

Organizing domain-specific information on the Web: An experiment on the Spanish business Web directory

Wingyan Chung^{a,*}, Guanpi Lai^b, Alfonso Bonillas^b, Wei Xi^b, Hsinchun Chen^b

^a*Department of Operations and Management Information Systems, Leavey School of Business, Santa Clara University, 500 El Camino Real, Kenna 323, Santa Clara, CA 95053, USA*

^b*Artificial Intelligence Laboratory, Department of Management Information Systems, University of Arizona, 1130 East Helen Street, McClelland Hall 430, Tucson, AZ 85721, USA*

Received 6 November 2006; received in revised form 18 May 2007; accepted 15 August 2007

Communicated by P. Zhang

Available online 31 August 2007

Abstract

Web directories organize voluminous information into hierarchical structures, helping users to quickly locate relevant information and to support decision-making. The development of existing ontologies and Web directories either relies on expert participation that may not be available or uses automatic approaches that lack precision. As more users access the Web in their native languages, better approaches to organizing and developing non-English Web directories are needed. In this paper, we have proposed a semi-automatic framework, which consists of anchor directory boosting, meta-searching, and heuristic filtering, to construct domain-specific Web directories. Using the framework, we have built a Web directory in the Spanish business (SBiz) domain. Experimental results show that the SBiz Web directory achieved significantly better recall, *F*-value, efficiency, and satisfaction rating than the benchmark directory. Subjects provided favorable comments on the SBiz Web directory. This research thus contributes to developing a useful framework for organizing domain-specific information on the Web and to providing empirical findings and useful insights for end-users, system developers, and researchers of Web information seeking and knowledge management.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Internet; Web; Browsing; Ontology; Business intelligence; Spanish; Non-English Web browsing; Knowledge management

1. Introduction

Although the Internet has greatly facilitated searching for information, users are often overwhelmed with a large amount of semi-structured and unstructured information. Aside from searching, browsing well-structured Web directories may help users to explore interesting yet unfamiliar domains. For example, business analysts can browse a Web directory on technology news Web sites to find the latest technological developments in their field as well as in related fields. Customers can obtain new product information or company profiles from business Web directories, which provide links to high-quality company Web sites and online stores. These Web directories play an

important role in facilitating Web browsing and decision-making. Examples of general Web directories include the Yahoo and Directory Mozilla (DMOZ) directories. Relying on the directory's labels and categorization, users can spend less time in finding information and experience a better Web navigation. Nevertheless, constructing a high-quality Web directory without much expert knowledge and extensive human efforts has challenged developers of Web portals. On the other hand, as the Internet grows in popularity worldwide, more users access Web content in their native languages (Chung et al., 2004). A survey shows that the majority of the total global online population (64.2%) lives in non-English-speaking regions (Global Reach, 2004), where Internet usage has grown the most (Miniwatts International, 2006). Better approaches to organizing non-English Web content will improve the browsing experience of a large number of people in the world.

*Corresponding author. Tel.: +1 408 554 4329; fax: +1 408 554 5157.
E-mail address: wchung@scu.edu (W. Chung).

In this paper, we describe a semi-automatic framework for constructing high-quality Web directories. Human knowledge of domain and language was used to guide the establishment of a Web directory framework and to enhance the quality of the directory. Automated meta-searching of high-quality information sources allowed us to efficiently fill the framework with meaningful and relevant items. To demonstrate the value and usability of the framework, we have built a Web directory for the Spanish-speaking business community. We leveraged the wide coverage of search engines and preciseness of human knowledge to obtain highly relevant directory content. An experiment involving native Spanish speakers was conducted to compare the directory with an existing Web directory in the Spanish business (SBiz) domain. Lessons learned and implications are discussed.

The remaining sections are structured as follows: Section 2 reviews prior work related to this research. Section 3 describes the research gaps and questions. Section 4 proposes the research framework and Section 5 details the research test bed. Section 6 explains the evaluation methodology and experimental design. Section 7 reports experimental findings and discusses the implications. Section 8 concludes the study and presents future directions.

2. Literature review

Organizing knowledge for domain-specific applications has challenged academics and practitioners due to the abundance of information and the difficulty of categorizing the information. Traditionally, ontologies are used in libraries to perform this task and have been extended to other fields. As the Web becomes a common platform for information storage and retrieval, Web directories built upon ontologies also are available. Developments of the Web in non-English languages have further fueled the demands for Web directories. It is therefore useful to review previous work in ontologies, Web directory development, and non-English Web resources. In particular, we review existing Web resources in Spanish, the language chosen for building our research test bed.

2.1. Ontologies and their developments

The term “ontology” originated from the field of philosophy to indicate a systematic account of existence. Through the centuries, philosophers such as Plato, Aristotle, Kant, and Hegel tried to classify all existing entities to develop ontologies (Sowa, 2000). John Dewey (1925) said that “knowledge is classification”. Langridge (1992) stated that without classification there could be no human thought, action, or organization. Nowadays, librarians use the term to refer to knowledge representation in a domain and to classify library materials. For example, the Dewey Decimal Classification (DDC) System and the US Library of Congress Classification system are heavily used by universities and public libraries.

Ontologies have been used in other areas than libraries to provide a consistent knowledge base for a specific domain and to combine different information sources (Grüninger and Lee, 2002). For example, the US National Library of Medicine developed the UMLS Semantic Network for the biomedical domain (<http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html>). With its 134 semantic types and 54 links, the network categorizes all concepts represented in the UMLS Metathesaurus and represents important relationships in the domain. Hovy (2003) described the way an ontology was used to access different government data sources. Wache et al. (2001) described methodologies to integrate ontologies from different information systems. Holsapple and Joshi (2002) proposed a collaborative approach to ontology design by using a Delphi technique to develop ontologies. Based on their study on designing knowledge management ontology, they found panelist attrition to be a major problem because the panelists were not willing to participate in multiple rounds of time-consuming review. In another study, Kim (2002) predicted that XML use would be preferable to ontology use and innovation of modeling tools would allow knowledge workers to codify idiosyncratic information. However, his prediction was based only on conceptual analysis and more research on ontology development is thus needed. Although ontologies have been developed for many applications, there is a lack of ontologies for developing Web directories. In particular, we did not find previous attempts to construct non-English Web directories based on existing ontologies or information hierarchies. A review of previous work on Web directories would provide a better understanding of how existing Web directories are constructed.

2.2. Constructing Web directories

Previous work in constructing Web directories falls into two categories: (1) manual identification and categorization of Web resources; and (2) automatic construction of directories using machine learning or Web mining techniques.

2.2.1. Manual categorization

Manual identification and categorization have been used in various domains, ranging from general search engines to domain-specific Web portals. The Open Directory Project, also known as DMOZ (<http://dmoz.org>), is constructed and maintained by a large, global community of volunteer editors. With 71,053 human editors, it lists more than 5,199,707 sites classified into over 590,000 categories (as of January 2006). The rationale of DMOZ is to use extensive human work to combat growth of human-created Web resources, which often grow with the size of the online population. Currently, DMOZ powers the core directory services of many search engines, including Netscape Search, AOL Search, Google, Lycos, HotBot, and DirectHit. However, the quality of the directory constructed by this method depends highly on the volunteer

editors' domain knowledge, which usually varies from person to person. Moreover, the approach is not scalable because many Web resources are generated automatically, making its growth more rapid than the growth of the online population. It may not suit the needs of other less dominant communities on the Web (e.g., the SBiz domain) because of the lack of expert knowledge and involvement.

The most well-known human-created Web directory is the Yahoo! Directory (<http://dir.yahoo.com/>), which is built and maintained by a team of paid editors who organize Web sites into categories and subcategories. The hierarchy has 4–5 levels and contains 14 main categories (e.g., Arts & Humanities, Business & Economy, Computers & Internet, Education, etc.), each having around 20–45 subcategories. Some categories contain more than 100,000 Web sites in them. Although Yahoo! Directory contains a lot of Web sites, the small team of editors, with limited knowledge and time, may be ineffective and inefficient in identifying and evaluating Web resources. They may be particularly limited in finding and categorizing Web resources that are not as popular as those already listed in the directory (e.g., non-English Web sites).

Based in California, the Librarian's Index to the Internet (LII, <http://lii.org/>) provides a searchable, annotated subject directory of more than 12,000 Internet resources selected and evaluated by librarians for their usefulness to users of public libraries. Over 100 contributors from libraries in California and Washington State participate in building and maintaining the index. The process of building and updating the directory is facilitated by a Web-based system, through which human indexers can edit existing records, create new records, and preview the edited records (Librarians' Internet Index, 2006). This gives flexibility to indexers when they edit the records. However, similar to DMOZ and Yahoo! directories, the index suffers from a lack of scalability and efficiency. It also does not cater well for less dominant communities.

2.2.2. Automated approaches to directory construction

Beside manual methods, automatic approaches to constructing directory and ontology have been proposed in previous research. Sato and Sato (1999) developed an automated editing system that generates a Web directory from a given category word without human intervention. Based on the category term, the system gathers instance names belonging to this category. For every instance name, it uses two search engines, Goo and InfoSeek, to collect related Web pages and then removes duplicated contents. Then the system generates a Web directory organized according to geographic regions. Although the system is efficient for generating a directory from a category label and runs without human intervention, the generated directory only has one level that restricts its use in more complicated browse tasks. Relying on only two search engines introduces bias in content collection.

In another research, Chuang and Chien (2003) propose a query-categorization approach to facilitate the construc-

tion of Web directories. Obtained from search engine log files, a total of 18,017 query terms were categorized by using a hierarchical agglomerative clustering algorithm into a predefined two-level hierarchical structure, consisting of 14 major categories together with 100 subcategories. Meta-searching from Google using the query terms was used to provide relevant documents, from which the title and description were extracted as their representations. Then the subject category of each query term is determined. The approach requires search engine log data that are typically not accessible by outsiders. The predefined directory structure also does not suit domains having relatively smaller coverage on the Web.

To automatically generate Web directory and identify directory labels, a self-organizing map approach was proposed that built up the relationships among Web pages and extracted category labels (Yang and Lee, 2003). The approach recursively generated super clusters via congregating neighboring neurons, then created the hierarchical structure of Web directories. However, the directory generated by this approach tends to include noisy content. Without precise filtering and editing, the directory is less reasonable and logical. Variations of the approach have been proposed. For example, Chen et al. (1996) proposed a self-organizing approach to Internet search and categorization. The approach used automatic textual analysis to categorize Web pages and then employed a multilayered neural network-clustering algorithm called Kohonen self-organizing feature map to classify Web pages based on their content. Subject-specific searching or browsing within the subject categories were supported by the approach. Focusing on the business intelligence domain, Marshall et al. (2004) also used the Kohonen self-organizing map algorithm to categorize and visualize retrieved Web pages onto a single-layer jigsaw map having different regions representing sets of categorized Web pages. Experimental findings showed that the visualizer contributed significantly to subjects' performance in browse tasks.

Kumar et al. (2001) presents a two-phase, semi-automatic approach to directory construction: (1) an ontologist uses a rich query language for a node to search from the *Clever* system (based on HITS algorithm (Kleinberg, 1999)) a high-quality set of links about his topic; and (2) the ontologist edits and annotates the resulting set of links to create an appropriate externally visible node about the topic. The approach combines human knowledge with search engine efficiency. But the quality of results depends highly on the ontologist's limited knowledge. Moreover, searching only the *Clever* system limited the coverage of results.

Stamou et al. (2005) developed an approach for automatically assigning Web pages to a directory framework based on the linguistic information in Web textual data. The approach leveraged a variety of lexical resources such as WordNet (Fellbaum, 1998), Suggested Upper Merged Ontology (SUMO) (Pease et al., 2002), Google directory (<http://dir.google.com/>), and WordNet Domains

(<http://wndomains.itc.it/>) to build a subject hierarchy and to define concepts in the hierarchy. Each Web page was reduced to a lexical chain consisting of 50 thematic words extracted from the page based on thematic scores of word pairs. Then each page represented by a lexical chain was mapped to the hierarchy's nodes based on a relatedness score that was higher than a threshold (0.5) to support non-exclusive categorization. Approximately 114 thousand Web pages were collected as a test bed, of which 23% (26,121 pages) could not be categorized by the approach. With that portion removed, an experiment showed that 75.1% of the pages categorized by the approach were in the Google directory from which the hierarchy was developed. Counting the 26,121 pages that failed to be categorized, only 58% of the collected pages were correctly classified. This suggested a serious weakness of this automated approach in dealing with noisy Web data. In addition, the approach could only be used to categorize English Web pages because of the English lexical resources it was based on. The highly predefined nature of the hierarchy combined with the unsatisfactory categorization accuracy makes the approach not promising for constructing Web directories, especially for non-English domains that have a rapidly growing Web content.

2.3. Domain-specific Web resources: a review of Spanish Web search engines

As more people use the Internet to search and browse for domain-specific information, major search engines have attempted to provide services targeted to specific communities. Regional search engines supporting more localized searching have emerged. In addition to English, these regional search engines typically accept queries in a user's native language and return pages from the regions being served.

In the following, we have chosen Spanish Web search engines and directories to be our focus of review because of several reasons. First, Spanish is the second most popular language in the United States and the primary language of Spain and some 20 countries. It was estimated that the Spanish-speaking population in the United States is over 22 million and Spanish is the fourth most widely used language in the world (Sí Spain, 2006). In 2005, there were more than 390 million people living in these mostly Latin American countries, consisting of more than 6% of the world's population (CountryWatch Inc., 2006).

Second, the population of these countries is expected to grow significantly in the coming decades. It is estimated that the proportion of Spanish-speaking population to the world's population will increase to 7.32% by the year 2030, with an average annual growth rate of 1.47% (higher than the 1.33% average growth rate of the world's population) (CountryWatch Inc., 2006). This higher growth rate is expected to fuel the growth of Spanish Web content, thereby promoting the need for a better browsing support for Spanish Web sites. Third, the North America Free

Trade Agreement (NAFTA) (Office of NAFTA and Inter-American Affairs, 2004) and the Central America Free Trade Agreement (CAFTA) (The Office of the United States Trade Representative (USTR), 2006) are expected to promote economic activities of most Latin American businesses, many of which use the Web to support their operations and to provide services. Therefore, the SBiz domain increasingly represents important segments of the Web that individual users and multinational organizations are interested in. Because of the importance of the SBiz Web domain, a survey of major search engines and Web directories in Spanish follows.

Major search engines in Spain are Terra and Wanadoo. Terra (<http://www.terra.com/>) offers services to more than 3.1 million Internet users in Europe and the Americas. According to a Gallup poll in 2002 (Gallup, 2002), Terra was voted the most popular search engine in Spain; Wanadoo (<http://www.wanadoo.com/>), a subsidiary of France Telecom, was rated second. Terra serves more than 3 million Internet users in Spain, Latin America, the United States, and many European countries. With 9.3 million customers in June 2004, Wanadoo is currently the leading Internet service provider in France and the United Kingdom.

Spanish search engines and Web directories serving Latin America include Yahoo Español, Ahijuna, Auyantepui, Quepasa, Bacan, and Conexcol. Yahoo Español (Spain, <http://espanol.yahoo.com/>), the Spanish version of Yahoo, provides a human-compiled Web directory developed by about 150 editors who categorized more than one million listed sites. Yahoo Español (YahooES) also supplements its results with those from Inktomi and Google. Inktomi matches also appear to users after all YahooES matches have first been shown. Established in 1995, BIWE (<http://www.biwe.com/>) was one of the first search engines and Web directories for searching and browsing Spanish information on the Web. It supports searching for news, products, images, and other information and provides a variety of services including Web directory, email, entertainment, and market information for Hispanics. Headquartered in the United States, Quepasa (<http://www.quepasa.com/>) was launched in 1997 and is a bilingual Web portal (Spanish and English) serving Hispanic populations in the United States and Latin America.

The following Spanish search engines and Web directories primarily serve their own or adjacent regions. Launched in 1998, Ahijuna (Argentina, <http://www.ahijuna.com.ar/>) provides searching of Argentina Web sites and other Spanish Web sites. It contains a Web directory with 14 categories having a total of 7578 hyperlinks. Based in Venezuela, Auyantepui (<http://www.auyantepui.com/>) provides a searchable Web directory of Spanish sites. It grew from 14 categories listing 117 Web sites in 1996—550 categories with over 18,000 Web sites in 2002. Launched in 1998, Conexcol (Colombia, <http://www.conexcol.com/>) provides a searchable Web directory containing 14

categories having 400 subcategories and 13,214 Web sites' URLs. With more than 150,000 unique visitors per month, it is one of the four most often visited sites in Colombia. Bacan (Ecuador, <http://www.bacan.com/>), which began its operations in 1996, provides services such as news, email, online chat, entertainment, and shopping guides. Every month Bacan has 80,000 individual visitors and generates more than 2 million hits. Ascinsa Internet (<http://www.ascinsa.com/>) is widely used in Peru and contains Web sites from Latin American countries and the United States. It also contains a directory listed by countries and then by domains.

3. Research gaps and questions

While the concept of ontology has been proposed for a long time, there has been little work on using existing ontologies or information hierarchies to construct domain-specific non-English Web directories. From our literature review, we found that previous approaches to building information directories have several limitations. On the one hand, manual approaches typically introduce biases due to limited knowledge of the group of directory editors. The fact that many Web pages are generated dynamically also makes this approach not scalable to the rapid growth of the Web. On the other hand, automatic approaches lack precision in identifying category terms and organizing items inside the directory. Previous efforts relying on such approaches typically exploit limited information sources, thus the quality of the resulting directory is limited. The lack of expert knowledge in many of these approaches also creates problems in the usability of the directory created.

Despite providing different types of information, existing search engines and Web directories in Spanish typically serve only a few regions but not the entire Spanish-speaking community. Major English search engines, though providing resources in Spanish, lack comprehensive Web directories to support Web browsing.

To our knowledge, no previous attempt has been made to develop an approach that cannot only be used to construct Web directories for English domains but also is applicable to domains having relatively smaller coverage

on the Web (e.g., non-English domains). Therefore, we are interested in answering three research questions as follows.

1. How can we combine human preciseness and machine efficiency in a framework to construct high-quality domain-specific directories in different languages to support Web browsing?
2. To what extent the directories developed by the framework improve Web browsing performance and assist in human analysis?
3. What is the user satisfaction achieved by using the directory for browsing domain-specific information on the Web?

4. A semi-automatic framework

To address the research questions, we have developed a semi-automatic framework for constructing high-quality domain-specific Web directories. The framework consists of three steps: anchor directory boosting, meta-searching, and heuristic filtering. The rationale of using these steps was to utilize the precision of human identification and the efficiency of automated Web searching to assist in the process of Web directory construction. Using the framework, we tried to overcome problems found in previous research by combining human knowledge and machine efficiency, while incorporating various information sources to ensure a high quality of content. Fig. 1 summarizes the steps of the framework.

4.1. Anchor directory boosting

This step tries to leverage existing general Web directories and human domain knowledge to construct a directory framework that consists of topical labels organized in a hierarchical structure. A comprehensive review encompassing domain-specific Web sites, search engines resources, public information providers (e.g., library Web sites, government sites, online newspapers), and other relevant resources should be used to identify an anchor directory. This anchor directory should be chosen based on

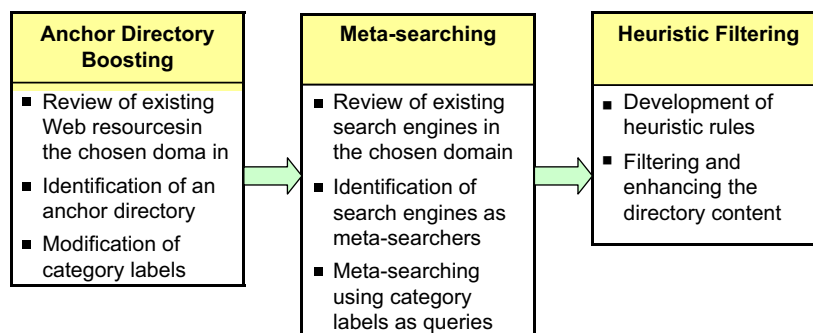


Fig. 1. A semi-automatic framework for constructing Web directories.

the comprehensiveness of domain coverage and richness of Web content. The directory labels collectively serve as a base framework for further modification, which includes changing category labels and enriching the domain-specific content.

4.2. *Meta-searching*

Because searching multiple high-quality search engines has been shown to provide higher quality of results than relying on only a small number of search engines (Mowshowitz and Kawaguchi, 2002), this step uses meta-searching to collect directory items (Web site URLs) automatically to fill in the directory framework (obtained from the previous step). By sending queries to multiple search engines and collating the set of top-ranked results from each search engine, meta-searching can greatly reduce bias in search results and improve coverage. The set of search engines to be used as meta-searchers should be chosen based on a comprehensive review of Web search engines. Category labels of the directory framework are used as input queries for meta-searching. Because search engines typically return a large number of duplicating results, only top-ranked results should be used (with duplicates filtered) in order to limit the scope of coverage.

4.3. *Heuristic filtering*

This step aims to enhance the quality of the automatically generated directory (from the previous step) by filtering out non-relevant items and by adding necessary items. Similarly to the first step, human domain knowledge is used to guide the filtering and enhancing of the directory. A number of heuristic rules must be established to ensure consistency in the work. Both general rules and domain-specific rules should be developed to remove non-relevant items and to include items that might have been missed in the meta-searching process. The rules also help to maintain scalability of the framework in constructing Web directories in different domains.

5. The Spanish business Web directory

Using the framework, we have developed a Web directory called the SBiz Intelligence Web directory. We define the SBiz domain as the segment of Web that serves the needs of Spanish-speaking communities in their economic and business activities. This domain includes business Web sites and other business-related Web sites in Spanish and was chosen because of several reasons. As explained in Section 2.3, Spanish is one of the most popular languages in the World. Because Latin America will have rapid economic growth in the coming decades (Caramelli, 2003), Internet usage of the region is growing rapidly as well. Better approaches to delivering Spanish Web content are likely to benefit lots of Spanish people. Furthermore, Spanish businesses frequently use the Web to provide

information to customers, but well-structured Spanish Web directories are not widely available. The chosen SBiz intelligence domain also has a relatively smaller coverage on the Web than its English counterpart (e.g., Chung et al., 2005).

In the following, we explain the steps of implementing the framework in the context of building the SBiz Intelligence Web directory.

5.1. *Anchor directory boosting: identification and modification of category labels*

Based on our review on Web search engines and directories summarized in Section 2.3, we chose the DMOZ directory as the anchor directory because of its comprehensive business directory and its wide acceptance, as seen in the fact that it is used by such major search engines as Google, AOL search, and Netscape Search. Since the DMOZ directory is mainly used in English-speaking regions, we removed 546 nodes from the original 779 nodes of its business sub-directory, leaving 233 nodes in the directory. In general, the following conditions would necessitate modification to category items: the category items did not reflect the information of the domain being considered, the items were not relevant to the domain being considered, and the category labels already had appeared in other places in the directory. While it is possible that a label may belong to more than one category or sub-categories, we considered exclusive categorization in which a category label appears only once in the directory.

To ensure that the resulting directory framework contained items specific to the SBiz domain (rather than English business domain), we reviewed a SBiz directory provided in BIWE (Buscador en Internet para la Web en Español, <http://www.biwe.es/>) to add 81 nodes to the 233 nodes, resulting in a 314-node SBiz directory framework. The 81 nodes were selected because they were directly relevant to the SBiz domain but do not appear in the DMOZ (English) Web directory. Based in Spain, BIWE is a major general Spanish search engine serving more than four million Web pages per month for the Spanish-speaking community. Its Web directory, with 16 main categories and 633 sub-categories, was chosen as it is more comprehensive, and contains more Spanish resources than existing Spanish Web directories such as Degerencia (<http://www.degerencia.com/>) and Gestipolis (<http://www.gestipolis.com/>). The following are some examples of the modification. A “Tourist companies” category was added to include more items related to tourism, a major economic activity in the Spanish-speaking business world. A “Business and Economic Information Sources” category was added so that more information about the Spanish-speaking business world would be included (e.g., financial magazines, stock market news, etc.).

Table 2 lists two fragments of the resulting SBiz Web directory framework translated into English.

5.2. Meta-searching: automatic generation of directory items

To fill in the SBiz directory framework, we used seven major search engines (Yahoo Español, Auyantepui, Teoma, Conexcol, Ambdirecto, Terra, and Ahijuna) as meta-searchers and the 314 category labels of the framework (from the previous step) as input queries. From our review (see Section 2.3), we found that these meta-searchers provided the richest Spanish resources on the Web. The following shows the pseudo code of the meta-searching algorithm:

0. Initiate an empty set named R to store all results. Let U be the set of all queries for meta-searching, let E be the set of all search engines chosen for meta-searching, let N be the number of top-ranked results to be retrieved as meta-search results.
1. While not all queries are used, select a new query Q from U
2. While not all search engines are used, select a new search engine S from E
3. Issue Q to S to obtain search results
4. Record the titles and URLs of the set of top-ranked (N) results returned from S
5. For each result in the set, check if it already appears in R. If yes, do nothing. Otherwise, add the result to R.
6. Return to Step 2
7. Return to Step 1
8. Return R as the set of all meta-searching results

In the meta-searching process, we used an automatic meta-search program to input these 314 queries to the seven search engines and collated the set of top ten results from each engine, with duplicate results removed. Through this process, we obtained 12,234 URLs of unique Web sites related to 296 category labels (non-empty nodes) out of the 314 nodes. The maximum depth of the resulting directory was 5.

5.3. Heuristic filtering: verification and enhancement

We developed a number of heuristic rules to filter and enhance the directories. The rationale for the development of these rules was two-fold. First, the addition or removal of directory items must be consistent with the topic of the directory. Second, a high level of relevancy of the directory items must be maintained. When building the SBiz directory, Web sites were removed if they were not relevant to the SBiz domain. Empty nodes were removed. Sub-topics of deleted nodes were removed as well. Web sites that contained too few links and pages (typically fewer than 10) were removed. Duplicated category labels were consolidated into one label. We have verified that these heuristics are comprehensive to ensure high quality and broad coverage of the directory items. The statistics of the completed SBiz Web directory are shown in Table 1. Two

Table 1
Summary statistics of the SBiz Web directory

Statistics	SBiz directory
Total number of categories	295
Total number of unique Web site URLs	4735
Average number of pages per category	16.1
Maximum depth	5

Table 2
Two fragments of the SBiz Web directory framework

SBiz directory framework fragment 1	SBiz directory framework fragment 2
Accounting	E-commerce
Associations	Associations
Business to business	By region
Firms	Conferences
News and media	Consultants
Tax negotiation and representation	Developers
Agriculture	Small business
Agricultural chemicals	Education and training
Tobacco farms	Employment
Aquaculture	News and media
Associations	Directories
Biologicals	Standards and protocols
Consulting	Strategy
Cooperatives	Technology vendors
Employment	Education and training
Farm and Ranch equipment	Business development
Farm real estate	Education
Horticulture	Masters and post doctorate
Import and export	Business schools
Industrial hemp	Finance
Livestock	Employment
News and media	By region
Trade shows	Careers
Automotive	Directories
Employment	Job search
Import and export	Energy and environment
Parts and accessories	Consulting
Electrical	Employment
Related services	Management
Retail	Marketing and advertising
Wholesale and distribution	News and media
Car insurance	Oil and gas
Insurance agents	Organizations
Insurance companies	Renewable
Insurance brokers	Hostels
Mutual funds	Hostel establishments
Workers insurance	Bars and cafes
Insurance lawyers	Restaurants
Car repair shops	Hotels
Biotechnology and pharmaceuticals	Tourist companies
Cancer research	Tourism companies
Consulting	Travel agencies
Employment	Online travel agencies
Pharmaceuticals	Travel agency franchises
Business services	
Auctions	
Virtual shopping centers	
Online auctions	
Public auctions	
AudioVisual	
Videoconferencing	
Communications	

fragments of the directory labels appear in Table 2. Fig. 2 shows screen shots of the directory.

6. Evaluation methodology

In this section, we describe our methodology to evaluate the usability of the directory developed by the framework. Our evaluation objectives were: (1) to study how the SBiz Web directory could assist human browsing of the SBiz domain on the Web; (2) to compare the SBiz Web directory with a benchmark directory on the Web in order to understand the effectiveness and efficiency of using the SBiz directory; and (3) to evaluate the user satisfaction achieved by using the SBiz directory.

To achieve objective (1), we invited human subjects to use the SBiz Web directory to browse the SBiz domain and record subjects' performance level and feedback. To achieve objective (2), we selected BIWE as the benchmark Web directory to compare against the SBiz Web directory. Compared with other Spanish Web directories, BIWE covered the most comprehensive resources in the SBiz domain on the Web (see Section 2.3). Despite BIWE's comprehensiveness, we believe that our framework can significantly enhance users' Web browsing experience by incorporating relevant resources and filtering out non-relevant ones. Therefore, the Web directory developed by using the framework should achieve significant (not only small) improvement over BIWE. The comparison was fair because, after applying our framework, the SBiz directory was significantly different than BIWE from which some of the directory labels were used in the SBiz. To achieve objective (3), we used a questionnaire to solicit subject ratings and comments after users explored the two directories.

6.1. Experimental design and tasks

We designed scenario-based browse tasks consistent with Text REtrieval Conference (TREC) standards (Voorhees and Harman, 1997) to evaluate the performance of our directories. For example, a scenario for testing SBiz directory was titled "America Online (AOL) in Latin America" and a browse task was "Find the URLs of financial portals where you can find stock quotes on America Online." In each task, the subject used the directory to find addresses (represented by URL links) of relevant Web sites or pages. To further validate the relevance of tasks, we did a pilot test with three subjects for each directory before conducting the actual experiment.

The subjects who voluntarily participated in the experiment were native Spanish speakers who could understand the content of the directories. We recruited 19 Spanish students as subjects from a major university in the United States to evaluate the browse performance of the SBiz Web directory. In the half-hour experiment, we introduced the two directories (the SBiz Web directory and the benchmark directory) to each subject and asked the subject to perform browse tasks using the SBiz Web directory (in one section)

or the benchmark directory (in another section). Each directory was randomly assigned to a scenario that consists of one browse task. The order in which the directories were used was randomly assigned to avoid bias due to their sequence.

After using a Web directory, the subject filled in a post-section questionnaire (see Appendix A) about his rating and comments on the directory. The experimenter recorded all verbal comments or behavioral observations that were later analyzed using protocol analysis (Ericsson and Simon, 1993), in which the verbal and written comments were categorized by their meanings. Upon finishing the study, the subject also filled in a post-study questionnaire (see Appendix B) to compare the two directories and to provide their demographic information, which was kept confidential in accordance with the Institutional Review Board Guidebook (Penslar, 2001).

6.2. Hypothesis testing

Because the SBiz Web directory developed by the proposed framework encompassed Web resources from different Spanish-speaking regions, we believed that they should provide richer content than that of the benchmark directory. Users could thus find relevant results more quickly from our directory. We therefore established the following three hypotheses about the effectiveness (H1), efficiency (H2), and usability (H3) of using the Web directories.

H1. The SBiz Web directory enables users to achieve higher effectiveness than the benchmark Web directory in performing browse tasks.

H2. The SBiz Web directory enables users to achieve higher efficiency than the benchmark Web directory in performing browse tasks.

H3. The SBiz Web directory achieves a higher user satisfaction than the benchmark directory.

As each subject was asked to perform similar tasks using the two directories, we used a one-factor repeated-measures design, which gives greater precision than designs that employ only between-subjects factors (Myers and Well, 1995).

6.3. Performance measure

We measured the efficiency of using a directory by recording the time the subject spent on each task. We also measured the effectiveness of using a directory by the following formulae:

$$\text{Precision} = \frac{\text{Number of relevant URLs identified by the subject}}{\text{Number of URLs identified by the subject}},$$

$$\text{Recall} = \frac{\text{Number of relevant URLs identified by the subject}}{\text{Number of relevant URLs identified by the expert}},$$

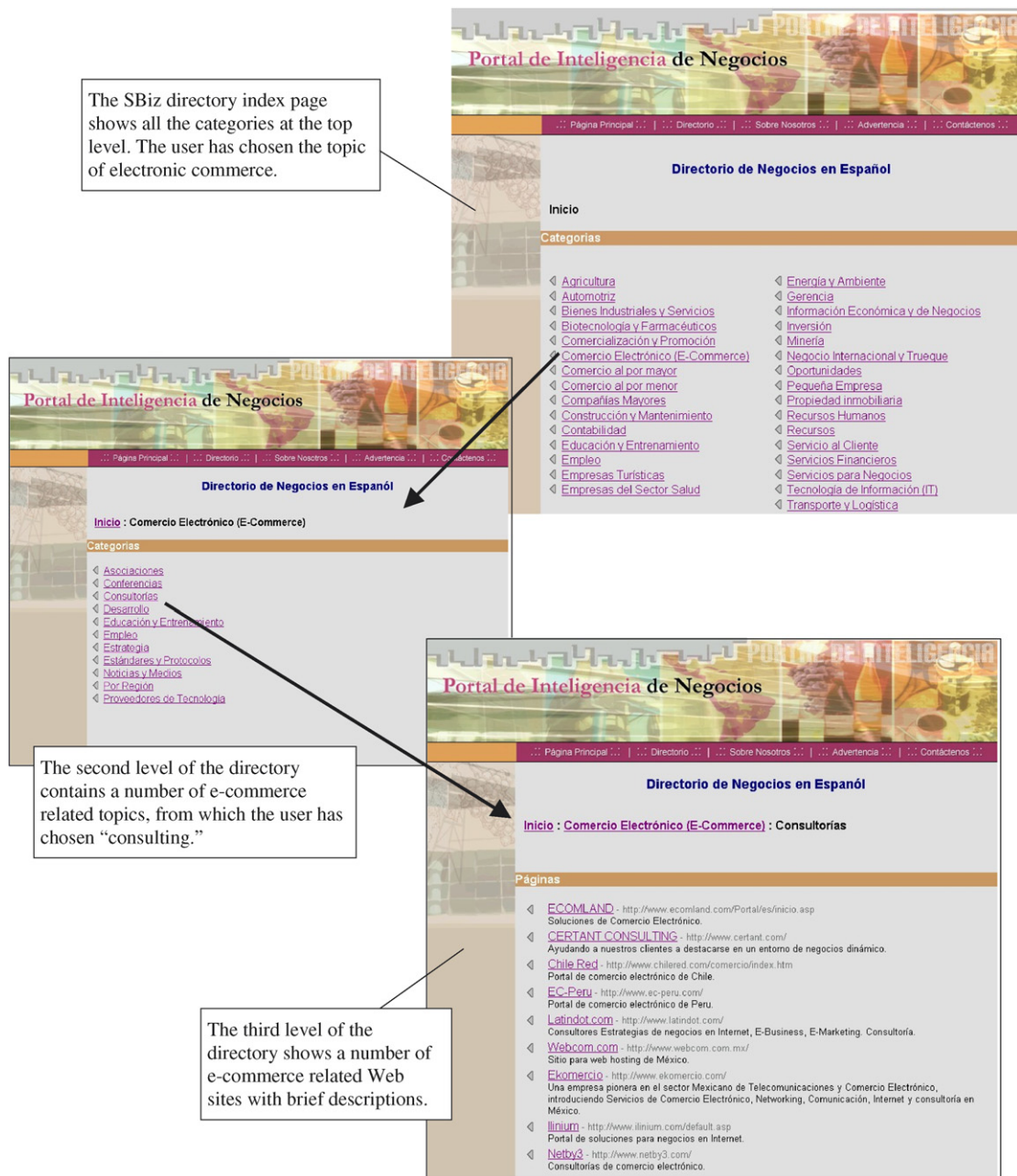


Fig. 2. Screen shots of the SBiz Web directory.

$$F\text{-value} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Precision reflected how well the directory helped users find relevant results and avoid non-relevant results. *Recall* reflected how well the directory helped users find all the relevant results. *F-value* was used to balance between recall and precision (Shaw et al., 1997), reflecting the performance achieved by the expert and subjects simultaneously. We recruited a SBiz expert to provide answers to evaluate subjects' performance in the tasks. This expert was a senior executive of a management consulting company in Mexico. Being a native Spanish speaker, he

had 24 years of experience in such areas as business development, raising capital, negotiations, finance and strategic planning. He also had worked as the Vice President of Business Development for the Gallup Organization in Mexico.

7. Experimental results and discussions

In this section, we report and discuss the results of our experiment on the Web directories. Tables 3 and 4, respectively, summarize the statistical results of hypothesis testing and subjects' demographic profiles.

Table 3
Statistical results of hypothesis testing

Hypothesis	Measure	SBiz directory		BIWE		<i>p</i> -Value	Result
		Mean	S.D.	Mean	S.D.		
H1	Precision	0.79	0.33	0.75	0.40	0.633	Partially supported
	Recall	0.29	0.14	0.17	0.13	0.002*	
	<i>F</i> -value	0.41	0.19	0.16	0.12	0.000*	
H2	Efficiency ^a	199	66	262	58	0.000*	Supported
H3	Satisfaction ^b	1.7	0.73	3.0	1.67	0.009*	Supported

* α error = 0.05, sample size = 19.

^aEfficiency was measured by the time (in seconds) used.

^bThe range of rating is from 1 to 7, with 1 being the best.

7.1. Effectiveness of the Web directories

We found that the SBiz Web directory achieved better precision, recall, and *F*-value than the benchmark directory. Pairwise *t*-tests show that the SBiz Web directory was significantly more effective than BIWE in terms of mean recall and *F*-value. We believe that the comprehensive coverage of the SBiz Web directory enabled it to provide a wider variety of Web resources for subjects to use. Collected by a comprehensive approach involving anchor directory boosting, meta-searching, and heuristic filtering, the resources came from not only major Spanish-speaking countries such as Spain and Mexico, but also some 20 Latin American regions. Subjects liked the fact that the SBiz Web directory provided relevant and precise information for their tasks. For example, subject #s6 said: “I found the information quickly and it was precise.” Subject #s10 said that “the information provided about the topic seems more related to it. (There are) more sites about the subject in search” and subject #12 said that he found “higher quality pages” using SBiz directory.

The relatively cleaner interface of the SBiz Web directory helped subjects perform the tasks more effectively and not to be distracted by advertisements or pop-up windows. Different from commercial directories, the SBiz Web directory does not have sidebar advertisements or many functionality tabs, thereby helping subjects to concentrate on their tasks and to find the information they wanted. For example, subject #s8 said “I found it (the SBiz directory) clear for browsing” and subject #s18 said that it was “easy to view retrieved data (with the SBiz directory).” We believe that SBiz directory’s interface provides a good example for existing Spanish Web directories to learn from.

However, we found no significant difference in the precision between the two Spanish Web directories ($p = 0.633$). This shows that both directories provided Web sites relevant to the subjects’ tasks and the SBiz directory did not outperform BIWE in identifying these relevant sites. We believe that the filtering process of SBiz directory needs to be improved to remove non-relevant items and to make relevant items more apparent. Never-

Table 4
Subjects’ demographic profiles

Demographic information	Subjects’ demographic profile (total: 19)
Country of origin	Mexico (12), USA (3), Panama (1), Puerto Rico (1), Colombia (1), Peru (1)
Education	Undergraduate (13), Bachelor earned (2), Master earned (3), Doctorate earned (1)
Age range	18–25 (14), 26–30 (2), 31–35 (2), 41–50 (1)
Gender	Female (10), male (9)
Hours of using computer per week	<5 (1), 5–10 (2), 10–15 (1), 15–20 (3), 20–25 (9), 30–35 (1), >35 (2)

theless, SBiz directory’s significantly better *F*-value and recall led us to conclude that H1 was *partially supported*.

7.2. Efficiency of the Web directories

We found that the SBiz Web directory achieved a significantly higher efficiency than the benchmark directory, *thereby supporting* H2. On average, a subject spent 3 min 19 s using SBiz directory to finish a task but a significantly longer time (4 min 22 s) using BIWE to complete a task. We believe that the comprehensive domain coverage and the clean user interfaces of the SBiz Web directory helped subjects to browse and to find the directory items more easily than the benchmark directories, as Spanish subject #s8 said: “I liked the (SBiz) directory domain, (which) is more easy to use than (the) other one (BIWE).” In addition, BIWE occasionally had pop-up windows showing advertisements to subjects. While these advertisements may be relevant to the topic of subjects’ tasks, some subjects did not like to see them or considered them hindering their work. For example, subject #s5 pointed out “pop-up windows” was a weakness of BIWE and subject #s3 said that “a lot of information (in BIWE) is not useful.” Another reason contributing to SBiz directory’s higher efficiency might be the domain analysis and extensive collection of the directory items, which ensured

Table 5
Analysis of subjects' comments on comparing the SBiz and BIWE directories

Comments on SBiz Web directory	Number of subjects expressing the meanings
Easy and nice to use	4
Useful and with relevant results	6
Good way to organization information	7
Not enough information for browsing	1
No comments/unclear comments	1
Comments on BIWE directory	Number of subjects expressing the meanings
Not useful/not good	5
Not clear for browsing/gives irrelevant results	6
Easy to navigate or browse	5
No comments/unclear comments	3

that relevant items had been included in the directory. Combining the advantages of automatic processes and human domain knowledge, the SBiz Web directory was able to provide a comprehensive coverage of the SBiz domain.

7.3. User ratings and comments on the Web directories

Subjects' ratings of their satisfaction on the SBiz Web directory were significantly better than those of the benchmark directory, thus confirming H3. In general, the subjects felt that the SBiz Web directory was helpful for browsing different topics in the SBiz domain. Table 5 summarizes the comments provided by the subjects in the post-study questionnaire comparing the two directories. We believe that the development of the directory had contributed to the higher usefulness in subjects' task performance.

The subjects' preferences on the SBiz Web directory were reflected in their verbal comments. Subjects who used the Spanish directories generally thought that the SBiz directory provided relevant content in a well-organized manner. Eight subjects said that the organization of the information in the SBiz directory was good. For example, subject #s1 said: "There are a lot of topics for browsing so as to find info on anything." Subject #s17 said that the SBiz directory has a "good color contrast" and was "easier to read." Six subjects said that SBiz directory was useful and provided relevant information. For example, subject #s7 said that SBiz directory "always gave relevant pages" and subject #s17 said that the "directory is very useful." Five subjects said that the directory was easy to view and browse. For example, subject #s2 said that SBiz directory was "easy to follow" and had "more accurate information."

On the other hand, the subjects' comments on the BIWE directory were mixed. Six subjects said that the directory was not clear and gave non-relevant results. For example,

Table 6
Subjects' preferences on different applications

Application	No. of subjects who preferred		
	SBiz	BIWE	Others
To search for high quality Spanish company information, I would use	19	0	0
To browse Spanish business-related Web resources, I would use	13	6	0
To achieve effective integration of information from different Spanish business sources, I would use	16	3	0
To analyze information about certain Spanish business issues, I would use	15	4	0

subject #s7 said that BIWE gave results from "other countries (that) I'm not interested." Subject #s9 said that using the directory was "time consuming at times." Five subjects said that the directory was not useful. For instance, subject #s12 said that BIWE provides "no real browsing support." Yet, five other subjects were satisfied with the BIWE directory. For example, subject #s18 said that it was "easy to navigate" and subject #s8 found it "clear for browsing." We believe that these mixed comments might have been caused by subjects' individual differences in Web browsing experience. Further research is needed to investigate this issue.

Subjects also expressed their preferences regarding the two Spanish Web directories they used for different applications. Table 6 shows that the subjects favored the SBiz Web directory on all listed applications. In particular, all subjects preferred using the SBiz Web directory to search for high-quality Spanish company information, probably due to its comprehensive coverage of Spanish company information. However, there were six subjects who preferred using BIWE to browse SBiz-related Web resources, perhaps because of BIWE's ability to search within its directory. In contrast, SBiz directory did not provide this function. These favorable preferences toward the SBiz Web directory demonstrated the usability of the framework to building these directories. Table 7 shows subjects' self-ratings on their knowledge and familiarity with Web searching. From these ratings, we believe that the subjects relied quite heavily on the Web to seek for information.

8. Conclusions and future directions

Building a high-quality Web directory without much expert knowledge and extensive human efforts has challenged developers of Web portals. As more users browse the Web in their native languages, better approaches to building Web directories in non-English

Table 7
Subjects' self-rating on their profiles

Profile	Mean	S.D.
"I am good at searching for information on the Web"	2.7	1.4
"I strongly rely on the Web to search for information"	2.0	1.4

Note: The rating was based on a 7-point Likert scale where "1" means "strongly agree."

languages are needed. In this paper, we have proposed a semi-automatic framework for constructing domain-specific Web directories. The framework, which consists of anchor directory boosting, meta-searching, and heuristic filtering, combined machine efficiency and human domain knowledge to assist in Web browsing and decision-making. Based on the framework, we developed the SBiz Web directory for the SBiz domain. Experimental results show that the SBiz Web directory significantly outperformed the benchmark Web directory in effectiveness, efficiency, and usability. Subjects provided favorable comments on the SBiz Web directory and preferred using it to search and browse for SBiz information. We therefore conclude that the proposed framework is highly usable and can support construction of high-quality domain-specific Web directories. This research thus contributes to developing a useful framework for organizing domain-specific information on the Web and to providing empirical findings and useful insights for end-users, system developers, and researchers of Web information seeking and knowledge management. The framework and its application to the SBiz domain are new. While previous work focused on Web searching in the English (Marshall et al., 2004) and non-English domains (Chung et al., 2004), this work addresses Web browsing—an area that has received relatively less attention from researchers.

This research was limited in a number of ways. We were limited by the scarce prior work on non-English Web browsing, which prevented a more comprehensive review on this topic that possibly would have offered better guidelines for developing the directories. As for the user study, the use of student subjects might have limited the external validity of the findings. Also, the numbers and diversity of Spanish subjects in the experiment were limited because we had difficulty recruiting more subjects. Future work may consider expanding the sample sizes and using professionals in the SBiz domain to establish a higher external validity for the experimental results. The relative lack of commercial Spanish Web directories (compared with English Web directories) also limited our choice of a benchmark directory in the experiment.

Our future directions include refining the SBiz Web directory and testing our framework in other domains and languages. We will investigate new techniques in constructing non-English Web directories. We will also consider more resources in different languages and address the specific needs of the online population using the languages.

Acknowledgments

This research was partly supported by fundings from National Science Foundation (NSF) Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999–March 2002, NSF Knowledge Discovery and Dissemination (KDD) Program #9983304 (June 2003–March 2004 and October 2003–March 2004), and The University of Texas at El Paso. We thank all the contributors of system development and user study, and the expert and human subjects in the experiment.

Appendix B

Post-Study Questionnaire

Participant Number: _____

For each of the following dimensions, please write 1, 2 or 3 to show your preference on the systems that you have used or would use, with 1 being the best.

Dimension	SBiz	BIWE	Others (please specify _____)
To search for high-quality Spanish business information, I would use			
To browse Spanish business-related Web resources, I would use			
To achieve effective integration of information from different Spanish business sources, I would use			
To analyze information about certain Spanish business issues, I would use			

Please compare SBizPort with BIWE and provide your comments below:

		BIWE
Usefulness for browsing		

I am good at searching for information on the Web

Strongly Agree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Strongly Disagree

How many years have you been working in the industries? _____ years

I strongly rely on the Web to search for information.

Strongly Agree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Strongly Disagree

Other CommentsDemographic Information

The country where you were born: _____

Is Spanish your primary language? _____ Yes _____ No

Number of hours spent on using computer per week (please check):

_____ Less than 5 hours

_____ 5 hours to less than 10hours

_____ 10 hours to less than 15 hours

_____ 15 hours to less than 20 hours

_____ 20 hours to less than 25 hours

_____ 25 hours to less than 30 hours

_____ 30hours to less than 35 hours

_____ Equal to or more than 35hours

Gender: M / F

Age range: 18-25____, 26-30____, 31-35____, 36-40____, 41-50____, 51-60____, 60 or above____

Education: Undergrad. Student____, Associate degree earned____, Bachelor earned____, Master earned____, Doctorate earned____

Thank you very much!

References

- Caramelli, P., 2003. The current and future rapid growth of older people in Latin America: implications in psychogeriatrics (keynote presentation). In: Proceedings of the Eleventh International Congress, International Psychogeriatric Association, Chicago, IL.
- Chen, H., Schuffels, C., Orwig, R., 1996. Internet categorization and search: a self-organizing approach. *Journal of Visual Communication and Image Representation* 7 (1), 88–102.
- Chuang, S.-L., Chien, L.-F., 2003. Enriching Web taxonomies through subject categorization of query terms from search engine logs. *Decision Support Systems* 35 (1), 113–127.
- Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T.-H., Chen, H., 2004. Internet searching and browsing in a multilingual world: an experiment on the Chinese Business Intelligence Portal (CBizPort). *Journal of the American Society for Information Science and Technology* 55 (9), 818–831.
- Chung, W., Chen, H., Nunamaker, J.F., 2005. A visual framework for knowledge discovery on the Web. *Journal of Management Information Systems* 21 (4), 57–84.
- CountryWatch Inc., 2006. [Online]. Available at <www.countrywatch.com/about/about.aspx>.
- Dewey, J., 1925. *Experience and Nature*. Open Court, Chicago.
- Ericsson, K.A., Simon, H.A., 1993. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA.
- Fellbaum, C., 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
- Gallup, 2002. Encuesta sobre portales 2002. [Online]. Available at <aui.es/estadi/gallup/gallup_portales_2002.htm>.
- Global Reach, 2004. *Global Internet Statistics (by Language)*. [Online]. Available at <www.glreach.com/globstats/>.
- Grüninger, M., Lee, J., 2002. Ontology: applications and design. *Communications of the ACM* 45 (2), 39–41.
- Holsapple, C.W., Joshi, K.D., 2002. A collaborative approach to ontology design. *Communications of the ACM* 45 (2), 42–47.
- Hovy, E., 2003. Using an ontology to simplify data access. *Communications of the ACM* 46 (1), 47–49.
- Kim, H.M., 2002. Predicting how ontologies for the semantic Web will evolve. *Communications of the ACM* 45 (2), 48–54.

- Kleinberg, J., 1999. Authoritative sources in a hyperlinked environment. *Journal of the Association of Computing Machinery* 46 (5), 604–632.
- Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., 2001. On semi-automated Web taxonomy construction. In: *Proceedings of the Fourth International Workshop on the Web and Databases*, Santa Barbara. ACM Press, CA, USA, pp. 91–96.
- Langridge, D.W., 1992. *Classification: Its Kinds, Elements, Systems, and Applications*. Bowker Saur, London.
- Librarians' Internet Index, 2006. About LII—Overview. [Online]. Available at <lii.org/pub/htdocs/about_overview.htm>.
- Marshall, B., McDonald, D., Chen, H., Chung, W., 2004. EBizPort: collecting and analyzing business intelligence information. *Journal of the American Society for Information Science and Technology* 55 (10), 873–891.
- Miniwatts International, 2006. Internet usage statistics—the big picture (updated on September 18, 2006). [Online]. Available at <www.internetworldstats.com/stats.htm>.
- Mowshowitz, A., Kawaguchi, A., 2002. Bias on the Web. *Communications of the ACM* 45 (9), 56–60.
- Myers, J., Well, A., 1995. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates Publishers, Hillsdale, NJ, USA.
- Office of NAFTA and Inter-American Affairs, 2004. North American Free Trade Agreement. [Online]. Available at <www.mac.doc.gov/nafta/>.
- Pease, A., Niles, I., Li, J., 2002. The Suggested upper merged ontology: a large ontology for the semantic Web and its applications. In: *Working Notes of the AAAI—2002 Workshop on Ontologies and the Semantic Web*, Edmonton, Canada.
- Penslar, R.L., 2001. *Institutional Review Board Guidebook*, Office for Human Research Protection, US Department of Health and Human Services. [Online]. Available at <ohrp.osophs.dhhs.gov/irb/irb_guidebook.htm>.
- Sato, S., Sato, M., 1999. Automatic generation of Web directories for specific categories. In: *Proceedings of the AAAI Workshop on Intelligent Information Systems*. AAAI Press, Orlando, FL.
- Shaw, W.M.J., Burgin, R., Howell, P., 1997. Performance standards and evaluations in information retrieval test collections: cluster-based retrieval models. *Information Processing and Management* 33 (1), 1–14.
- Sí Spain, 2006. The Spanish Language Worldwide. [Online]. Available at <www.sispain.org/english/language/worldwid.html>.
- Sowa, J.F., 2000. *Ontology*. In: *Knowledge Representation: Logical, Philosophical, and Computation Foundations*. Thomson Learning, pp. 51–131.
- Stamou, S., Krikos, V., Kokosis, P., Ntoulas, A., Christodoulakis, D., 2005. Web directory construction using lexical chains. In: *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*, Springer, Alicante, Spain.
- The Office of the United States Trade Representative (USTR), 2006. Central America-Dominican Republic Free Trade Agreement [Online]. Available at <www.ustr.gov/Trade_Agreements/Bilateral/CAFTA/Section_Index.html>.
- Voorhees, E., Harman, D., 1997. Overview of the Sixth Text Retrieval Conference (TREC-6). In: *NIST Special Publication 500-240: The Sixth Text Retrieval Conference (TREC-6)*. National Institute of Standards and Technology, Gaithersburg, MD, USA.
- Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S., 2001. *Ontology-based integration of information: a survey of existing approaches*. In: *The IJCAI 2001 Workshop on Ontologies and Information Sharing*, Seattle, WA.
- Yang, H.-C., Lee, C.-H., 2003. A text mining approach on automatic generation of Web directories and hierarchies. In: *Proceedings of the IEEE/WIC International Conference on Web Intelligence*. IEEE Computer Society, Halifax, Canada.