The Design and Implementation of a Web-based Information System Integrating Cooperation, Information Gathering and Classification

Hui Ma, Yong Tang, Lingkun Wu
Department of Computer Science, Zhongshan(Sun Yat-sen) University, Guangzhou, P.R. China
mahui.sysu@gmail.com, issty@mail.sysu.edu.cn

Abstract

The explosive increase of resources available on the World Wide Web (WWW) not only provides users with valuable information, but also makes the information usage difficult. On the other hand, nowadays it is a trend that more and more people cooperatively and collaboratively work together to complete a task. Efficient coordination and cooperation is required. Thus a cooperative information system is needed. This paper presents the design and implementation of a web-based cooperative information system. It automatically collects information from the WWW, intelligently classifies them into different categories and transforms them into structured records before storing them into the repository. Meanwhile, the system supports cooperative information query and provides a platform that allows users to exchange their ideas with their co-workers. The cooperative features of the system are outlined. Then, the design architecture and implementation techniques are described. Finally its application is illustrated.

Keywords: Cooperation, Information system.

1. Introduction

The prevalence of computer applications and the rapid development of internet related technologies provide a platform for information propagating and sharing. Nowadays, it is quite easy to get tremendous information from the World Wide Web （WWW）. Besides their own experience and existing data, users benefit a lot from extra information on the Internet, which could provide them valuable advices and implications. Meanwhile, a mass of information updates on the WWW every day makes the acquisition and management of information more difficult. On the other hand, with the change of working mode, it's a trend that more and more people are involved in a task to collaboratively complete it [1]. Application domain requires a platform for users to exchange ideas with other co-workers and cooperatively share their information. The traditional information systems gradually become unfit under the onslaught of the information explosion and new cooperative working mode. It is necessary to develop a new series of methodologies, frameworks, tools to fulfill these demands.

In this paper the system design architecture and the implementation of a web-based information system integrating cooperation, information gathering and classification are presented. The system is built based on the following considerations. 1. The popularity and easy-access of WWW. WWW browsers have been part of working environment. They offer simple interfaces to access information in a platform-independent way. 2. Information Storing. Some information is important, for example, it makes significance for future decisions, statistics or analysis. So it is required to provide an information repository along with its interfaces for users to manage information. It is suggested to keep the data organized and structured so as to improve the efficiency and effectiveness of information access. 3. Information gathering. Local information becomes more and more deficient for users along with their increasing demand for high quality knowledge. What they need is lots of fresh and quality information on the Internet. Further more, some users may need the latest information, for example, today's news, or the prices of stocks etc. It is a waste of time and human resources to collect such information now and then. Instead, an automatic search engine is preferred. 4. Communication among users and cooperative information recommendation. When several users collaboratively complete a task together, it is highly probable that the information useful for a user may be also helpful for another user with the same interest. Therefore, when a user inquires information, the system could suggest the most related information based on the evaluation of the interests of other users who share similar interests with this user. Besides, it should allow users to exchange opinions towards different information. 5. Information usage. The system should provide a platform for users to manage information, for example, to query, update or delete information etc. Hence, the system provides the functionalities that it can automatically gather information via search engine from the WWW, classify them into different categories, and put them into an information repository. It learns the user's interests and cooperatively provides the most related information to

1-4244-0963-2/07/$25.00 ©2007 IEEE.

them. Also, it offers a platform to support information usage.

The remainder of this paper is organized as follows: related studies of this paper are introduced in section 2. The system architecture, including its cooperative features, functionalities and design architecture are described in section 3. Some key techniques are discussed in section 4. An illustration is given in section 5. Finally conclusion and future work are given in section 6.

2. Related studies

There has been a tremendous amount of work on information systems. However, the cooperative information system is a relatively young research branch which starts in the early of 1990's. Since then, it attracts great attentions. The works focus on different facets. Jon Mylopoulos and Michael Pagazolou [2] present the shape of cooperative information system and its challenges. Wang et al [3] propose a usage control model in ubiquitous computing environment, allowing the access restrictions directly on services and object documents. Shakshuki et al [4] propose a multi-agent system that cooperatively assists different users to locate and retrieve information. Some work mainly focus on cooperative information collecting. Huang et al [5] propose a subject-oriented web information classification system. It can efficiently collect and classify web pages into different categories, reducing the collection redundancy. Diamadis and Polyzos [6] developed a web collaborative searching assistant, improving searching efficiency. Lin, Ho and Kou [7] developed a web-based portal environment for integrating digital archives and web resources, providing management, index and retrieval services. Some work [8] [9] [10] shed light on cooperative information retrieval, making the acquiring of information more cooperative and personalized.

3. The design architecture of the web-based cooperative information system

In this section the design architecture of the system is proposed. Starting with taking an overview of the information system, its cooperative features are presented subsequently. Then, the design architecture is proposed.

3.1 An overview

As is mentioned in section 1, the system automatically gathers information from the WWW, intelligently classifies them and stores them into a information repository. Meanwhile, it provides a workspace for a user to manage information and exchange ideas with other co-workers.

3.1.1. The workspace. Information is divided into two types according to its accessibility: public and private. Public information is shared among users while private one is only visible to its owner (who creates this piece of information). Therefore, every user has his/her individual workspace. In the workspace a user can manage information.

3.1.2. The search engine. Granted user can get more information by a search engine. User can configure designated key words; websites, the number of layers to be downloaded and searching frequency. Searching frequency means how often would a search engine works, for example, every week, or every day.

3.1.3. Information usage. A user can perform query action on the information repository with designated key words, or browse information within a specific category. The category is a tree-like hierarchy structure which is predefined upon the agreement of all the users. It classifies information into different sorts. The information is listed in descending order of the average ratings from the top most 20% correlated co-workers with the user. A user can select any interesting information from the result list. The chosen information would be retrieved and displayed as preview, showing key words and digests. If it is needed, the user could click on the hyper link and load the whole information page for details.

A user can add to favorite his/her interested information in his/her workspace.

A user can create his/her information and add it to the information repository. Also granted user can update and delete information.

3.1.4. Communication with other users. The communication provides a channel for users to exchange their ideas. A user can recommend information to other users to whom such information may be helpful. A user can also give comments or ratings on a piece of information.

3.2 Cooperative features

In this section the cooperative features are discussed. Cooperation helps to support and improve the efficiency and effectiveness of group work. The cooperative features could be summarized in three points:

3.2.1 Collaborative sharing of resources. As is mentioned in section 1, cooperatively working mode requires users share information, collaboratively complete a task. Collaborative sharing reduces labors of repeated information gathering.

3.2.2. Asynchronous communication. The human channel of information exchange plays a critical part in CSCW. User would exchange ideas during a work process in order to coordinate their work. When a user posts comments or ratings on information, creates new information, or recommends the information to other users, he/she is expressing his/her opinions, giving valuable suggestions to other co-workers. These asynchronous communications help to improve the efficiency of information acquiring.

3.2.3 Cooperative information recommendation. In a cooperative group, different user plays different roles so that they have different interests. Although the comments/ratings and recommendation mechanism could improve cooperation to some extent, it needs users' active actions. The system itself should offer active cooperative mechanism. It could learn the users' interests and provide personalized information query. When a user queries information, the ones which are likely to matches his/her interests are listed in front. Thus different users performing the same query could result in different orders of information lists. This active cooperative mechanism could make the queries more precise, and help users to find their interested information much more quickly.

3.3 The design architecture

In this section the design architecture and functionalities of some key modules are presented.

For the sake of allotting the functional modules coherently and allowing distributed users collaboratively join in this system, the B/S structure is adopted. The interface between client side and server side is browser, using HTTP and HTML, which is common and platform-independent. No more software or updates are needed for the client side. Other complicated components, logical mechanism and data repository is integrated on the server side. The design architecture is shown in figure 1. The functionalities of each module are mainly discussed in three parts.

3.3.1 The information processing. The transformation of raw information from WWW to the organized one stored in information repository must go through several phases.

(1) Search Engine Module. The Search engine searches resources on WWW. Its configuration information resides in an XML file. Before working, XMLParser parses its configuration XML file and initializes the environment. The search engine is designed to be multi-threaded, because there are thousands of millions of resources and it's time-consuming to establish a search connection. Each single thread is in charge of a searching task. If it finishes its task or it expires, another thread is waked up.

(2) Encapsulation Module. The downloaded information data should be transformed into structured one. It is decomposed into an eight tuple form: <id,

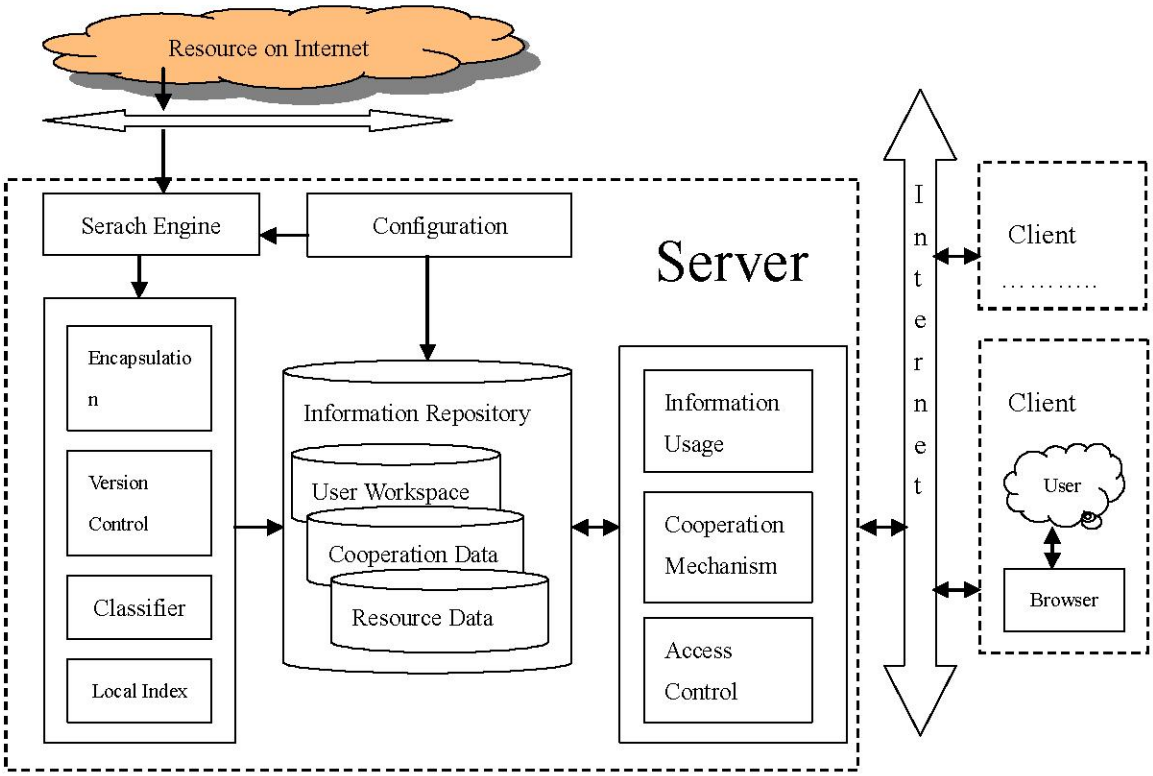

Figure 1. The design architecture of the web-based information system

author, date, source, url, key word, digest, content>. The "id" is the identity of the information. "Author" means the one who creates the information, such as the search engine or the user. "Date" is the date on which this information is created. "Source" means where it comes from and "url" is its hyper link address. "Key word", "digest" and "content" is respectively its designated key words, its digest and its main body. The content will be put into file system while other structured parts of the information will be stored in the DBMS.

(3) Version Control Module. It's probable that the downloaded information is already existed in the information repository, or it's an updated one but the old version is stored in the information repository. The version control module detects such problems and performs corresponding save, update or disregard actions.

(4) Classifier Module. The classifier intelligently classifies the information into a certain category. The information would be stored in the corresponding category folder. The classification method adopts the maximum entropy method [11]. The content of the information are compared with the pre-defined keywords of each category. The information would be classified into the most matched category.

(5) Local Index Module. The downloaded resource is stored in its corresponding category folder. However, the user may usually query information from information repository by key words. If each key word is compared with the file stream, it would greatly hamper the efficiency. Thus an index is built for each resource file to improve the query efficiency.

Through the above phases raw data is transformed into structured information and is put into information depository.

3.3.2. Information Repository. As is illustrated, the information depository stores Users' workspaces, cooperation data and resource data. User workspace part includes user's profiles, his/her private information, favorites and the recommended information he/she receives. Cooperation data part manipulates data such as ratings/comments, which contribute to provide the cooperative mechanism. Resource data part stores the information files downloaded from WWW.

3.3.3. User Manipulation.

(1) Information Usage Module. It provides interfaces for user actions such as query, update, delete, give comments etc.

(2) Cooperation Mechanism Module. It observes users' interests, explores their interest similarities, and offers cooperative recommendation. It acts like a filter that after filtrating it puts ahead the most "related" information with a user's interest. The filtering algorithm learns a user's interest by observing his/her given ratings towards all information. Then it explores the interest similarities among users by comparing their ratings on the same information. The closer their ratings to the same information are, the more similar their interests are. When a user queries information, the system considers the attitudes of his/her similar-interest co-workers. The ones with high average ratings of the similar-interest co-workers will be listed on top [12].

(3) Access Control Module. It manipulates the access control, guaranteeing user's privacy.

4. The implementation

Considering that different users in different working environment will collaboratively join in this system, and future changes may occur, J2EE architecture is chosen because it offers platform independent, flexible and abundant supports of tools and services. The programming language is Java. Database adopts Sqlserver 2000, and BEA Weblogic is chosen for application web container. The system is divided into five layers as depicted in figure 2.

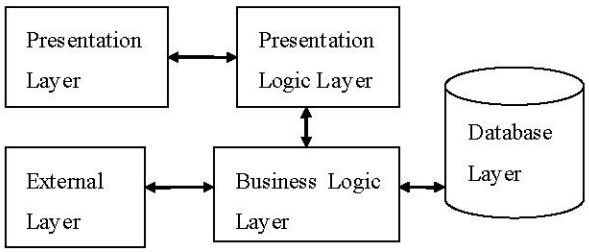

Figure 2. Five layers of the cooperative information system

The external layer offers data exchange interfaces with WWW. The business logic layer implements the business logic and flow control of the system. It contains business logic components and business data components. It uses EJB technologies to reduce the implementation complexity of transactional needs of some business service. The business logic layer provides proxies to interact with database layer. The presentation logic layer handles the interaction between user request and the business logic layer. It includes session management, navigation, providing interfaces with presentation layer from business logic layer. It adopts Struts which is based on MVC pattern, because it provides powerful taglibs, and perfectly separates presentation layer and logic layer, making the maintenance become very convenient. Presentation layer includes JSPs, HTML web pages, CSS, Javascript etc. Users interact with the system via generated HTML web pages.

5. Illustration

In this section some illustrations are given. Figure 3 is a desktop view of a user's workspace.

Figure 3. A user's workspace.

A user's workspace shows his/her created information, favorite ones, and the recommended ones he/she receives. A user can choose any interested information for details.

Figure 4. An information browsing page.

Figure 4 is a screenshot of the information browsing page. The information is shown in preview, giving the keywords and digests. The user can click the hyperlink on the title to load the whole detail information. Comments and ratings of this information are presented below.

6. Conclusion

In this paper we present the design and implementation of a web-based information system integrating cooperation, information gathering and classification. It can automatically gather information from WWW, intelligently classifies them into different categories, and provides a platform for information usage and user communication. It also learns users' interests to provide cooperative recommendations. The cooperative features are discussed and the design architecture and implantation are presented. Finally an application illustration is given. The future work is to improve the information recommendation filtering algorithm, and to make the information query more precise.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (60373081, 60673135); Guangdong Provincial Science and Technology Foundation (04105503, 05200302, 5003348); the Program for New Century Excellent Talents in University

References

[1] Y. Tang, Y. Pan, L. Liang, H. Ma and N. Tang, "Concepts, Model and Framework of Cooperative Software Engineering", *IEEE Transactions on System, Man, and Cybernetics*. 2004, 34, pp. 229-242.

[2] J. Mylopoulos and M. Papazoglou, "Cooperative Information Systems", *IEEE Expert Intelligent Systems & Their Applications*, 1997, 12(5), pp. 28-31.

[3] H. Wang, Y. Zhang and J. Gao, "Ubiquitous Computing Environment and Its Usage Access Control", *Proceedings of the 1st International Conference on Scalable Information Systems*, Hongkong, 2006.

[4] E. Shakshuki, H. Ghenniva and M. Kamel, "An Architecture for Cooperative Information Systems", *Knowledge-Based Systems*, 2003, 16, pp. 17-27.

[5] Y. Huang, Q. Wang, et al, "The Design and Implementation of a Subject-oriented Web Information Classification System", *Proceedings of the Ninth International Conference on Computer Supported Cooperative Work in Design*, 2005, pp. 836-840.

[6] E.T. Diamadis and G.C. Polyzos, "Efficient Cooperative Searching on the Web: System Design and Evaluation", *International Journal of Human Computer Studies*, November, 2004, 61(5), pp. 699-724.

[7] S. Lin, J. Ho and P. Kou, "A Web-Based Resource Discovery Architecture for Digital Archive Systems", *Proc. of the 2002 International Conference on Digital Archive Technologies, Taipei, Taiwan,* 2002, pp.85-90.

[8] N. C. Romano, Jr., D. Roussinov et al, "Collaborative Information Retrieval Environment: Integration of Information Retrieval with Group Support Systems", *Proc. of the 32nd Hawaii International Conference on System Sciences*, Maui, Hawaii, Jan. 5-8, 1999, pp.1053.

[9] H. Peng and Z. Lin, "Learning and Cooperating in Information Gathering", *Proc. of the 24th International Conference on Information Technology Interfaces ITI 2002,* Cavtat, Croatia, June 24-27, 2002, pp. 211-216.

[10] J. Wang, X. Jin, et al, "Distributed and Cooperative Information Retrieval on the World Wide Web", *Journal of Computer Science and Technology*, 2000, 15(6), 611-618.

[11] R. Li, J. Wang, and X. Chen, "Using Maximum Entropy Model for Chinese Text Categorization [J]", *Journal of Computer Research and Development*, 2005, 42(1), pp. 94-101.

[12] J.L. Herlocker, J.A. Konstan et al, "An Algorithmic Framework for Performing Collaborative Filtering", Proc. of the 22nd annual international ACM SIGIR conference on research and development in information retrieval table of contents, Berkeley, California, 1999, pp.230-237.