Building a Directory for the Underdeveloped Web: An Experiment on the Arabic Medical Web Directory

Wingyan Chung[1] and Hsinchun Chen[2]

[1] Department of Operations and Management Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
[2] Department of Management Information Systems, Eller College of Management, The University of Arizona, AZ 85721, USA
wchung@scu.edu, hchen@eller.arizona.edu

Abstract. Despite significant growth of the Web in recent years, some portions of the Web remain largely underdeveloped, as shown in a lack of high quality content and functionality. An example is the Arabic Web, in which a lack of well-structured Web directories has limited users' ability to browse for Arabic resources. In this research, we proposed an approach to building Web directories for the underdeveloped Web and developed a proof-of-concept prototype called Arabic Medical (AMed) Web Directory that supports browsing of over 5,000 Arabic medical Web sites and pages organized in a hierarchical structure. We conducted an experiment involving Arab subjects and found that AMed directory significantly outperformed a benchmark Arabic Web directory in terms of browsing effectiveness and user ratings. This research thus contributes to developing a useful Web directory for organizing information of the Arabic medical domain and to better understanding of supporting browsing on the underdeveloped Web.

Keywords: Browsing, information seeking, Arabic, medical domain, meta-searching, Web directory, underdeveloped Web, user study, experiment.

1 Introduction

Internet usage has been growing rapidly in recent years, especially in many developing countries and regions where more and more people are getting access to the Internet. For example, between 2000 and 2007, the online populations of the Middle East grew by 491.4% [1]. However, the functionality and quality of content in the Web sites of these regions often lack behind the growth of their user base. Users are challenged to browse over a large number of Web sites without well-designed Web directories. Here we define the "underdeveloped Web" as the portion of the World Wide Web in which the growth of its usage far outpaces the growths of its content and functionality, thus users' needs for Web resources are largely not satisfied. An example of the underdeveloped Web is the Arabic Web, which consists of Web sites from the Middle-Eastern regions and is used by over 19 million Arabic-speaking people. Despite the rapid growth in Arabic Web usage, a lack of well-structured Web directories has limited users' ability to browse the Arabic Web. Although some search engines

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 468–477, 2007.

support searching in Arabic, the relatively little content of the Arabic Web (compared with other languages such as English) makes it difficult for search engines to provide a comprehensive indexing of Arabic content. The unique characteristics of the Arabic language further complicate the problems.

To address the needs, we developed an approach to organizing information on the underdeveloped Web by using a combination of manual and automatic methods. Based on the approach, we developed the Arabic Medical (AMed) Web Directory as a proof-of-concept prototype that facilitates browsing of Arabic medical information on the Web. To understand the usability and effectiveness of the AMed Web directory, we conducted an experiment involving Arab subjects to compare AMed directory with a benchmark Web directory. In the following, we review prior work on Web directory development, describe our work on developing the AMed Web Directory to address the needs, report findings of an experiment on studying the effectiveness and usability of the directory, and finally conclude our work and outline future directions.

2 Literature Review

Research into Web searching and browsing on the non-English Web has been growing in recent years [e.g., 2, 3, 4]. While many efforts are related to Web searching, relatively little attention has been paid on Web browsing, an activity that users often perform when seeking information. Web directory is often the starting point of users' Web browsing and different approaches are used to develop Web directories. In this section, we review previous research on browsing and on Web directory development. We also review Web resources in Arabic, the language we chose to study browsing on the underdeveloped Web.

2.1 Web Browsing

Browsing is a major activity that users frequently engage when seeking information on the Web. It has been defined as an exploratory information seeking process characterized by the absence of planning, with a view to forming a mental model of the content being browsed [5]. In exploratory browsing, a user first transforms his general information need into a problem. He then articulates his needs as search terms or hyperlinks that appear on the system interface, searches using those terms or explores hyperlinks using browse supports such as Web directories, and finally evaluates the results by scanning through them. As the Internet evolves as a major information-seeking platform for many developing regions, supporting Web browsing by using Web directories have become increasingly important. Consequently, the development of these directories has drawn attention from researchers and practitioners.

2.2 Web Directory Development

Previous work in developing Web directories falls into two categories: (1) Extensive manual identification and categorization of Web resources; and (2) Automatic construction of directories using machine learning or Web mining techniques. We review the work done in these two categories below.

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 Directory Mozilla (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. 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 online population. Currently, DMOZ powers the core directory services of many search engines, including Netscape Search, AOL Search, Google, Lycos, HotBot, and DirectHit. There are several other Web directories developed by paid or volunteer editors. The Yahoo! Directory (http://dir.yahoo.com/) is built and maintained by a team of paid editors who organize Web sites into categories and subcategories. 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. The UMLS Semantic network is one of three UMLS Knowledge Sources being developed by the National Library of Medicine (http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html). Based on 134 semantic types and 54 links, the network provides a categorization of all concepts represented in the UMLS Metathesaurus and represents important relationships in the biomedical domain. The construction of the above-mentioned Web directories relies heavily on expert participation and their domain knowledge. Moreover, the quality of the directory constructed by this method depends highly on the volunteer editors' domain knowledge, which usually varies from person to person.

Beside manual methods, automatic approaches to constructing directory and ontology have been proposed in previous research. For example, Sato and Sato developed an automated editing system that generates a Web directory from a given category word without human intervention [6]. Chuang and Chien propose a query-categorization approach to facilitate the construction of Web directories [7]. 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 [8]. Chen and colleagues proposed a self-organizing approach to Internet search and categorization [9]. Stamou et al. developed an approach to automatically assigning Web pages to a directory framework based on the linguistic information on the Web textual data [10]. In general, these automatic approaches yielded high efficiency at the expense of accuracy. The typically unsatisfactory categorization accuracy shows the deficiency of these approaches in constructing Web directories, especially on the underdeveloped Web where the usage is growing rapidly nowadays.

2.3 Arabic Web Resources

The Arabic Web serves as a good example of the underdeveloped Web. Arabic is spoken by more than 284 million people in about 22 Middle-Eastern and North African countries. Although Arabic is the fifth most frequently spoken language in the world, the Arabic Web is still in its infancy, constituting less than 1% of the total Web content and having a low 2.2% penetration rate [11]. The cross-regional use of Arabic and the exponential growth of Arabic Web [12] nevertheless have highlighted the

necessity of supporting better Web browsing. Here we review several Arabic Web portals to understand their support on Web browsing.

Ajeeb (http://www.ajeeb.com/) is a bilingual Web portal (English/Arabic) launched in 2000 by Sakhr Software Company. Its database contains over one million searchable Arabic Web pages, which can be translated to English using the online version of Sakhr's machine-translation software. In addition, Ajeeb has a multilingual dictionary and is known for its large Web directory, "Dalil Ajeeb," which the company claims is the world's largest online Arabic directory. Albawaba (http://www.albawaba.com/) is a consumer portal offering comprehensive services including news, sports, entertainment, e-mail, and online chatting. The portal supports searching for both Arabic and English pages and the results are classified according to language and relevancy. Albawaba also provides metasearching of other search engines (Google, Yahoo, Excite, Alltheweb, Dogpile) and a comprehensive directory of all Arab countries. Launched in 2000, UAE-based Albahhar (http://www.albahhar.com/) provides a wide range of online services such as searching, news, online chatting, and entertainment. The portal searches its 1.25 million Arabic Web pages and provides Arabic speakers a wide range of other online services like news, chat, and entertainment. Based in New Hampshire, Ayna (http://www.ayna.com/) is a Web portal providing an Arabic Web directory, an Arabic search engine, and other services such as a bilingual (English/Arabic) email system, chat, greeting cards, personal homepage hosting, and personal commercial classifieds. In July 2001, Ayna had over 700,000 registered users and provided access to more than 25 million pages per month. Due to Ayna's popularity, Alexa Research ranks it among the top three leading Web sites in the Arab World.

3 The Arabic Medical Web Directory

As shown in the review, previous approaches to building information directory 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 underdeveloped 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. On the underdeveloped Web (such as the Arabic Web), a lack of comprehensive Web directory further aggravate the problems. To our knowledge, no previous attempt has been made to develop an approach to developing Web directories for the underdeveloped Web.

To address the research gaps, we have developed a generic approach to facilitating Web directory development for the underdeveloped Web. Our approach 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. Searching multiple high-quality search engines, which has been shown to provide higher quality of the results than relying on only a few search engines [13], and manual filtering of results are the main components. In the following,

Table 1. Summary statistics of the Arabic Medical Web Directory

Statistics	AMed Directory
Total number of categories	232
Total number of Web pages	5,107
Average number of pages per categories	22.1
Maximum depth	5

Fig. 1. Screen shots of AMed Web Directory

we explain our approach in the context of building the Arabic Medical Intelligence (AMed) Web directories. We chose the Arabic Web as our research test bed because of its high growth rate yet limited development in its resources. Providing timely and accurate Web directory of a domain that most Arabic people concern is very important; and the Arabic medical domain serves as a good example in this regard.

In the first step, we identified an existing Web directory as the base directory and modified its category labels as queries for meta-searching. We used the DMOZ directory as the base directory because of its comprehensiveness in the English medical domain. We removed 46 nodes from the original 356 nodes of its medical subdirectory, leaving 310 nodes in the directory. Then, 11 nodes were manually added by including cultural specific items such as Islamic medicine, resulting in a 321-node Arabic medical directory framework.

In the second step, we filled in the directory framework (obtained from the first step) with items obtained by automatic meta-searching. 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. To fill in the Arabic medical directory, we used six major search engines (Ba7th, Arabmedmag, Google, Ayna, Sehha, and ArabVista) as meta-searchers and the 321 category labels of the framework (from step 1) as input queries. To our knowledge, these meta-searchers provide the richest Arabic resources on the Web. After running an automatic meta-search program adaptated to the Arabic language, we obtained 8,040 unique URLs related to 292 category labels (non-empty nodes) out of the 321 nodes. The maximum depth of the resulting directory was 5.

In the third step, we manually filtered out non-relevant items and added necessary items. We followed a number of heuristics to filter and enhance the directories. URLs were removed if they were not relevant to the topic or they were not related to the domain (i.e., Arabic medical domain) being considered. 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. The statistics of the two resulting directories are shown in Table 1. Figure 1 show screen shots of the AMed Web directory.

4 Experimental Results and Discussion

In this section, we describe an experiment to evaluate the usability of the AMed Web directory and report the experimental findings. We selected Albawaba (http://www.albawaba.com/) as the benchmark directory to compare against AMed directory because of Albawaba's high stability and reliability. To our knowledge, Albawaba's medical Web directory provides the most comprehensive listing of medical topics and Web sites in Arabic. We designed scenario-based browse tasks consistent with Text Retrieval Conference standards [14] to evaluate the performance of the two directories. For example, a browse task related to prevention and treatment of cancer was: "Find articles about healthy diet and cancer prevention." In each task, the subject used the directory to find addresses (represented by URL links) of relevant Web sites or pages.

The subjects who voluntarily participated in the experiment were native Arabic speakers who could understand the content of the Web directories. Seven subjects participated in the experiment and each subject performed five different browse tasks using each of the two directories, making a sample size of $n = 35$ (5×7). In the half-hour experiment, we introduced the two directories (our directory and the benchmark directory) to each subject. Each subject worked on five tasks in the first section (using one directory) and five different tasks in the second section (using another directory), thus preventing learning effect in their performance. The order in which the directories were used was randomly assigned to avoid bias due to their sequence. We also randomly assigned the two sets of five tasks to evaluate the two directories. Because each subject used the two directories, a repeated-measure design was used in the experiment.

After using a Web directory, the subject filled in a post-section questionnaire about his rating and comments on the directory. The experimenter recorded all verbal comments or behavioral observations that were later analyzed using protocol analysis [15]. Upon finishing the study, the subject also filled in a post-study questionnaire to compare the two directories and to provide their demographic information, which was kept confidential in accordance with the Institutional Review Board Guidebook [16].

We measured the effectiveness of using a directory by precision, recall, and F value as shown in the formulas below. Serving in this research as a medical expert, a final-year Arabic medical student in a Middle East country graded the subjects' answers to calculate the values of the measures. The expert also verified all the experimental tasks to be appropriate for the experiment.

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

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

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

The hypotheses we tested are as follows:

H1: The AMed Web Directory achieves a significantly higher precision than the benchmark Web directory.
H2: The AMed Web Directory achieves a significantly higher recall than the benchmark Web directory.
H3: The AMed Web Directory achieves a significantly higher F value than the benchmark Web directory.
H4: The AMed Web Directory achieves a significantly better rating on its helpfulness to browsing than the benchmark Web directory.
H5: The AMed Web Directory achieves a significantly better rating on user satisfaction than the benchmark Web directory.

Table 2 summarizes the results of testing the five hypotheses. The AMed Web Directory (AMedDir) achieved a significantly better effectiveness than Albawaba, as shown in the significantly higher precision, recall, and F value. We believe that these favorable results were due to AMedDir's well-organized information hierarchy and comprehensive coverage of Arabic medical resources on the Web. The approach we used to create the AMedDir has incorporated the preciseness of manual methods and the efficiency of automatic methods, thus giving the benefits of high effectiveness and broad content coverage. Subjects were able to obtain answers to their tasks from AMedDir effectively. Therefore, hypotheses H1, H2, and H3 were supported.

In addition to the favorable results on effectiveness, AMedDir achieved significantly better ratings on helpfulness to browsing and on user satisfaction than Albawaba. On a 7-point Likert scale where "1" refers to the most favorable option, AMedDir obtained on average 2.14 on helpfulness to browsing and 2.43 on user satisfaction, while Albawaba obtained ratings of 5.57 and 5.86 respectively. We believe that the wide margins between the two directories' ratings were due to AMedDir's high effectiveness in supporting subjects' task performance and relatively cleaner

interface, and Albawaba's problems in browsing support. Some subjects complained that Albawaba's collection was too small and had inadequate organization of information, thus providing insufficient results or non-relevant results. Based on the significant differences in subjects' ratings, we conclude that hypotheses H4 and H5 were supported.

Table 2. Results of hypothesis testing

Hypothesis	AMedDir		Albawaba		*p*-value	Result
	Mean	SD	Mean	SD		
H1: Precision	**0.74**	0.44	0.09	0.28	0.000	Supported
H2: Recall	**0.32**	0.29	0.06	0.20	0.000	Supported
H3: F value	**0.43**	0.30	0.07	0.23	0.000	Supported
H4: Helpfulness to browsing*	**2.10**	1.07	5.57	0.54	0.000	Supported
H5: Satisfaction*	**2.43**	1.27	5.86	0.69	0.000	Supported

* The subject rating was based on a 7-point Likert scale, where "1" refers to the most favorable option.

The favorable experimental results bring about several implications. First, the development of AMedDir not only helped support browsing, but also demonstrated an effective way to improve organization of vast and growing amounts of information on the underdeveloped Web. The hierarchical structure helped organize information effectively, facilitating concept categorization and topic classification. Second, the experimental findings revealed the need for better Arabic Web directories for browsing. While the Arabic Web is growing much faster than other parts of the Web, Arabic users are not getting the same level of services that many non-Arabic Web users currently enjoy. In our review, we found that many Arabic portals are unstable and contain limited information. This study highlights the need for researchers and practitioners to enrich the Arabic Web with better content and functionality. Third, the review and findings from this research contribute to the growing body of research on non-English Web browsing, which becomes increasingly important as it demonstrates a strong potential of growth in the coming decades. Rapidly emerging issues such as browsing Web directories, information organization, and directory development were addressed in this research.

5 Conclusions

While the developments of many parts of the World Wide Web have matured over the past decade, there are still some portions of the Web that are largely underdeveloped. This underdeveloped Web is often characterized by a lack of high quality content and functionality, despite having a rapid growth in usage. These problems reveal the needs for better organization and presentation of information on many Web sites of the underdeveloped Web. In this research, we developed an approach to building Web directories for the underdeveloped Web and used the approach to construct the Arabic Medical Web Directory (AMedDir) that supports browsing the Arabic medical

domain. Findings from an experiment involving Arab subjects show that AMedDir significantly outperformed a benchmark Web directory in terms of effectiveness and subject ratings. We conclude that the Arabic Medical Web Directory has high effectiveness and usability for supporting browsing the Arabic Web. This research thus contributes to developing a proof-of-concept prototype for organizing information of the Arabic medical domain and to better understanding of supporting browsing in the underdeveloped Web.

The research was limited in a number of ways. First, the AMed directory was a research prototype and hence lacked the scalability of commercial Web portals. Second, scarce literature on browsing the underdeveloped Web made it difficult for us to perform a more comprehensive review of the field. Third, we were limited by our ability to recruit more Arab subjects for the experiment.

Future directions of this research include developing Web directories for other parts of the underdeveloped Web and studying their impacts on browsing. For example, Spanish is the language of many Hispanic communities and Latin American countries whose populations are growing rapidly. Yet many Web sites in Spanish still lack browsing support and information and can be improved by introducing better Web directories or other functionality. Another direction is to develop automatic techniques to generate Web directory framework and to collect information on the Web. A third direction is to enhance human filtering and classification process to increase the precision and accuracy. These efforts will yield more useful and effective Web directories and enhance the browsing experience of many people in the world.

Acknowledgements

This research was partly supported by NSF Knowledge Discovery and Dissemination (KDD) program #9983304 (June 2003 – March 2004 and October 2003 – March 2004) and Santa Clara University. We thank the subjects and expert who participated in the experiment and the members of the system development team.

References

[1] Miniwatts International.: Internet Usage Statistics - The Big Picture (2007), http://www.internetworldstats.com/stats.htm

[2] Chung, W.: Studying information seeking in the non-English Web: An experiment on a Spanish business Web portal. International Journal of Human-Computer Studies 64, 811–829 (2006)

[3] Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T.-H., Chen, H.: 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, 818–831 (2004)

[4] Spink, A., Ozmutlu, S., Ozmutlu, H.C., Jansen, B.J.: U.S. versus European Web Searching Trends. SIGIR Forum 36 (2002)

[5] Chung, W., Chen, H., Nunamaker, J.F.: A visual framework for knowledge discovery on the Web. Journal of Management Information Systems 21, 57–84 (2005)

[6] Sato, S., Sato, M.: Automatic generation of Web directories for specific categories. In: Proceedings of the AAAI Workshop on Intelligent Information Systems, Orlando, FL (1999)
[7] Chuang, S.-L., Chien, L.-F.: Enriching Web taxonomies through subject categorization of query terms from search engine logs. Decision Support Systems 35, 113–127 (2003)
[8] Yang, H.-C., Lee, C.-H.: A text mining approach on automatic generation of Web directories and hierarchies. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence, Halifax, Canada (2003)
[9] Chen, H., Schuffels, C., Orwig, R.: Internet categorization and search: a self-organizing approach. Journal of Visual Communication and Image Representation 7, 88–102 (1996)
[10] Stamou, S., Krikos, V., Kokosis, P., Ntoulas, A., Christodoulakis, D.: Web directory construction using lexical chains. In: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (2005)
[11] Abbi, R.: Internet in the Arab World, UNESCO Observatory on the Information Society, p. 3 (2002)
[12] Norton, L.: The Expanding Universe: Internet Adoption in the Arab Region. World Markets Research Centre, Report (2001)
[13] Mowshowitz, A., Kawaguchi, A.: Bias on the Web. Communications of the ACM 45, 56–60 (2002)
[14] Voorhees, E., Harman, D.: Overview of the Sixth Text Retrieval Conference (TREC-6), In: NIST Special Publication 500-240: The Sixth Text Retrieval Conference (TREC-6), Gaithersburg, MD, USA (1997)
[15] Ericsson, K.A., Simon, H.A.: Protocol analysis: verbal reports as data. MIT Press, Cambridge (1993)
[16] Penslar, R.L.: Institutional Review Board Guidebook,Office for Human Research Protection, U.S. Department of Health and Human Services (2006), http://www.hhs.gov/ohrp/irb/irb_guidebook.htm