DGPort: A Web Portal for Digital Government

Chun Q. Yin, L. Dwayne Nickels, Charles Zhi-kai Chen, T. Gavin Ng, Hsinchun Chen
Artificial Intelligence Lab
The University of Arizona
Tucson, Arizona 85721, USA
{cyin, nickels, czhikai, tgn, hchen}@bpa.arizona.edu

Abstract

This paper provides a summary of the initial development of a Web portal for the digital government domain. Information retrieval techniques commonly used to find information on the Internet are discussed along with the problems associated with these techniques that led to the development of the Digital Government Web portal (DGPort). We also discuss the advantages that DGPort could have for researchers in the digital government domain as well as the value-added features that this portal provides. Future evaluation plans for the portal are also described.

1. Introduction

Over the past several years there has been tremendous growth in areas related to digital government. The success of digital government researchers depends on their being current on research results and industrial trends. Knowledge of what others may be doing helps researchers avoid wasteful duplication and obtain hints for improving their own research techniques. This research project aims to explore technical, educational, and social issues relevant to supporting intelligent Web searching in this domain and to develop a Web portal to provide a one-stop search service for researchers and practitioners.

2. Research Motivation

Information regarding research results and industrial trends is critical to the success of digital government researchers. While the Web provides a convenient means for information searching, the discipline is facing the problem of information overload (Bowman et al., 1994), in which a search on a general-purpose search engine such as Google (www.google.com) results in thousands of hits.

In addition, the discipline encompasses a diversity of research perspectives and application areas. This has resulted in terminology/vocabulary differences (Furnas et al., 1987, Chen, 1994). Researchers in different disciplines use different terminologies to describe their findings, and user search terms may be different from the indexing terminologies used in databases.

Information sources and quality on the Web are equally diverse (Lawrence & Giles, 1999). Information useful to digital government researchers can include scientific papers, journals, technical reports, and Web pages of varied quality. As a result, researchers need to go to multiple information sources and conduct significant manual analysis to identify quality, time-critical information. Thus, the lack of an integrated, "one-stop shopping" service is an urgent problem to resolve.

3. Existing Approaches

Several Web-searching approaches to retrieving information have been adopted in academia and industry:

3.1 General-purpose search engines and portals: Most prevailing general-purpose search engines, such as Yahoo or Google, are keyword based (Arasu et al., 2001). Each has its own characteristics and

employs its preferred algorithm in indexing, ranking and visualization of Web documents. Their search speeds are fast; however, their results are often overwhelming and not precise.

3.2 Meta search engines: Empirical studies show that every search service returns a different set of documents for the same query. Meta search engines such as Copernic, MegaSpider, and MetaCrawler "search the search engines" (Selberg and Etzioni, 1995). Although the information returned is comprehensive, the problem of information overload worsens.

3.3 Post-retrieval analysis: Some search engines attempt to alleviate the information overload problem by performing post-retrieval analysis where documents returned are classified into different subject folders previously created by subject experts. The quality of post-retrieval analysis is often directly derived from the quality of the folders. For fine-grained, time-critical searches that do not have suitable topics in the taxonomy, retrieval results can be poor (Zamir and Etzioni, 1999).

4. Domain Analysis

The realm of digital government is concerned with how governments use information technology, specifically the Internet, to provide services to citizens, businesses and government agencies (Chen, 2002). The digital government domain can be traced back to the early 1990s. Ever since then government has done much to leverage information technology to deploy e-government services; however, much work remains before the vision of e-government can be fully realized (National Academies Press, 2002).

We began our initial analysis by collecting major topics within the domain through the latest publications, journals, papers, and research projects. The compiled list of topics runs across diverse fields. The researchers' interests also vary to a large extent. Because such information is scattered over countless places on the Internet, online searching typically yields a huge collection of results but with low precision. Thus one of our goals is to build a search portal that can fill the gap between existing approaches and researchers' needs.

Surveys were used to assist us in further refining our domain. They were sent out to researchers in this field to collect information such as valuable academia database resources, funding resources, etc. Interviews and conference meetings with domain experts were conducted. This gave us an opportunity to ask questions, solve ambiguities, and correct misconceptions. Reviewing recent publications and related papers helped us get a sense of what the researchers are doing and where the research is heading in the near future.

5. DGPort System Architecture and Features

DGPort follows typical search engine architecture, which relies on vertical spiders to traverse the Web to collect pages by following hyperlinks. The main technologies used include HTML, Java, Java Server Page (JSP), JavaBeans and Java Servlet (Chau et al., 2002). The pages collected are then processed by the Indexer to generate index files. Information in the index files is loaded into a MS SQL server database.

DGPort serves dynamic content via JSP pages. Search queries from the User Interface are passed to the Middleware, which generates corresponding SQL statements through a JavaBeans file to retrieve relevant Web pages from the database. Afterward the results are returned to the User interface, users can perform further analysis on the retrieval set, using post-retrieval analysis features, to find more precise results. Table 1 illustrates the DGPort architecture, which consists of the following features:

5.1 **Vertical Searching:** The AI Lab SpidersRUs Digital Library Toolkit (ai.bpa.arizona.edu) is used to collect the content. We customized the spider by equipping it with heuristics from domain experts to filter out irrelevant and/or poor-quality content. After collecting, the spider indexes Web content

systematically. We used 4000 URL as seeds and aimed to collect one million high-quality Web pages in our fully developed system.

5.2 **Meta-searching:** We incorporated meta searching ability in our service by connecting with different search engines and combining the search results. These information sources include online journals - FirstGov, GovTech; presses and news - CNN, BBC; and general-purpose search engines - Yahoo, Alta Vista. The meta searching offers researchers more extensive information.

5.3 **Noun Phrasing:** Browsing the ranked results from general-purpose search is time-consuming and mentally exhausting. The Arizona Noun Phraser extracts the key phrases from the search results of a session (Tolle & Chen, 2000). These phrases are used to create a folder structure that categorizes search results based on key concepts, making it much easier for users to grasp the results.

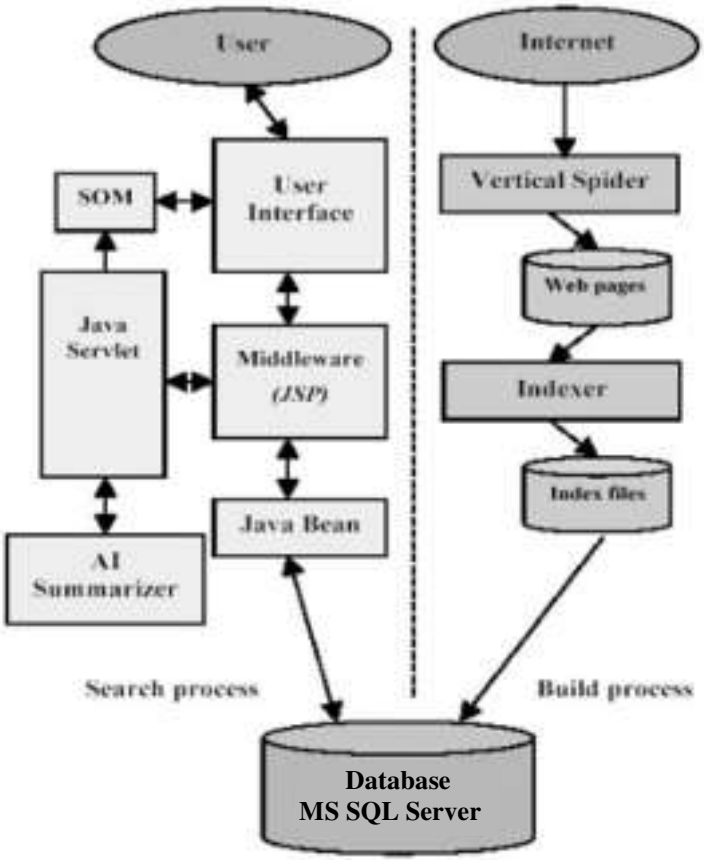

Figure 1: DGPort Architecture

5.4 **Concept Space:** Based on selective noun phrasing and co-occurrence analysis techniques, we adopted a text-mining approach to extract terminologies and their weighted relationships in creating a large, precise automatic thesaurus: Concept Space. Using Concept Space, researchers are able to find more relevant documents with less time and effort (Leroy & Chen, 2001).

5.5 **Self-organizing Maps (SOM):** SOM, a learning method that clusters objects with multi-dimensional attributes into a lower–dimension space, is applied to generate a 2D topic map representing a key concept in the set of documents. More important concepts dominate larger regions and similar concepts are grouped in a neighborhood. The documents are categorized into different regions for easier concept identification.

5.6 **Automatic Summarization:** The AI summarizer is also employed, which can summarize any Web page from the search result into a few sentences. The summarized sentences are highlighted in the location of their original search result pages. Such summarization can help the user decide quickly whether a page is interesting or not.

6. Conclusion and Future Direction

In this paper we discuss our experience in developing DGPort and present the challenges associated. We report on the research and development work that we have done for the portal thus far and describe the search tools that the portal contains. We also discuss the benefits that DGPort has over traditional search engines for researchers in the field of Digital Government.

Our future research plan includes an evaluation of the portal in two parts. The first part will consist of assessing search, collection and analysis capabilities of the portal. This evaluation will be based mainly on precision and recall measures. Lab experiments will be conducted to compare DGPort to other popular search tools. The users will be encouraged to comment on the usability of the portal, most useful features of the portal, and possible areas for improving the portal. The second part of the evaluation will consist of assessing the portal's educational and social impact. Project evaluation activities will be developed for formative and summative evaluation. Formative evaluation will seek to determine how the search portal can enhance collaborative and instructional activities in the domain of Digital Government. Summative evaluation will focus on key project goals, including the effectiveness of the portal for educators and students, overall usability and value of the portal, and increased accessibility to Web-based resources.

Acknowledgements

We would like to acknowledge the many digital government researchers who assisted us in our research efforts. We would like to especially thank Dr. Chris Demchak of the University of Arizona and Dr. Sharon Dawes of the Center for Technology in Government for their contributions to this project. This project was supported by the National Science Foundation Digital Government Research Program under grant No. 0302353.

References:

- A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan (2001). "Searching the Web," *ACM Transactions on Internet Technology*, 1(1), 2-43.
- C. M. Bowman, P.B. Danzig, U. Manber, and F. Schwartz (1994). "Scalable Internet Resource Discovery: Research Problems and Approaches," *Communications of the ACM*, 37(8): 98-107.
- Michael Chau, H. Chen, Jialun Qin, Yilu Zhou, Yi Qin, Wai-Ki Sung, Daniel McDonald (2002). "Comparison of Two Approaches to Building a Vertical Search Tool: A Case Study in the Nanotechnology Domain," Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02), Portland, Oregon.
- H. Chen (1994). "Collaborative Systems: Solving the Vocabulary Problem," *IEEE Computer*, Special Issue on Computer-Supported Cooperative Work (CSCW), 27(5): 58-66.
- H. Chen (2002). "Digital Government: technologies and practices," *Decision Support Systems*, 34: 223-227.
- G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais (1987). "The Vocabulary Problem in Human-System Communications," *Communications of the ACM*, 30(11), 964-971.
- S. Lawrence and C. Giles (1999). "Accessibility of Information on the Web," *Nature*, 400, 107-109.
- G. Leroy and H. Chen (2001), "Meeting Medical Terminology Needs: The Ontology-Enhanced Medical Concept Mapper," *IEEE Transactions on Information Technology in Biomedicine*, 5(4), 261 – 270.
- National Academies Press (2002). "Information Technology Research, Innovation, and E-Government," available at http://www.nap.edu/catalog/10355.html.
- E. Selberg and O. Etzioni (1995). "Multi-service Search and Comparison using the MetaCrawler," *Proceedings of the 4th World Wide Web Conference (WWW4).*
- K. M. Tolle and H. Chen (2000), "Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools," *Journal of the American Society for Information Science,* Special Issue on Digital Libraries, 51(4), 352-370.
- O. Zamir and O. Etzioni (1999). "Grouper: A Dynamic Clustering Interface to Web Search Results," *Computer Networks*, 31(11-16), 1361-1374.