keep only information from their own regions, while it is desirable for users to track medical information in all regions. For example, during the SARS epidemic, users who wanted to find detailed information about SARS outbreaks and control in all regions had to use different systems. These factors greatly hinder the medical information sharing among mainland China, Hong Kong, and Taiwan and result in information gaps between these regions.

5. Proposed approach

5.1. An integrated knowledge portal approach

We propose to use an “integrated knowledge portal” approach to supporting Internet searching and browsing in a multilingual world. Our portal approach adopts the common architecture of most search engines that collects Web documents, indexes them, and makes them searchable to information seekers. In addition, we add three key extensions to this basic structure: the generic language process ability, the integration of multiple information resources, and post-retrieval analysis ability.

5.1.1. Generic language process ability

Based on our review in Section 2.2.3, we adopted the character-based indexing technique in our research. In addition, the positional information on the words or

characters within a document was captured and stored such that when the query was a phrase, documents containing the exact phrase could be retrieved and given higher ranking than pages with separated words.

In addition to the basic character-based indexing, the ability to extract meaningful phrases from documents is also desired because such phrases are often useful for other analyses. For this purpose, we adopted a mutual information approach. The mutual information approach is a statistical method that identifies significant patterns as meaningful phrases from a large amount of text in any language [12,13,15,37]. The approach is an iterative process of identifying significant lexical patterns by examining the frequencies of word co-occurrences in a large amount of text. We experimented with this approach in processing Chinese, Spanish, and Arabic text collections and got satisfactory results.

5.1.2. Integration of multiple information resources and post-retrieval analysis

As reviewed in a previous section, relying on a single document collection could result in low information coverage. In the proposed approach, our own collections were complemented by information collated from different regions and resources using meta-searching. Such an integration of multiple information resources has been shown to offer precise and diverse information and facilitate efficient information seeking [9].

Post-retrieval analysis is another important feature in our design that provides added value to searching and browsing. We adopted text summarization technique to provide a preview of the search results and document categorization techniques to provide an overview of the search results. These techniques have been successfully applied in previous research [10,18].

5.2. A research prototype in the Chinese medical domain: CMedPort

Based on our proposed approach, we developed CMedPort as a research prototype to investigate whether integrated techniques can help improve Internet searching and browsing in languages other than English. It uses a three-tier architecture (as shown in Fig. 1). The main components are: (1) Content Creation; (2) Meta-search Engines; (3) Encoding Converter; (4) Chinese Summarizer; (5) Categorizer; and (6) User Interface. In this section, we discuss each major component in depth.

5.2.1. Content creation

As a cross-regional search engine that covers medical information from mainland China, Hong Kong,

and Taiwan, CMedPort needs to be able to collect information from all three regions. The AI Lab's SpidersRUs toolkit (<http://ai.bpa.arizona.edu/spidersrus/>), a digital library development tool developed by our research group, has been used to build collections for the Web portal. The toolkit contains components that support document fetching, document indexing, collection repository management, and document retrieval. It is able to deal with different encodings of Chinese (GB2312, Big5, and UTF8) and index different document formats, including HTML, SHTML, text, PDF, and MS Word. SpidersRUs also supports other languages, including English, Spanish, Arabic, etc.

5.2.1.1. Spidering. Based on suggestions from medical domain experts in the regions, 210 starting URLs were manually selected, including 87 from mainland China, 58 from Hong Kong, and 65 from Taiwan. They cover a large variety of medicine-related topics, from public clinics to professional journals, and from drug information to hospital information. Beginning with these medically related URLs, the SpidersRUs toolkit searched the Internet using a breadth-first search algorithm. It is assumed that medical pages included in the list will be likely to point to sites that they consider useful [10]. The three regional Web page collections created contain more than 300,000 Web pages in total. Web pages from mainland China, Hong Kong, and Taiwan were collected separately in order to differentiate among sources and identify encoding schemes. During the spidering process we found more medical Web sites in mainland China and Taiwan than in Hong Kong. This observation is reasonable because mainland China and Taiwan have much larger online populations than Hong Kong, and Hong Kong residents very often use English medical Web sites.

5.2.1.2. Indexing. In CMedPort we used character-based indexing with positional information for document retrieval. Character-based indexing is known to be efficient and achieve high recall. In our approach, the positional information about words or characters within a Web page was captured and stored such that, when the query was a phrase, Web pages containing the exact phrase could be retrieved and given higher ranking than pages with separated words. This also ensures the precision of Chinese document retrieval and could be useful in working with languages lacking explicit word boundaries.

To perform advanced information retrieval techniques, such as document categorization and summariza-

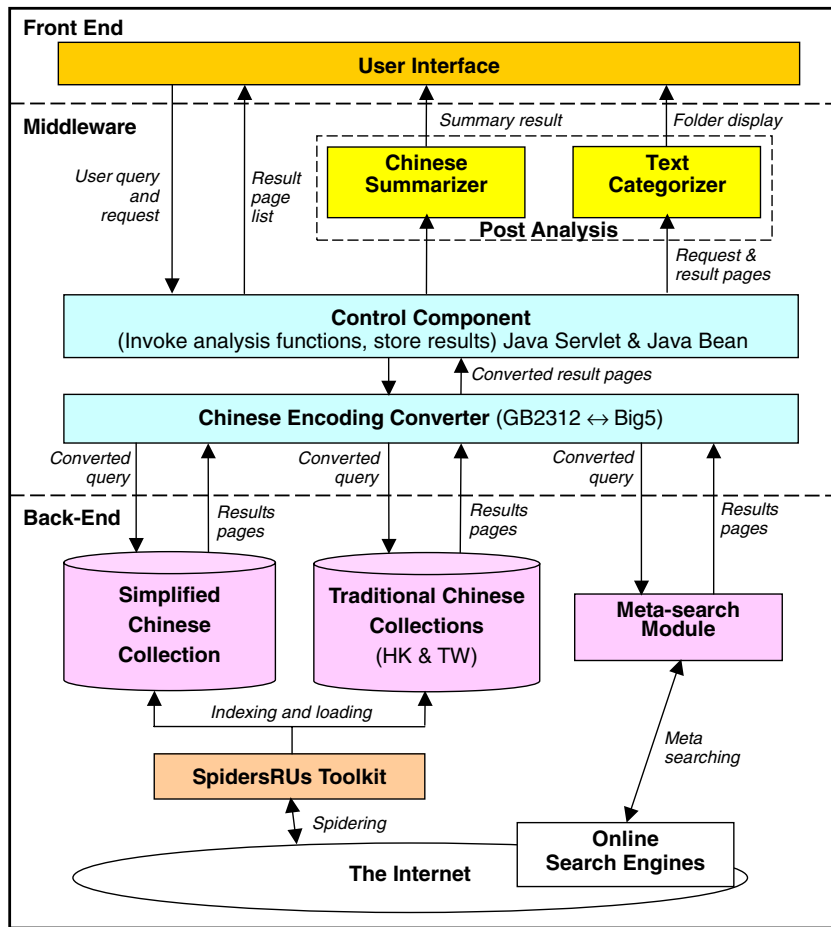


Fig. 1. The CMedPort system architecture.

tion, we extract meaningful Chinese phrases from textual information. In order to capture up-to-date phrases in our collection, we adopted the mutual information approach, a statistical method to identify significant lexical patterns examining the frequencies of word co-occurrences in a large amount of text [37]. This approach computes how frequently a pattern appears in the corpus, relative to its sub-patterns. Based on the algorithm, the MI of a pattern c (MI_c) can be found by

$$MI_c = \frac{f_c}{f_{\text{left}} + f_{\text{right}} - f_c}$$

where f stands for the frequency of a set of words. Intuitively, MI_c represents the probability of co-occurrence of pattern c , relative to its left sub-pattern and right sub-pattern. Phrases with high MI are likely to be extracted and used in automatic indexing. For example, if the Chinese phrase “乙肝病毒” (hepatitis B virus, HBV) appears in the corpus 100 times, the left sub-pattern (乙肝病) appears 110 times, and the right sub-

pattern (肝病毒) appears 105 times, then the mutual information (MI) for the pattern “乙肝病毒” is $100 / (110 + 105 - 100) = 0.87$. Phrases of all lengths were examined in the MI program. Furthermore, Stop words like “的” (of), “了” (function word, no meaning), and “及” (and) are removed. The included word list, which has priority over the stop-word list, allows users to have the flexibility to retain words that appear in the stop-word list. For example, the Chinese phrase “目的” (aim) can be listed in the included words although the word “的” (of) appears in the stop-word list. Using this approach we created simplified and traditional Chinese lexicons. Indexing against MI lexicon were then saved into a separate index file and later used by the CMedPort Categorizer.

The indexed files were loaded into a MS SQL Server database in which the data were separated into the three regions such that, during retrieval, the system could tell which region a Web page came from. Pages from each region were ranked by $tf*idf$ during retrieval. $Tf*idf$

combines the frequency of occurrence of every word in a document as well as the word's total occurrences in the collection, which indicated correlation between documents and a particular keyword.

5.2.2. Meta-search engines

In addition to the regional collections, CMedPort also integrates information from different sources by meta-search engines. As discussed, few reputable medical-domain Chinese search engines are available for incorporation as information sources. Although some medical databases have been developed, online versions such as China Academy CBM (Chinese Biomedical Database) (<http://www.imicams.ac.cn/cbmdisc/cbmdisc.html>) are not stable. Based on suggestions from domain experts, six key Chinese search engines were chosen for meta-searching, two from each region (as shown in Table 1). These search engines contain a large portion of medically related information, usually from different parts of the Internet. Access to these systems provides a richer representation and a fresh coverage of information to supplement our local collection. More meta-search engines could be incorporated into CMedPort as they become available on the Internet.

5.2.3. Encoding converter

In order to share medical information in different forms of written Chinese (simplified Chinese and traditional Chinese) among all three regions, an encoding conversion program is employed in CMedPort. The encoding converter uses a dictionary of 6737 entries that map between simplified and traditional Chinese characters. Since some simplified characters map to multiple traditional equivalents, the conversion from simplified characters to traditional ones is sometimes ambiguous.

We have picked the candidate character that is most frequently selected as equivalent to the original one.

In the simplified Chinese version of CMedPort, when a user enters a query in simplified Chinese the query is sent to all mainland China information sources using simplified Chinese. At the same time, the query is converted into traditional Chinese and sent to all information sources from Hong Kong and Taiwan that use traditional Chinese. When displaying results, the encoding conversion program is invoked again to convert results from traditional Chinese into simplified Chinese. The whole process is transparent to the user. The encoding conversion program enables cross-regional search and addresses the problem of dissimilar Chinese character forms.

5.2.4. Chinese summarizer

Automatic summarization has been applied as a document preview tool in many English retrieval systems. In CMedPort, a Chinese Summarizer was developed based on the sentence extraction approach. The Chinese Summarizer is a modified version of the AI Lab TXTRACTOR, a summarizer for English documents developed in our previous research [31]. Its major components include: 1) sentence evaluation, 2) segmentation or topic boundary identification, and 3) segment ranking and extraction. First, a sentence evaluation component parses the original Web page and extracts all sentences. These sentences are evaluated based on linguistic heuristics including presence of cue phrases (e.g. “总而言之” (in summary), “所以” (therefore)), $tf*idf$ score normalized for the sentence length, sentence position, and sentence length. Second, the TextTiling algorithm [19] is used to analyze the Web page and determine where the topic boundaries are

Table 1
Major Chinese search engines in the three regions

Region	Information source	Description
Mainland China	Baidu (www.baidu.com)	The biggest Internet search service provider in mainland China and has indexed millions of medical Web pages.
	Sina China (www.sina.com.cn)	The biggest Web portal in mainland China, containing more than 100,000 manually classified medical and health related Web sites.
Hong Kong	Yahoo! Hong Kong (hk.yahoo.com)	The most popular directory-based search engine in Hong Kong. Its “Health and Medicine” directory contains both public health and professional medical information.
	Hong Kong Government Information Center (search2.info.gov.hk)	A high quality search engine provided by the Hong Kong government, providing information including Hong Kong Health Department news, policies, etc.
Taiwan	Yam (www.yam.com)	The biggest Chinese search engine in Taiwan with a Health directory of more than 100,000 manually classified Web sites.
	Taiwan (www.sina.com.tw)	One of the biggest Chinese Web portals in Taiwan containing information about hospitals, traditional Chinese medicine, etc.

located. The Web page is thus segmented into its main topics. Third, the Summarizer ranks the document segments based on the scores given to the sentences and extracts high-ranking sentences from different segments as summary sentences.

The Chinese Summarizer was embedded in the CMedPort system and can be invoked at real time. Users can start the summarizer by choosing the number of sentences to be summarized under each returned result. The Chinese Summarizer dynamically retrieves the Web page on the Internet, processes the content, and presents a summary in a pop-up summarizer window. On the summarizer page, summary sentences are displayed on the left-hand side and the original Web page is displayed on the right-hand side with summary sentences highlighted. Users can click on any summary sentence on the left-hand side and go to the location of that sentence in the original page on the right-hand side. This feature is especially useful for browsing long documents where the Summarizer can help a user quickly determine whether or not a Web page is of interest.

5.2.5. Categorizer

Another component of CMedPort is the categorizer, which organizes returned results into various folders labeled by key phrases. When the categorizer is invoked all returned results are processed, and key phrases that have appeared in their titles and summaries are extracted by matching to the Chinese lexicons obtained from the Mutual Information program. As described in Section 5.2.1, the lexicons were constructed from fresh Web page collections and are more up-to-date and highly related to the medical domain than previous Chinese lexicons. Key phrases with high occurrences in returned results are extracted as folder topics. Web pages that contain a folder topic are included in that folder. One Web page may appear in multiple folders if it contains multiple folder topics. We are using only title and summary to extract keywords because it is practical and permits dynamic categorization. Previous research has shown that clustering based on snippets is almost as effective as clustering based on a whole document [45].

5.2.6. User interface

CMedPort has two versions of User Interface to accommodate users from different regions: the traditional Chinese version and the simplified Chinese version. They look the same and provide the same functionalities, except that they use different encoding schemes and Chinese characters (simplified vs. traditional).

On the search page (see Fig. 2.a), users can begin searching by typing keywords in the search box and

indicating which local database and meta-search engines are to be searched. Multiple keywords can be entered into the search box at the same time, one keyword per line. Available information sources are organized into three columns by the region to which they belong and can be chosen by selecting the checkbox in front of a name.

On the result page, the top 20 results from each information source are displayed as ranked lists. Results from the CMedPort local collection are displayed in sorted order of relevancy to the query as measured by $tf*idf$ score. For each result in the lists, the title and a short summary are displayed (see Fig. 2b). All results of different encodings are converted into the same encoding as the interface and displayed together (see Fig. 2c). By clicking on the name of a particular information source in the navigation bar at the top right-hand side of the page, users can go to the first result from that information source.

There is a drop-down box beneath each result in the list that users can use to select a sentence length and let the system automatically generate a 1-to-5-sentence summary of a Web page (see Fig. 2d). Users can also click on the ‘Analyze Results’ button to go to the analyzer page where all the results are categorized into folders with extracted topics. Clicking on the folders of interest produces a list of URL titles that is displayed under the relevant folder for him/her to browse (see Fig. 2e).

6. Evaluation methodology

In order to evaluate our system’s performance in assisting Internet searching and browsing, a user experiment was designed and conducted. Our study mainly addressed the following questions: (1) whether our integrated approach in CMedPort can facilitate searching and browsing of Chinese medical information more effectively and more efficiently than other existing Chinese search engines; (2) whether the summarizer and categorizer in CMedPort are effective and efficient support tools for browsing; and (3) whether CMedPort generates higher user satisfaction than other Chinese search engines. In this section, we discuss the experimental design and the results of our study.

6.1. Search and browse tasks

Consistent with the Interactive track in TREC (Text Retrieval Conference) evaluations [44], we gave users a list of questions. They are required to find answers using the given systems with their own queries. During the experiment, they were allowed to change their

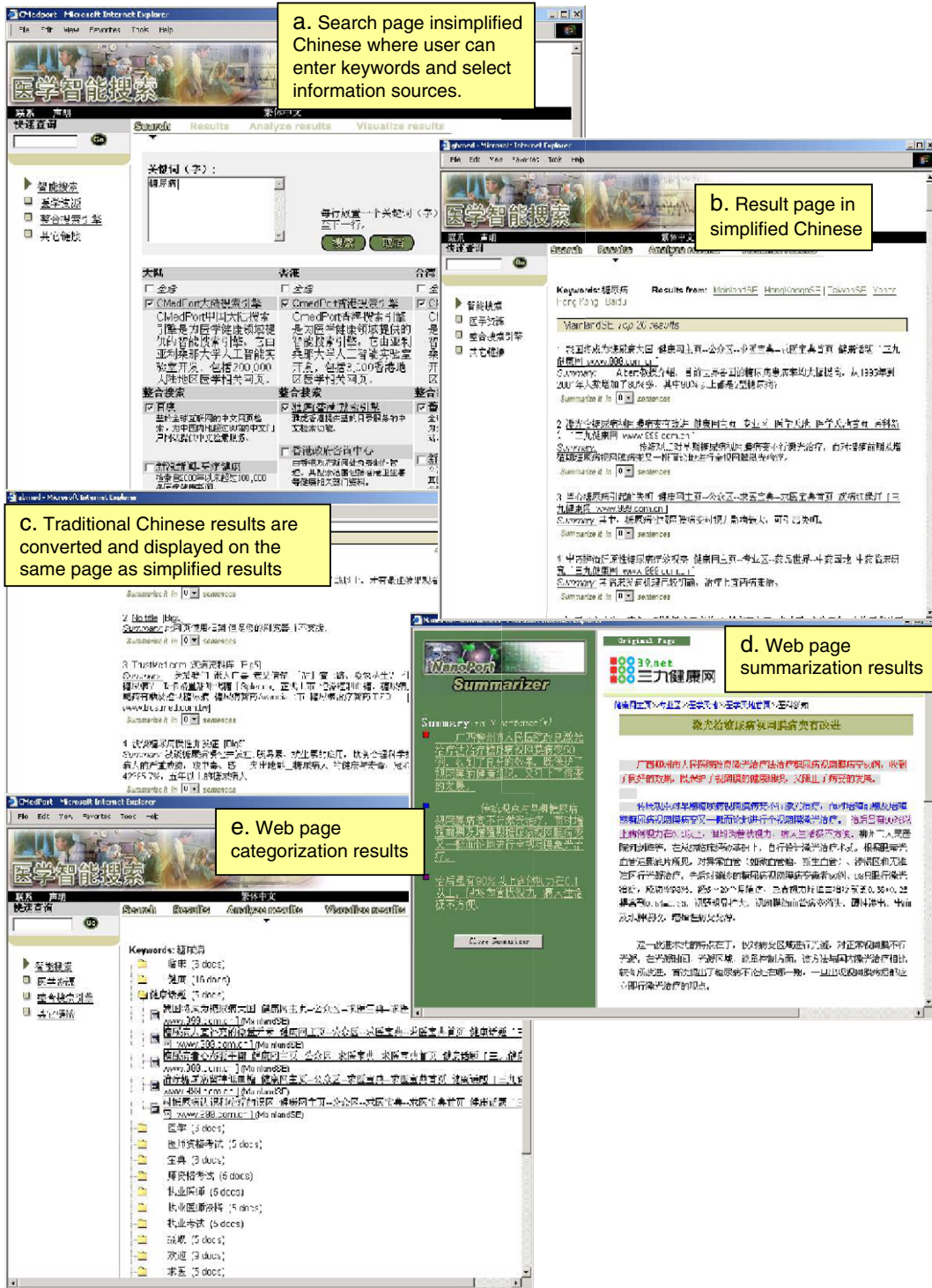


Fig. 2. The CMedPort user.

query words and do multiple retrievals. Since CMedPort has been designed to facilitate both searching and browsing, two types of tasks were designed: search tasks and browse tasks.

Search tasks in our user study were short questions that required specific answers. This type of task was designed to have just one correct answer. An example of a searching task would be “Which disease was

Cocktail Therapy FIRST used on?” Since there is always just one correct answer to the question, accuracy was used as the primary measure of effectiveness in searching tasks as follows:

Accuracy

$$= \frac{\text{Number of correct answers given by the subject}}{\text{Total number of questions asked}}$$

In browse tasks, subjects were given a topic that defined an information need accompanied by a short description regarding the task. Subjects were expected to summarize the findings of their Web browsing session as a number of themes. A theme was defined as “a short phrase, which describes a certain topic” (Chen et al., 2001). Different from search tasks, there are always multiple themes that a user needs to address in browse tasks. An example of a browse task would be:

Topic: Liu Wei Di Huang Wan (A traditional Chinese medicine)

Description: Please summarize the effect of Liu Wei Di Huang Wan and diseases that it can be used on.

Theme identification has been used to evaluate performance of browse tasks [9,11]. Theme precision and theme recall were used as the primary measures of effectiveness in browsing tasks. They are defined as follows:

Theme precision

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

Theme recall

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

A theme is considered correct if it matches any of the themes identified by experts. By examining the themes that subjects came up with using different searching tools, we were able to evaluate how effectively each system helped a user locate relevant information from the Web.

Efficiency in both tasks was directly measured by the time that subjects spent on the tasks using different systems.

6.2. Benchmarks

To compare the performance of CMedPort with existing Chinese Search Engines, we selected one search engine from each region as a benchmark system. Existing Chinese medical portals were not considered suitable for benchmarks because they do not have adequate

search functionalities and they usually search only inside their own Web sites. Thus, CMedPort was compared with three major commercial Chinese search engines from the three regions: Sina, Yahoo! Hong Kong, and Openfind. Although their content is not focused on the medical domain, their indexes of Chinese Web pages cover a large variety of medical information. Among the existing Chinese search engines they are the ones having the most functions comparable with those of CMedPort and are the most popular Chinese search engines in their respective regions.

6.3. Hypotheses

We tested three groups of hypotheses in our Chinese medical-domain search engine, CMedPort.

In Hypotheses 1, we hypothesized that CMedPort would be more effective and efficient than benchmark search engines in search tasks where specific information is needed.

H1a. CMedPort is more effective than the existing benchmark Chinese search engines in searching.

H1b. CMedPort is more efficient than the existing benchmark Chinese search engines in searching.

In Hypothesis 2, we hypothesized that CMedPort would be more effective and efficient than benchmark search engines in browse tasks where users need to become familiar with a topic. CMedPort integrates information sources from three regions and provides broad coverage of results. At the same time, we believed Chinese summarization and categorization would be post-retrieval analysis tools that would further improve effectiveness and efficiency in CMedPort.

H2.1a. CMedPort is more effective than the existing benchmark Chinese search engines in browsing.

H2.1b. CMedPort is more efficient than the existing benchmark Chinese search engines in browsing.

H2.2a. CMedPort’s Chinese summarization further improves effectiveness in browsing.

H2.2b. CMedPort’s Chinese summarization further improves efficiency in browsing.

H2.3a. CMedPort’s Chinese categorization further improves effectiveness in browsing.

H2.3b. CMedPort’s Chinese categorization further improves efficiency in browsing.

In Hypothesis 3, we hypothesized that CMedPort would achieve higher user satisfaction in terms of usability. CMedPort provides a user-friendly interface with clearly organized information sources. The ability to do cross-regional search, summarization, and categorization would ease the process of information seeking.

H3. CMedPort achieves higher user satisfaction than existing benchmark Chinese search engines.

6.4. Experimental design

Forty-five subjects, 15 from each region, were recruited for the experiment. Each subject was required to perform four search tasks and eight browse tasks. We specified more browse tasks because we also were interested in the performance of post-retrieval analysis tools in browsing. A time limit of 10 min was given for each task. Among the search tasks, two were performed using CMedPort and another two using the benchmark search engine from the subject's own region. Among the browse tasks, two were performed using CMedPort with general search function without summarizer or categorizer, two used CMedPort with summarizer, two used CMedPort with categorizer, and two used the benchmark search engine. The order of the questions used in the tasks and the systems used in the experiment were rotated to avoid any potential bias. During the experiment, the tasks performed by subjects were timed and their answers were judged by domain experts. In order to test the CMedPort system as a whole, we did not specify the queries. The users chose their own query words. They could modify their query words anytime during the experiment.

At the end of each experiment, subjects were required to fill out a user satisfaction questionnaire. The questionnaire included 1) 19 Computer System Usability Questionnaire (CSUQ) items from [27], specifically designed to assess aspects of usability, (Table 2); 2) questionnaire on individual components of CMedPort to evaluate the user perspective on cross-regional search, summarizer, and categorizer; and 3) open-ended positive or negative comments on CMedPort.

Three Chinese graduate students from the medical school at the University of Arizona, one from each region, were recruited as domain experts. They helped design the tasks used in our experiment and provided the standard answers for search and browse tasks. We did not control the type of websites experts got the answers from. They can use any search engines that they are comfortable with to provide the correct answers. The

Table 2
Computer system usability questionnaire ([27])

Overall

Overall, I am satisfied with this system

Easiness of use

Overall, I am satisfied with how easy it is to use this system

It was simple to use this system

I feel comfortable using this system

It was easy to learn to use this system

The information provided for the system is easy to understand

The information (such as online help, on-screen messages, and other documentation) provided with this system is clear

It is easy to find the information I needed

Effectiveness and efficiency

I can effectively complete my work using this system

I am able to complete my work quickly using this system

I am able to efficiently complete my work using this system

I believe I became productive quickly using this system

The information is effective in helping me complete the tasks and scenarios

Error recovery

The system gives error messages that clearly tell me how to fix problems

Whenever I make a mistake using the system, I recover easily and quickly

Interface

The organization of information on the system screens is clear

The interface of this system is pleasant

I like using the interface of this system

Functionality

This system has all the functions and capabilities I expect it to have

final version of standard answers was an agreed-upon version from the experts.

7. Experimental results

In this section, we report the evaluation results and observations based on the user study.

7.1. Results from search tasks

Table 3.1 summarizes the systems' search-task performance by regions. In order to determine whether there were significant differences among the performances of the systems, paired *t*-tests were performed for each pair of methods. The statistical results (*p*-values) are shown in Table 3.2. Table 3.3 shows average improvement achieved in CMedPort compared to benchmark systems.

The results of testing hypotheses H1a and H1b showed that in terms of effectiveness CMedPort per-

Table 3.1

Searching performance of CMedPort and benchmark systems by regions

Region	System used	Accuracy	Time spent (s)
Mainland China	Sina China	0.625	149.039
	<i>CmedPort</i>	0.917	97.962
Hong Kong	Yahoo HK	0.857	117.967
	<i>CMedPort</i>	0.929	95.033
Taiwan	Openfind	0.846	114.767
	<i>CMedPort</i>	0.962	72.433

formed significantly (at $\alpha=0.05$) better than Sina and comparably to Yahoo HK and Openfind in search tasks. In terms of efficiency, CMedPort was significantly (at $\alpha=0.05$) better than Sina and Openfind and comparable to Yahoo HK. On average CMedPort achieved 20% higher accuracy and was 30% more efficient than the benchmark systems.

7.2. Results from browse tasks

Table 4.1 summarizes the system performance in browse tasks. Hypotheses test results are shown in Tables 4.2 and 4.3. Table 4.4 shows average improvement achieved in CMedPort compared with the benchmark systems and illustrates improvement gained using summarizer and categorizer.

The results of testing hypotheses H2.1a and H2.1b (Table 4.2) showed that CMedPort achieved significantly (at $\alpha=0.05$) higher theme precision than Yahoo HK and Openfind, and significantly (at $\alpha=0.05$) higher theme recall than all three benchmark systems for browse tasks. Meanwhile, users spent significantly (at $\alpha=0.05$) less time in CMedPort than in Sina and Openfind. On average we found 22.2% improvement in theme precision, 113% improvement in theme recall, and 19.5% improvement in efficiency comparisons between CMedPort and benchmark Chinese search engines. The results from browse tasks are very encouraging. The high theme precision probably resulted from our focused collection building technique, including our character-based indexing approach for Chinese.

Table 3.2

Hypotheses testing for search tasks by regions

H1: search task	Hypotheses	<i>p</i> -value	Result
H1a: effectiveness	CMed>Sina	0.008*	Confirmed
	CMed>Yahoo	0.163	Not confirmed
	CMed>Openfind	0.092	Not confirmed
H1b: efficiency	CMed>Sina	0.040*	Confirmed
	CMed>Yahoo	0.194	Not confirmed
	CMed>Openfind	0.045*	Confirmed

* Significant level is at 0.05.

Table 3.3

Searching performance of CMedPort and benchmark systems with combined regions

	Accuracy	Improvement	Time spent	Improvement
Benchmark systems	0.776	–	127.3	–
CMedPort	0.935	20.5%	88.5	30.5%

The theme recall achieved is greater than we expected. This probably is because CMedPort has the ability to search information in all three regions and integrates results from different search engines by meta-searching.

The results of testing hypotheses H2.2a and H2.2b (Table 4.3) showed that there were no significant (at $\alpha=0.05$) differences in theme precision and recall when using or not using CMedPort's summarizer. No significant (at $\alpha=0.05$) differences in efficiency were found. Surprisingly, the Chinese summarizer did not further improve performance. We observed that the Chinese summarizer is not always fast. Before processing the page, the summarizer needs to fetch Web pages from remote servers and processing time is largely affected by the length of the Web page. Users usually lost patience after 10 s of waiting and would try other results. These factors affected the performance of the Chinese summarizer as a browsing support tool.

The results of testing hypotheses H2.3a and H2.3b (Table 4.3) showed that there were no significant (at $\alpha=0.05$) differences in theme precision and recall when using or not using CMedPort's categorizer. However, using CMedPort's categorizer was significantly (at $\alpha=0.05$) more efficient than not using the categorizer. Results suggested that, as a document overview tool, the categorizer could help users identify topics of interest more quickly when browsing.

Table 4.1

Browsing performance of CMedPort and benchmark systems by regions

Region	System used	Theme precision	Theme recall	Time spent (s)
Mainland China	Sina China	0.675	0.250	412.231
	<i>CMedPort Basic</i>	0.819	0.472	312.961
	<i>CMedPort Summarizer</i>	0.849	0.385	270.231
	<i>CMedPort Categorizer</i>	0.859	0.510	265.039
Hong Kong	Yahoo HK	0.651	0.228	376.700
	<i>CMedPort Basic</i>	0.790	0.524	360.400
	<i>CMedPort Summarizer</i>	0.878	0.517	454.800
	<i>CMedPort Categorizer</i>	0.825	0.506	286.133
Taiwan	Openfind	0.636	0.215	318.267
	<i>CMedPort Basic</i>	0.789	0.480	218.100
	<i>CMedPort Summarizer</i>	0.739	0.450	255.933
	<i>CMedPort Categorizer</i>	0.845	0.514	230.733

Table 4.2
Hypotheses testing for browse tasks of CMedPort and benchmark systems

H2: browse tasks		Hypotheses	<i>p</i> -value	Result
H2.1a: effectiveness	Theme precision	CMed>Sina	0.071	Not confirmed
		CMed>Yahoo	0.050*	Confirmed
		CMed>Openfind	0.031*	Confirmed
	Theme recall	CMed>Sina	<0.001*	Confirmed
		CMed>Yahoo	<0.001*	Confirmed
		CMed>Openfind	<0.001*	Confirmed
H2.1b: efficiency		CMed>Sina	0.003*	Confirmed
		CMed>Yahoo	0.290	Not confirmed
		CMed>Openfind	<0.001*	Confirmed

* Significant level is at 0.05.

7.3. Results from usability questionnaire

The results of testing hypotheses H3 show that CMedPort was rated significantly higher than all benchmark systems. CMedPort achieved an average rating of 5.83 out of 7, while all three benchmark Chinese search engines were rated below 5. Among all the evaluation criteria, CMedPort consistently gained higher ratings than benchmark Chinese search engines. Users appreciated the easiness, effectiveness, and interface of CMedPort. Among the three benchmark systems, Yahoo HK and Openfind were rated higher than Sina. This is consistent with what we observed in the performance of search and browse tasks.

Questionnaires on easiness and usefulness of featured tools in CMedPort showed that the categorizer achieved 6.174 out of 7, the highest among the three features, while the cross-regional search feature was rated 5.949 and the summarizer was rated 5.665. No paired *t*-tests were performed on individual tools in CMedPort because benchmark systems do not provide functions comparable to those in CMedPort (Tables 5.1 and 5.2).

7.4. Subjective feedback

The results of the searching and browsing experiments and usability questionnaire show that CMedPort achieved significant improvement on three regional

benchmark systems. Users' subjective feedback provides further information about the differences in the system performance. We summarize several aspects that received the most comments.

7.4.1. Information quality and coverage

Twenty-one out of 45 subjects indicated that CMedPort gave more relevant results compared to benchmark systems. They expressed that "it is easier to find useful information using CMedPort," while "Sina provides a lot of commercial company information as top results." One Taiwanese student said, "CMedPort is especially useful when looking for information from other regions (mainland China and Hong Kong)." Two users pointed out that the benchmark systems gave few results when they searched for "抗肿瘤药" (anti-tumor drug), while CMedPort gave plenty of useful results.

7.4.2. Categorizer

Among the 45 subjects, 32 subjects expressed that they liked the Categorization function. One subject said, "Categorizer is very powerful and useful. I like this function very much." Others said, "It (categorizer) makes searching more quickly. I can skip a lot of irrelevant information in results and focus on relevant ones." Four subjects mentioned that the categorizer topics were not very relevant to what they were looking for sometimes. We observed that category topics sometimes contain noise from advertisements.

Table 4.3
Hypotheses testing for browse tasks of CMedPort summarizer and categorizer

H2: browse tasks		Hypotheses	<i>p</i> -value	Result
H2.2a: effectiveness	Theme precision	CMed Summ>CMed basic	0.214	Not confirmed
	Theme recall	CMed Summ>CMed Basic	-0.201	Not confirmed
H2.2b: efficiency		CMed Summ>CMed basic	-0.290	Not confirmed
H2.3a: Effectiveness	Theme precision	CMed Categ>CMed Basic	0.073	Not confirmed
	Theme recall	CMed Categ>CMed basic	0.307	Not confirmed
H2.3b: efficiency		CMed Categ>CMed basic	0.022*	Confirmed

* Significant level is at 0.05.

Table 4.4
Browsing performance of CMedPort and benchmark systems with combined regions

	Theme precision	Improvement	Theme recall	Improvement	Time spent (s)	Improvement
Benchmark systems	0.654	–	0.231	–	369.066	–
CMedPort	0.800	22.2%	0.492	113.0%	297.154	19.5%
CMedPort with summarizer	0.822	25.6%	0.451	95.2%	326.988	11.4%
CMedPort with folder display	0.843	28.8%	0.510	120.8%	260.635	29.4%

7.4.3. Summarizer

Twenty-five subjects expressed their preference for CMedPort's Chinese summarizer. For example, one subject claimed that "I also like summary function. It becomes more useful when the article is very long." Four of them mentioned that they liked the feature that summary sentences are highlighted in the original Web page. Eight subjects gave neutral comments on the summarizer. One said, "summarizer is useful sometimes, but not all the time." Nine subjects complained about the processing speed of the summarizer, although some of them liked the idea of summarizing Web pages.

7.4.4. Cross-regional search

Eleven subjects commented that cross-regional search is useful. They complained that "it is difficult to find information from other regions in Openfind." Similar comments were made on the other benchmark regional search engines. One Hong Kong user said: "It is helpful to convert simplified Chinese into traditional Chinese."

7.4.5. User-interface

Seventeen users mentioned that they liked the clear user-interface of CMedPort. They commented that "CMedPort looks professional and is easy to understand," while benchmark search engines "always have advertisements floating around and make me lose focus."

7.4.6. Speed

Nine subjects commented that benchmark search engines had faster processing speed when compared to

CMedPort, which saved them information seeking time. However, as CMedPort is still a research prototype and does not have a powerful server to support the process, it would not be difficult to improve the speed of the system.

In general, subjects' overall opinion on CMedPort tended to be positive, although some complained about its speed. In contrast, benchmark systems received relatively fewer positive comments.

8. Conclusions and future directions

In this paper, we have proposed a general framework for supporting multilingual Internet searching and browsing on the Web. The framework is consistent with the architecture of most search engines. However, our approach features three extensions to this basic structure: generic language processing ability, integration of multiple information resources, and post-retrieval analysis.

We have discussed the development of CMedPort, a Chinese medical portal to serve the information seeking needs of Chinese users. A systematic evaluation has been conducted to study the effectiveness and efficiency of CMedPort in assisting human analysis. Our experimental results show that CMedPort achieved significant improvement in searching and browsing performance compared to three benchmark regional search engines, Sina, Yahoo! Hong Kong, and Openfind. We believe that CMedPort's collection building method, meta-searching, and cross-regional searching contributed to the improvement in information seeking. Although post-retrieval analysis methods, such as categorizer and summarizer, did not further improve browsing performance significantly, users' subjective evaluation and verbal comments revealed that they appreciated these analysis functions. Overall, the experimental results are promising.

Table 5.1
User satisfaction rating of CMedPort and benchmark systems

	Sina	Yahoo HK	Openfind	CMedPort (Avg)
Overall	4.040	5.069	4.857	6.033
Easiness	4.813	5.322	5.402	6.081
Effectiveness	4.078	4.372	4.804	5.890
Error recovery	4.440	4.552	4.750	5.363
Interface	5.026	5.477	5.095	5.988
Functionality	4.083	4.643	4.571	5.626
Average	4.413	4.906	4.913	5.830
Cross-regional search	–	–	–	5.949
Summarizer	–	–	–	5.665
Categorizer	–	–	–	6.174

Table 5.2
Hypotheses testing for user satisfaction

H3: user satisfaction	<i>p</i> -value	Result
CMedPort>Sina	<0.001*	Confirmed
CMedPort>Yahoo HK	0.003*	Confirmed
CMedPort>Openfind	0.001*	Confirmed

* Significant level is at 0.05.

However, this research has several limitations. Since our study is based on subjects from different Chinese-speaking regions, a large sample size is infeasible. We therefore reported results based on a limited number of 45 subjects. Additionally, because there are few medical domain Chinese search engines available, we were only able to compare with general Chinese search engines that contain medical content as our benchmark systems. Since these Chinese search engines do not have post-retrieval analysis components in them, we were only able to compare basic functions in CMedPort with existing search engines.

In the future we plan to study the semantic differences between Simplified and Traditional Chinese to provide cross-regional information search functionalities for Chinese users. We are also in the process of applying our integrated approach to search engines in more languages such as Spanish and Arabic. Meanwhile, we plan to add multilingual information retrieval function to these portals, and a dynamic, graphic knowledge map that could categorize multilingual documents retrieved from the Web is in development. We also plan to conduct a larger-scale evaluation study with Chinese medical practitioners that could address the limitations of our current user study.

Acknowledgement

This project was supported in part by an NSF Digital Library Initiative-2 grant, PI: H. Chen, “High-performance Digital Library Systems: From Information Retrieval to Knowledge Management,” IIS-9817473, April 1999-March 2002. We would like to thank the following people for their help and comments: Zan Huang, Yiwen Zhang, Wingyan Chung, Gang Wang, Michael Chau, Daniel McDonald, Byron Marshall, Alan Yip, Mark Chen, Gondy Leroy, Thian-Huat Ong, Wai-Ki Sung, Chienting Lin, and Lu Tseng. We would also like to thank the AI Lab team members who developed the AI Lab SpidersRUs toolkit, the Mutual Information software, and the Web Weaver package. Finally, we also want to thank the domain experts and all our subjects who took part in the evaluation study.

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] N.J. Belkin, P.G. Marchetti, C. Cool, BRAQUE: design of an interface to support user interaction in information retrieval, *Information Processing & Management* 29 (3) (1993) 325–344.
- [3] D.C. Blair, M.E. Maron, An evaluation of retrieval effectiveness for a full-text document-retrieval system, *Communications of the ACM* 28 (3) (1985) 289–299.
- [4] E. Carmel, S. Crawford, H. Chen, Browsing in hypertext: a cognitive study, *IEEE Transactions on Systems, Man, and Cybernetics* 22 (5) (1992) 865–884.
- [5] S. Chakrabarti, M. van den Berg, B. Dom, Focused crawling: a new approach to topic-specific web resource discovery, *Proceedings of the 8th International World Wide Web Conference*, Toronto, Canada, 1999.
- [6] M. Chau, H. Chen, Comparison of three vertical search spiders, *IEEE Computer* 36 (5) (2003) 56–62.
- [7] M. Chau, H. Chen, J. Qin J., Y. Zhou, Y. Qin, W. Sung, D. McDonald, Comparison of two approaches to building a vertical search tool: a case study in the nanotechnology domain, *Proceedings of JCDL’02*, ACM Press, Portland, Oregon, USA, 2002, pp. 135–144.
- [8] H. Chen, A.L. Houston, R.R. Sewell, B.R. Schatz, Internet browsing and searching: user evaluations of category map and concept space techniques, *Journal of the American Society for Information Science* 49 (7) (1998) 582–603.
- [9] 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.
- [10] H. Chen, A. 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.
- [11] H. Chen, H. Fan, M. Chau, D. Zeng, Testing a cancer meta spider, *International Journal of Human-computer Studies* 59 (5) (2003) 755–776.
- [12] L. Chien, PAT-tree-based keyword extraction for Chinese information retrieval, *Proceedings of the 1997, ACM SIGIR*, Philadelphia, PA, USA, 1997, pp. 50–58.
- [13] L. Chien, H. Pu, Important issues on Chinese information retrieval, *Computational Linguistics and Chinese Language Processing* 1 (1) (1996) 205–221.
- [14] J. Cho, H. Garcia-Molina, L. Page, Efficient crawling through URL ordering, *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 1998.
- [15] K. Church, P. Hanks, Word association norms, mutual information, and lexicography, *Proceedings of the 27th Annual Meeting of Association for Computational Linguistics*, 1989, pp. 76–83 Vancouver, BC, Canada.
- [16] CNNIC, Statistical reports on the Internet development in China, The 12th Survey Report (June, 2003). [Online]. Available at <http://www.cnnic.net.cn/download/manual/en-reports/12.pdf>.
- [17] D. Ellis, A behavioral approach to information retrieval systems design, *Journal of Documentation* 45 (3) (1989).
- [18] S. Greene, G. Marchionini, C. Plaisant, B. Shneiderman, Previews and overviews in digital libraries: designing surrogates to support visual information seeking, *Journal of the American Society for Information Science* 51 (4) (2000) 380–393.
- [19] M.A. Hearst, Multi-paragraph segmentation of expository text, *Proceedings of the 32th annual meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, 1994, pp. 9–16.
- [20] M.A. Hearst, J.O. Pedersen, Reexamining the cluster hypothesis: scatter/gather on retrieval results, *Proceedings of the 19th International ACM SIGIR Conference on Research and Development*

- in Information Retrieval (SIGIR '96), ACM Press, New York, 1996, pp. 76–84.
- [21] E. Hovy, C.Y. Lin, Automated text summarization in SUMMARIST, *Advances in Automatic Text Summarization*, MIT Press, 1999, pp. 81–94.
- [22] C.C. Kuhlthau, Inside the search process: information seeking from the user's perspective, *Journal of the American Society for Information Science* 42 (5) (1991) 361–371.
- [23] K. Kwok, Comparing representations in Chinese information retrieval, *Proceedings of ACM SIGIR*, Philadelphia, PA, 1997, pp. 34–41.
- [24] S. Lawrence, C.L. Giles, Accessibility of information on the web, *Nature* 400 (1999) 107–109.
- [25] G. Leroy, H. Chen, MedTextus: an ontology-enhanced medical portal, *Proceedings of the Workshop on Information Technology and Systems (WITS)*, Barcelona, Spain, 2002.
- [26] J.R. Lewis, IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use, *International Journal of Human-Computer Interaction* 7 (1) (1995) 57–78.
- [27] G. Marchionini, Information seeking in electronic environments, in: J. Long (Ed.), *Cambridge Series on Human-Computer Interaction*, vol. 9, Cambridge University Press, Cambridge, 1995, 10 vols.
- [28] G. Marchionini, B. Shneiderman, Finding facts vs. browsing knowledge in hypertext systems, *IEEE Computer* 21 (1) (1988) 70–80.
- [29] A. McCallum, K. Nigam, J. Rennie, K. Seymore, Building domain-specific search engines with machine learning techniques, *Proceedings of AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.
- [30] D. McDonald, H. Chen, Using sentence selection heuristics to rank text segments in TXTRACTOR, *Proceedings of JCDL'02*, Portland, Oregon, 2002, pp. 28–35.
- [31] P. McLellan, A. Tombros, J. Jose, I. Ounis, M. Whitehead, Evaluating summarisation technologies: a task oriented approach, *Proceedings of the 3rd European Conference on Digital Libraries*, 2001, pp. 198–214, Paris, France.
- [32] L.I. Meho, R. Tibbo, Modeling the information-seeking behavior of social scientists: Ellis's study revisited, *Journal of the American Society for Information Science and Technology* 54 (6) (2003) 570–587.
- [33] W. Meng, Z. Wu, C. Yu, Z. Li, Highly scalable and effective method for metasearch, *ACM Transactions on Information Systems (TOIS)* 19 (3) (2001) 310–335.
- [34] E. Nielsen, Global Internet Population Grows an Average of Four Percent Year-over-year (2003) [Online] Available at: http://www.nielsen-netratings.com/pr/pr_030220.pdf.
- [35] T. Ong, H. Chen, Updatable pAT-tree approach to Chinese key phrase extraction using mutual information: a linguistic foundation for knowledge management, *Proceedings of the Second Asian Digital Library Conference*, 1999, pp. 63–84, Taipei, Taiwan.
- [36] C.J. van Rijsbergen, *Information Retrieval*, 2nd ed., Butterworths, London, 1979.
- [37] J.E. Rush, R. Salvador, A. Zamora, Automatic abstracting and indexing: production of indicative abstracts by application of contextual inference and syntactic coherence criteria, *Journal of the American Society for Information Science* 22 (4) (1964) 260–274.
- [38] E. Selberg, O. Etzioni, Multi-service search and comparison using the metacrawler, *Proceedings of the 4th World Wide Web Conference*, 1995, pp. 195–208, Boston, Mass, USA.
- [39] E. Selberg, O. Etzioni, The metaCrawler architecture for resource aggregation on the web, *IEEE Expert* 12 (1) (1997) 8–14.
- [40] E.H. Shortliffe, The evolution of health-care records in the era of the internet, *Medinfo* 9 (1998) 8–14.
- [41] K. Tolle, H. Chen, Comparing noun phrasing techniques for use with medical digital library tools, *Journal of the American Society for Information Science* 51 (2000) 352–370.
- [42] E. Voorhees, D. Harman, Overview of the sixth text rEtrieval conference, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, 1997, pp. 1–24, Gaithersburg, MD, USA.
- [43] O. Zamir, O. Etzioni, Grouper: a dynamic clustering interface to web search result, *Proceedings of the Eighth World Wide Web Conference*, Toronto, 1999, pp. 1361–1374.
- [44] D. Sullivan, Nielsen NetRatings Search Engine Ratings (2003) [Online] Available at: <http://searchenginewatch.com/reports/article.php/2156451>.



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.



Jialun Qin is a Ph.D. candidate in the Department of Management Information Systems at the University of Arizona. Before joining the Artificial Intelligence Lab, he received the B.S. degree in computer science from Shanghai Jiaotong University in China. His research interests include knowledge management, Web knowledge discovery and extraction, digital library, and social network analysis. Contact him at qin@u.arizona.edu.



Dr. 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, MBA from the SUNY Buffalo, and the Ph.D. degree in Information Systems from New York University. He is author of seven books and more than 120 SCI journal articles covering intelligence analysis, data/text/web mining, digital library, knowledge management, medical informatics, and Web computing. He serves on five editorial boards including: *Journal of the American Society for Information Science and Technology*, *ACM Transactions on Information Systems*, *IEEE Transactions on Systems, Man, and Cybernetics*, *IEEE Transactions on Intelligent Transportation Systems*, and *Decision Support Systems*.