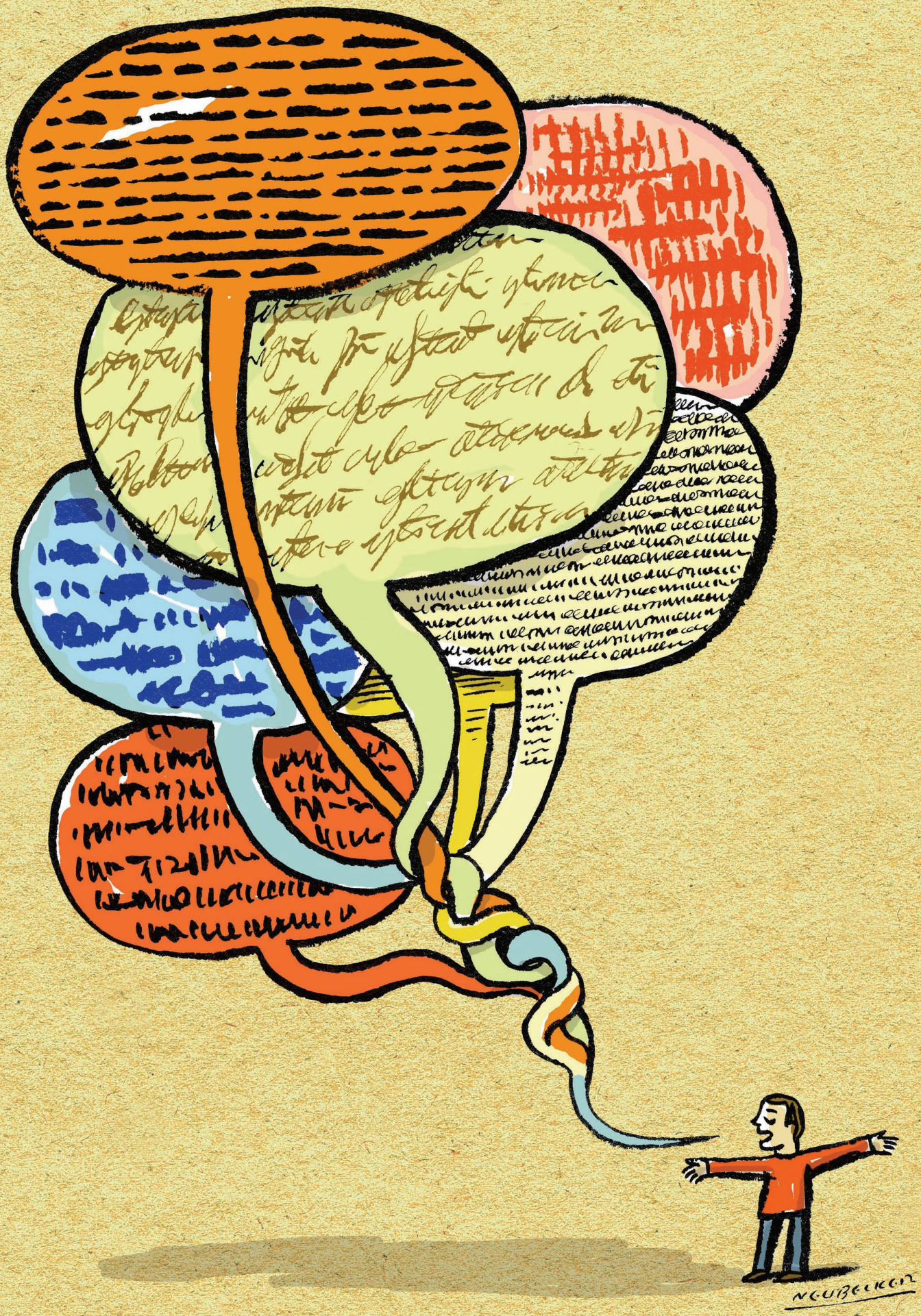
Studies of three prototype Web search portals—in Chinese, Spanish, and Arabic—reveal how to best support non-English Web searching.

WEB SEARCHING IN A MULTILINGUAL WORLD

By Wingyan Chung

Worldwide Internet use has grown tremendously in recent years, most rapidly in non-English-speaking regions. For example, from 2000 to 2007, the online populations in Latin America and the Middle East grew by 577.3% and 920.2%, respectively [9]. At the same time, the number of registered domain names in mainland China (.cn) surged by 137.5% annually [2], fueling the growth of Web pages in Chinese, the second most popular language on the Web. Meanwhile, Arabic Web content was estimated to be doubling every year [1]. Such growth has created demand for better Web searching and browsing in some non-English languages. However, existing Web portals may be unable to meet it because they primarily serve English-speaking users.

ILLUSTRATION BY ROBERT NEUBECKER



Arabic, the fifth most popular language in the world, is spoken by more than 284 MILLION PEOPLE IN SOME 22 COUNTRIES, yet the Arabic Web is still in its infancy, constituting less than 1% of total Web content.

While many research findings on Web searching are available, little has been done on the theoretical and empirical aspects of non-English Web searching. Here, I review Web-search engines in a multilingual world and describe a framework that tries to address these issues. Experimental studies of three prototype Web search portals—in Chinese, Spanish, and Arabic—reveal how to best support non-English Web searching.

English has been the dominant language for information seeking on the Web. But this is not the case for many non-English-speaking users who rely on their native languages to search and browse the Web. The process of information seeking consists of various stages of problem identification, definition, resolution, and solution presentation [12]. Two major information-seeking activities are searching and browsing. In searching, users first decompose their goal into smaller problems, then formulate keyword queries, and finally evaluate the results through serial search or systematic sampling. In browsing, users first transform their general information needs into a problem, then explore the Web content and hyperlinks through such browse-support tools as automatic summarization, clustering, visualization, and Web directories, ultimately evaluating the results by scanning through them.

Techniques proposed to support Web searching and browsing include meta-searching and Web-page preview and overview. Because different search engines employ different methods for page collecting, indexing, and ranking, they may include systematic bias in their search results [10]. Meta-searching is a promising method for alleviating this problem [4]. By sending queries to multiple search engines and collating the set of top-ranked results from each engine, meta-searching can greatly reduce bias in search results and improve coverage. In addition, post-retrieval analysis provides added value to results returned by search engines. Text-categorization techniques help filter Web-page content and provide previews of individual Web pages in the form of summaries. Document-categorization techniques help group Web pages, and document visualization techniques help amplify human cognition in browsing Internet search results. Though used in some search engines, including excite.com and vivisimo.com, meta-searching and information previews and overviews are rarely applied in non-English search engines.

Web searching in a multilingual world is characterized by cross-region and cross-country use of a language, producing regional effects in Web-site design and functionality. For example, Spanish is widely used in Europe, North America, and South America. Arabic is the primary language in the Middle East and North Africa. Chinese is the primary language in mainland China, Hong Kong, and Taiwan. The users of the Fast search engine (www.fastsearch.com), mostly European, input queries more frequently than Excite search-engine users, who focus more on e-commerce topics [11]. These results suggest regional differences on the Web.

SEARCH ENGINES

Several major search engines provide search services to non-English-speaking users. Having more than 160 local domains, Google allows users to restrict search results to pages in 117 languages, providing translation services between English and eight European languages (Dutch, French, German, Greek, Italian, Portuguese, Russian, and Spanish), three Oriental languages (Chinese, simplified and traditional, Korean, and Japanese), and Arabic. AltaVis-

ta's Babel Fish (babel.altavista.com) provides more pairwise translation services between languages (except Arabic). Similar translation services are also provided by Yahoo!, which has regional sites in 24 countries supporting Web search in 37 languages used by 411 million unique users each month. Yahoo!'s diversified services, including online shopping, auctions, email, news, blogs, partnerships with content providers, and instant messaging, enable it to fit comfortably into most aspects of users' lives. Meanwhile, MSN Search has 42 regional sites located in different countries. Its U.S. site supports a local search service for searching information in the user's geographic area. Like Yahoo!, MSN also provides such services as email, instant messaging, news, and entertainment information. Its connection with Microsoft Windows and Internet Explorer has helped it earn an important share in the search market (monitored by SearchEngineWatch.com, a Web site that provides lists and reviews of major and specialized search engines).

Region	Greater China Regions						U.S, Europe, and Latin America								Middle East			
Main language	Simplified Chinese		Traditional Chinese				Spanish								Arabic			
Search engine or portal (headquarters) / Content and Functionality Features	Baidu.com (mainland China)	Sina.com.cn (mainland China)	Yahoo! HK (hk.yahoo.com, Hong Kong)	Timway.com (Hong Kong)	Yam.com (Taiwan)	Openfind.com.tw (Taiwan)	Terra.com (Spain)	Orange.es (France)	Auyantepui.com (Venezuela)	Conexcol.com (Colombia)	Bacan.es (Ecuador)	Yahoo! Telemundo (telemundo.yahoo.com, Spain)	BIWE.com (Spain)	Quepasa.com (Mexico & U.S.)	Ajeeb.com (Kuwait)	Albawaba.com (Jordan)	Weyak.ae (U.A.E.)	Ayna.com (U.S.)
Membership services	✓	✓	✓		✓	✓	✓	✓				✓	✓	✓	✓	✓		✓
Newsgroup/ Weblog search	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Web directory		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Search for Web sites	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Search stock prices		✓	✓		✓			✓				✓						
Online translation			✓									✓			✓			
Search for news	✓	✓	✓		✓	✓		✓				✓	✓		✓	✓		✓
Multimedia search (image, music, video, software)	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Size of collection	Very good	Good	Very good	Fair	Good	Very good	Very good	Very good	Fair	Fair	Fair	Very good	Good	Good	Very good	Good	Very good	Very good
User interface	Good	Fair	Very good	Fair	Very good	Good	Very good	Very good	Good	Good	Fair	Very good	Good	Good	Very good	Very good	Good	Good

Table 1. Search engines and portals in Chinese, Spanish, and Arabic.

While an exhaustive review of search engines in all languages is beyond my scope here, I have reviewed major Web search engines in three emerging languages: Chinese, Spanish, and Arabic. Table 1 lists major search engines and portals in these languages, highlighting important content and functionality features. Chinese is the primary language used by people in mainland China, Taiwan, and Hong Kong. Language encoding, vocabularies, economies, and societies of these regions differ significantly. In mainland China, Baidu.com is a major search engine serving many large enterprises, including Dell (China), Lenovo, and Yahoo! China. It has collected more than a billion Chinese Web pages from mainland China, Hong Kong, Taiwan, and other regions, a collection that grows by several hundred thousand Web pages per day. Another major Web portal in China, Sina.com.cn, provides comprehensive services, including Web search, email, news, business directory, entertainment, and weather forecasts. Leveraging on its rich content and large user base, Sina has its own search engine, iAsk.com, that uses both Web content and usage information to rank Web pages. Other search engines in mainland China include Sogou.com and Zhongsou.com.

The two major search portals in Taiwan are Openfind and Yam. Openfind.com.tw, established in 1998, suggests relevant terms to refine users' search queries, allowing them to find other related items from each search result. Yam.com, established in 1995, provides comprehensive online services involving search in various media in Taiwan, including Web sites and pages, news, forum messages, and local activities. Since 2000, Yam.com has partnered with Google to provide search services.

Due to Hong Kong's bilingual culture, people there rely on both English and Chinese when searching the Web. Yahoo! Hong Kong (hk.yahoo.com) returns results in a variety of categories, including Web sites, pages, and news. Established in 1997, Timway.com searches more than 30,000 Hong Kong Web sites categorized into more than 3,000 groups that attract 2.6 million visits per month.

Spanish is the second most popular language in the U.S., as well as the primary language in Spain and some 22 Latin American countries where regional search engines provide search and browse services. With 19 regional sites, Terra.com offers its services to more than 3.1 million Internet users in the U.S., Spain, and Latin America. A Gallup poll in 2002 described Terra as the most popular search engine in Spain; Orange.es (formerly Wanadoo), a subsidiary of

I recommend that system developers and IT managers INCORPORATE BROWSE SUPPORT AND ANALYSIS TOOLS into their online search systems and portals to augment traditional textual list displays.

France Telecom, was second. Yahoo! Telemundo (Spain, telemundo.yahoo.com), the Spanish version of Yahoo! serving the U.S. and Latin America, provides a Web directory compiled by human editors categorizing millions of listed sites. Yahoo! Telemundo supplements its results with those from Inktomi and Google. Established in 1995 as one of the first search engines to search Spanish information on the Web, BIWE.com provides a variety of services, including a Web directory, email, entertainment, and market information for Spanish-speaking users. Meanwhile, Quepasa.com, with headquarters in the U.S., is a bilingual Web portal (Spanish/English) serving Spanish-speaking populations in the U.S. and Latin America.

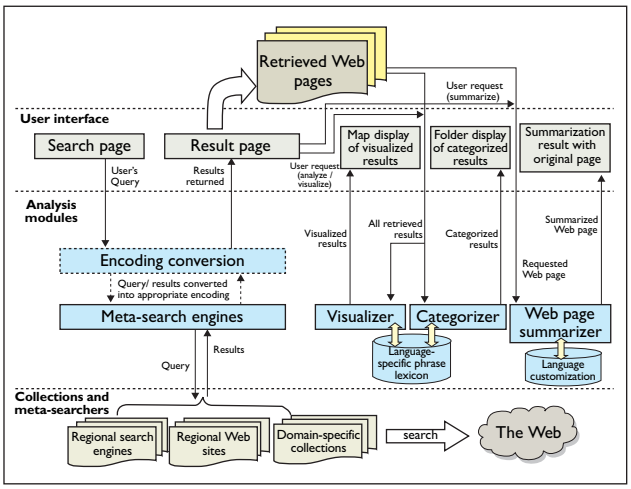


Figure 1. Framework for Web searching in a multilingual world.

Arabic, the fifth most popular language in the world, is spoken by more than 284 million people in some 22 countries, yet the Arabic Web is still in its infancy, constituting less than 1% of total Web content. Four major search engines offer Arab speakers comprehensive services and extensive content coverage. Ajeeb.com, a bilingual Web portal (English/Arabic) launched in 2000 by Sakhr Software Company, includes a multilingual dictionary (Arabic/English/ French/Turkish/German) and a Web directory, "Dalil Ajeeb," which Ajeeb claims to be the world's largest online Arabic directory. Albawaba.com, another Arabic search portal providing comprehensive services, supports searching for both Arabic and English pages; results are classified according to language and relevance. It also meta-searches other search engines—Google, Yahoo!, Excite, Alltheweb, and Dogpile—and provides a comprehensive directory related to 22 Arab countries. Launched in 2000, United Arab Emirates-based Weyak (www. weyak.ae/) offers a range of online services covering more than 1.25 million Arabic Web pages. Based in the U.S. (in New Hampshire), Ayna.com provides an Arabic Web directory, an Arabic search engine, and other services, including a trilingual (Arabic/English/French) email system, chat, greeting cards, personal homepage hosting, and commercial classified ads. Claiming more than 700,000 registered users, Ayna provides access to more than 25 million pages per month. Alexa Research ranks Ayna among the three leading Web sites in the Arab world.

FRAMEWORK FOR WEB SEARCHING

My review found that existing search engines in Chinese, Spanish, and Arabic typically present results in the form of long lists of textual items. While such presentation is convenient for viewing, it may limit users' ability to understand and analyze the results. The collections searched by the search engines are often region-specific and lack a comprehensive understanding of the environment in which they operate. Major English-language search engines, including Google, support searching through non-English resources but fall short of covering domain- and region-specific information. There is a need to better support Web search in some emerging non-English languages. Here, I

describe a framework that addresses some of the needs for Web searching in a multilingual world. As outlined in Figure 1, the framework consists of domain collections, meta-search, statistical language processing, and Web-page summarization, categorization, and visualization.

Reflecting regional and language differences, a careful domain analysis must be conducted by any prospective search-engine developer before a Web portal is built in any particular language. To ensure comprehensive coverage, the analysis should review existing Web portals and technologies, including the characteristics of the language, and select an area or theme for which significant Web resources in the language have been developed. The review should cover regional search engines, government and business Web sites, and news Web sites to select the relevant Web content needed to build a domain-specific collection or for meta-search. Important keywords and URLs relevant to the chosen domain are gathered as seed queries or hyperlinks to build the collection.

Managing a large user base and growing Web content, many non-English Web search engines and portals are challenged to organize their content properly to support convenient browsing and searching. For example, Sina.com.cn includes more than 700 hyperlinks on its home page, each annotated with long textual descriptions in a small font, making browsing difficult, especially for inexperienced Web users. Pre- and post-retrieval analysis is thus needed to alleviate information overload.

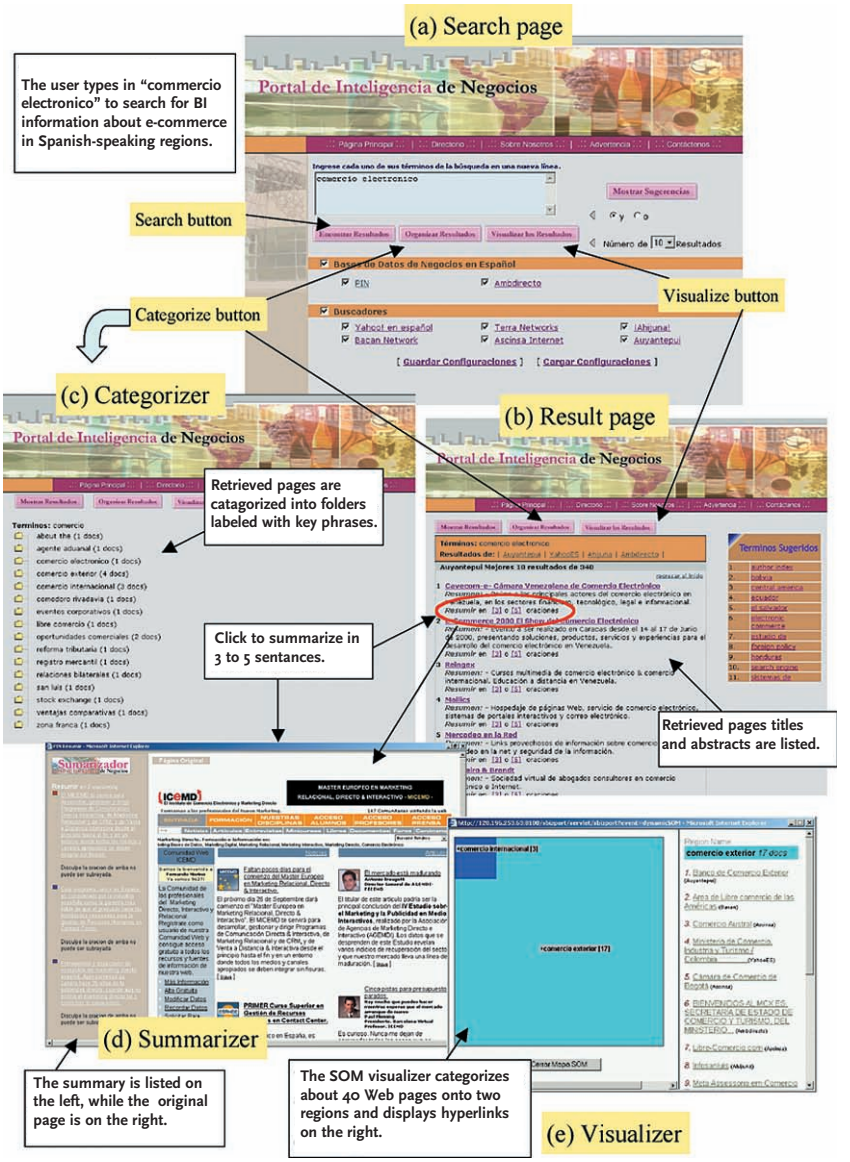


Figure 2. Screenshots from SBizPort.

Modules supporting such analysis include encoding conversion, summarization, categorization, and visualization. Encoding conversion is necessary when a language is used by people in multiple regions and countries using different versions of the same language. For example, the traditional and simplified versions of Chinese differ enormously in written formats, leading to two different input formats in information-retrieval systems; hence, they require encoding conversion for searching across the two language versions. Web-page summarization uses linguistic and heuristic techniques to extract key sentences from the page to represent a summary of the article [8].

Categorization helps organize search results in different groups that are understood more easily. To assist the categorization process, lexicons built by a statistics-based mutual-information approach can provide meaningful phrases in different languages. A neural-network approach called "Kohonen self-organizing map" can be used to categorize and visualize Web pages, helping users navigate on a 2D jigsaw map to identify the set of similar pages or find relevant pages.

Based on the framework, three prototype search portals—in Chinese, Spanish, and Arabic—were developed [3, 6]. The Chinese Web portal (CBizPort) helps users search and browse for business intelligence (BI) in mainland China, Hong Kong, and Taiwan.

Here, BI refers to the product of acquisition, interpretation, collation, assessment, and exploitation of information in the business domain [4]. CBizPort includes two versions of its user interface—one for simplified Chinese, one for traditional Chinese—each with the same look and feel. Relying on a conversion dictionary with 6,737 Chinese characters in each of the two encodings (Big5 and GB2312), the encoding converter converted all Chinese characters into the encoding of the interface version. The eight information sources used in the portal's meta-searching are major Chinese search engines and business-related portals from the three regions. Relying on two Chinese-phrase lexicons to extract phrases, the portal's categorizer organizes retrieved Web pages into various folders labeled with the key phrases in the page summaries and titles.

The Spanish Web portal (SBizPort) supports searching and browsing of business information from 22 Spanish-speaking regions. In addition to keyword searching, summarization, and categorization, like those in CBizPort, SBizPort provides a comprehensive collection of business Web pages for searching and supports visualization of retrieved pages (see Figure 2). Users visualize Web pages by clicking on a region to see a list of pages on the right and open pages by clicking the link-embedded titles.

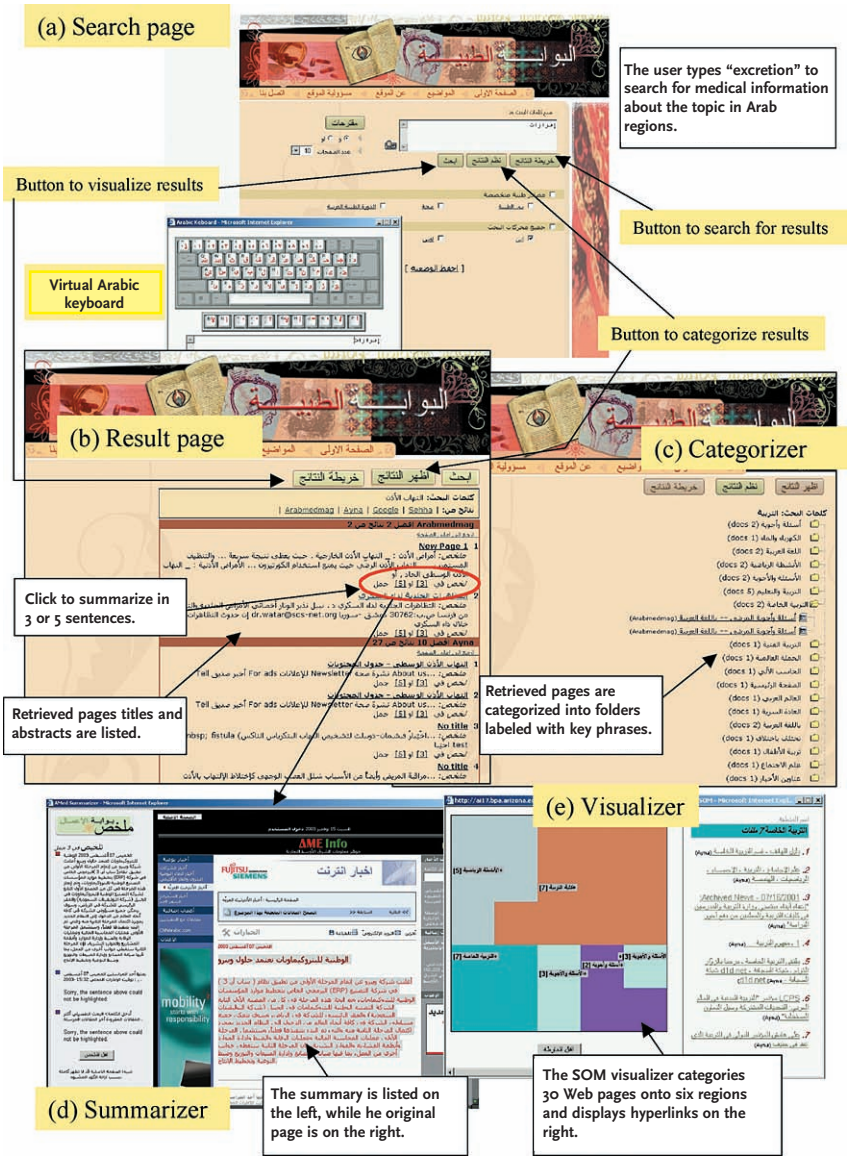


Figure 3. Screenshots from AMedPort.

The Arabic Web portal, AMedPort (see Figure 3) focuses on the medical domain in some 22 Arab regions and supports all search and browse functions available in SBizPort. AMedPort includes a customized user interface with a right-to-left text display and a virtual keyboard to facilitate Arabic input.

EXPERIMENTAL FINDINGS

Sixty native speakers participated in the experiments (detailed in [3, 6]) of the three Web portals to evaluate the framework's usability in supporting Web searching in a multilingual world (see Table 2). In each experiment (about an hour), a Web portal was compared with a benchmark search engine in each of the three languages. Each subject was introduced to the portal and the benchmark search engine and was randomly assigned different task scenarios (one scenario per system). Each scenario contained three or four search and browse tasks based on standards developed through the National Institute of Standards and Technology Text Retrieval Conference (trec.nist.gov). All questionnaires used in the experiment were administered in the subjects' native language. Subjects spent an average of three minutes to finish a search task and eight minutes to finish a browse task. The order in which the systems were used was randomly assigned to avoid bias due to the sequence of use.

Several information-retrieval measures—precision, recall, and F value—revealed the search and browse effectiveness of the system by computing the ratios between the number of relevant results found by a

Users must be cautioned that THE TOOLS ARE STILL PRONE TO ERROR due largely to ambiguities in natural-language processing and high computational costs.

subject and the number of all results found by the same subject or by an expert. Domain experts provided answers for judging the subjects' browse-task performance. After using a system, a subject filled out a questionnaire with comments and satisfaction ratings on a seven-point Likert scale.

In the CBizPort experiment with 30 Chinese subjects from mainland China, Hong Kong, and Taiwan and with three Chinese business academics and practitioners serving as experts, the results show that the effectiveness of existing Chinese search engines can be improved significantly by adding CBizPort. However, no significant difference was found in the effectiveness of the summarizer and categorizer in the Chinese portal and in their user-satisfaction ratings. Despite this, 11 subjects commented that the summarizer and categorizer facilitated their understanding and searching. These results indicate that CBizPort could augment these search engines in searching and browsing Chinese business information but also that the summarizer and categorizer need further improvement in precision and browse support.

In the SBizPort experiment with 19 Spanish subjects from six countries—Colombia, Mexico, Panama, Peru, Puerto Rico, and the U.S.—and a veteran Spanish business consultant serving as the expert, the SOM visualizer achieved significantly better browse effectiveness than BIWE, the benchmark search engine, showing that the tool helps alleviate information overload and supports browsing effectively. The use of a domain-specific collection achieved higher mean accuracy and search efficiency than not using it, though the differences were not significant. Subjects rated SBizPort significantly better than BIWE, citing the precision and relevance of returned results. These results further indicate that the information-visualization tool can be an alternative to presenting search results as text in a list.

In the AMedPort experiment with 11 Arab subjects from five countries—Iraq, Jordan, Lebanon, Mauritania, and Morocco—and an Arab microbiology researcher serving as the expert, the AMedPort achieved a significantly higher mean accuracy, efficiency, and satisfaction rating than (and comparable browse effectiveness to) Ayna, the benchmark search engine. Nine subjects said AMedPort was useful and provided more topics and information than the benchmark. The portal provided high-quality information from many sources but needs improvement in both its summarizer and categorizer.

There was a probability of 0.05 (or less) that the confirmed results were actually not statistically valid in the experiments in which the best search engines in the respective languages were used as benchmarks.

Portal	CBizPort	SBizPort	AMedPort
Research questions	• How does the framework support Web searching in a multilingual world? • How can Web searching and browsing be made more effective by using the portals developed by the framework? • What are the lessons learned and implications for non-English Web searching?		
Languages	Traditional Chinese is used in Hong Kong and Taiwan, while simplified Chinese is used in mainland China. They have different encoding formats. Word segmentation is a problem.	Spanish is used in many regions and countries in North and South Americas and Europe; Catalan (another version of Spanish) is widely used as well.	Arabic is used in more than 22 countries in the Middle East and North Africa. It is written from right to left. Web content in Arabic is generally not rich.
Regions	Mainland China, Hong Kong, Taiwan	Mexico, Honduras, Costa Rica, Panama, Colombia, Venezuela, Ecuador, Peru, Chile, Argentina, Spain, Bolivia, Paraguay, Uruguay, Guatemala, Nicaragua, U.S.	Lebanon, Saudi Arabia, Bahrain, Canada, Tunisia, Kuwait, Egypt, Switzerland, United Arab Emirates, Russia, U.K., U.S.
Benchmarks	Sina.com.cn, Openfind.com.tw, HK.yahoo.com	BIWE.com	Ayna.com
Result highlights	CBizPort's meta-searching, summarization, and categorization are helpful for searching and browsing Chinese Web pages in the three regions. CBizPort needs improvement on search precision and information quality.	SBizPort achieved a better browse performance and satisfaction rating than the benchmark but needs to be improved on its domain-specific collection.	AMedPort achieved better search accuracy than comparable browse performance in Ayna. AMedPort obtains better user ratings in user satisfaction.

Table 2. Experimental studies on developing Web search portals in various languages.

LESSONS LEARNED

The results from these experiments demonstrate that

the framework supports Web searching in a multilingual world. Post-retrieval analysis techniques (such as summarization and visualization) were found to alleviate information overload but also that the extent of such improvement varies across domains. Summarization and categorization did not achieve significant improvement in the CBizPort study. In the SBizPort and AMedPort studies, information visualization achieved significant performance improvement in Web-search results. The ability to visualize a large number of search results was essential for good performance in all three portals.

I recommend that system developers and IT managers incorporate browse support and analysis tools into their online search systems and portals to augment traditional textual list displays. Such tools can be used to summarize Web-page textual descriptions [6], support query formulation [7], visualize emerging events related to their environment and organizations [5], and categorize search results into hierarchies or maps [4]. However, users must be cautioned that the tools are still prone to error due largely to ambiguities in natural-language processing and high computational costs that may not be economical for small Web sites.

Factors to be considered when adopting the tools include the extent to which the Web-page collection provides sufficient statistical information for machine learning, adequate hardware and software to support intensive computation, availability of a work force to improve the Web-site interface and accommodate new presentation choices, characteristics of the language used, and user IT literacy.

Across a variety of languages and domains, I found significant differences in the development of Web-search portals, technologies, and language use. For instance, the growth of Internet use in mainland China (but relative lack of comprehensive Web search and browse support) strongly suggests the need for future improvements. While Web-search technologies in Taiwan are more mature, there is likely room for new technologies developed specifically for processing Chinese. The strong growth of the Chinese- and Spanish-speaking online populations will likely persist in the coming years, further emphasizing the need for better, more integrated Web-search portals that deliver results in a variety of formats and provide richer information for the regions and the communities that use the languages. The increasing amount of Arabic Web content and online population, along with economic and political developments in Arab regions, will continue to fuel the growth of many Arabic Web sites that remain mostly underdeveloped today. The research I've reported here will likely contribute to a better understanding of related developmental and experimental issues.

My ongoing work includes developing scalable techniques to collect, analyze, and visualize Web information in different languages, studying user needs in non-English Web search, and exploring the effect of new techniques in information exploration and analysis. This effort will contribute to Web searching and browsing in a multilingual world. C

References

1. Abbi, R. *The Current Status of the Internet in the Arab World.* UNESCO Observatory on the Information Society (2002); www.unesco.org/cgibin/webworld/portal_observatory/cgi/jump.cgi?ID=2329.
2. China Internet Network Information Center. *The 20th Statistical Survey Report on the Internet Development in China;* Beijing, China, 2007; www.cnnic.net.cn/uploadfiles/pdf/2007/7/18/113918.pdf.
3. Chung, W., Bonillas, A., Lai, G., Xi, W., and Chen, H. Supporting non-English Web searching: An experiment on the Spanish business and the Arabic medical intelligence portals. *Decision Support Systems 42,* 3 (Dec. 2006), 1697–1714.
4. Chung, W., Chen, H., and Nunamaker, J. A visual framework for knowledge discovery on the Web. *Journal of Management Information Systems 21,* 4 (Spring 2005), 57–84.
5. Chung, W., Chen, H., Chaboya, L., O'Toole, C., and Atabakhsh, H. Evaluating event visualization: A usability study of the COPLINK spatio-temporal visualizer. *International Journal of Human-Computer Interaction 62,* 1 (Jan. 2005), 127–157.
6. Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T.-H., and Chen, H. Internet searching and browsing in a multilingual world: An experiment on the Chinese Business Intelligence Portal. *Journal of the American Society for Information Science and Technology 55,* 9 (July 2004), 818–831.
7. Leroy, G., Xu, J., Chung, W., Eggers, S., and Chen, H. An end-user evaluation of query formulation and results review tools in three meta-search engines. *International Journal of Medical Informatics 7,* 11–12 (Nov.–Dec. 2007), 780–789.
8. McDonald, D. and Chen, H. Summary in context: Searching versus browsing. *ACM Transactions on Information Systems 24,* 1 (Jan. 2006), 111–141.
9. Miniwatts International. *Internet Usage Statistics: The Internet Big Picture* (updated Nov. 30, 2007); www.internetworldstats.com/stats.htm.
10. Mowshowitz, A. and Kawaguchi, A. Bias on the Web. *Commun. ACM 45,* 9 (Sept. 2002), 56–60.
11. Spink, A., Ozmutlu, S., Ozmutlu, H., and Jansen, B. U.S. versus European Web searching trends. *SIGIR Forum 36,* 2 (Fall 2002).
12. Wilson, T. Models of information behavior research. *Journal of Documentation 55,* 3 (June 1999), 249–270.

Wingyan Chung (wchung@scu.edu) is an assistant professor in the Department of Operations and Management Information Systems of the Leavey School of Business at Santa Clara University, Santa Clara, CA.

DOI: 10.1145/1342327.1342335