Rating Information Collecting and Distributing in PICS Protocol

Yao-Tsung Lin, Shian-Shyong Tseng, Yu-Xiu Zhong, and Mon-Fong Jiang*

Department of Computer and Information Science
National Chiao Tung University
Hsinchu 300, Taiwan, R.O.C.

Abstract

Content selection is used to solve the prblem of inappropriate informaiton, and PICS protocol, which developed by W3C, is widely used in many content selection and rating services. In order to solve several issues in a PICS-based rating service, a new mechanism for rating information collecting and distributing is proposed in this paper. A collaborative environment is proposed to solve the problem in rating information collecting. Users' opinions about web content will be collected and analyzed to conclude objective rating information for rating service. In addition to rating information collecting, rating information distribution is another important issue as designing a rating system. In order to solve the issues in storage, performance, and integrity of rating information for a label bureau in PICS protocol, an *LDAP-based Distributed Rating Service*, named *LDRS*, is also proposed. We also design a questionnaire to understand the opinions concerning the Internet content selection of users, and some results will be shown in the experiment. Besides, LDAP-based Distributed Rating Service is implemented and some performance evaluations are also shown in the experiment.

1. Introduction

In recent years, due to the rapid growth of the amount of web pages, web becomes an important media for information communication. However, much inappropriate information is transferred through Internet, so does useless information. It seems that some regulation or content selection for Internet is necessary.

A lot of educational organizations and companies are facing the problem of inappropriate content on Internet, and censorship is a significant method to avoid. However, censorship is quite country-specific, and different countries with different culture seem to apply different strategies on Internet content regulating and many researches are proposed to solve the problems [2, 3, 13, 14, 15, 16].

The concept of content selection is first proposed in many researches to provide a solution to the problem of inappropriate content without using coaction, including PICS [14] protocol proposed by W3C [16], which use content rating to provide information for content selection. However, although some of these researches provide systematic and complete architecture for content selection and rating, there are still some problems in both rating information collecting and distributing.

In this work, a collaborative environment is proposed to solve the problem in rating information collecting. Users' opinions about web content will be collected and analyzed to conclude objective rating information for rating service. The architecture of the proposed collaborative environment consists of the collecting phase and the extracting process. In the collecting phase, we first collect the ratings made by the huge amount of users on Internet. And then, in extracting phase, the rating information or label of each page is extracted from the rating data obtained by using our *Weight Adjusting Algorithm* and *Voting Algorithm*.

In addition to rating information collecting, rating information distribution is another important issue as designing a rating system. In order to solve the issues in storage, performance, and integrity of rating information for a label bureau in PICS protocol, an *LDAP-based Distributed Rating Service*, named *LDRS*, is proposed. *LDRS* consists of *LDAP-based Label Bureau* and *PICS Compliant Client Software* (*PCCS*). *LDAP-based Label Bureau* is responsible for the storing and distributing of rating information, and *PCCS* is responsible for web content filtering according to the rating information from *LDAP-based Label Bureau*.

We also design a questionnaire to understand the opinions concerning the Internet content selection of users, and some results will be shown in the experiment. Besides, LDAP-based Distributed Rating Service is implemented and some performance evaluations are also shown in the experiment.

2. Background

In this section, some related works will be introduced. First, we will introduce the concept of Internet content selection, and several related researches, including PICS [18] proposed by W3C [27], will also be described. After that, some background knowledge used in this work to solve problems in content selection will be presented [9, 10, 11, 12].

2.1 Internet Content Selection

Due to the rapid growth of Internet contents, content regulation on Internet seems to become a public debate. A survey by Georgia Tech concluded that censorship on Internet is mostly concerned by Internet users [3].

2.2 PICS, Platform for Internet Content Selection [14]

To solve the problems of selecting appropriate or desired content via the Internet, many researches have been proposed. In these researches, the PICS [14] protocol is proposed by W3C [16] organization and provided a systematic and complete architecture for document rating system. In addition to the syntax of rating labels, PICS provides the methods of rating and labeling for users.

In PICS protocol, the rating information is provided by two methods, self-labeling or third-party labeling. In self-labeling method, the rating information and rating label is provided by the content providers, and in third-party labeling, the rating information is provided by voluntaries or non-profit organizations The rating information will be coverted into PICS syntax-compliant labels.

After the labels for some pages are determined, the following two methods are used to distribute the labels. First, the rating labels can be added into the META tag of the HTML. Second, the labels may be stored in an existing or a newly created labeling bureau, which is a specific server used to distribute rating label.

In order to use rating information for content selection, a filtering software will be used. According to the corresponding rating information of web pages, the rating filtering software on the user' s computer will decide the access rights to users.

2.3 What Is Directory?

A directory is a specific database containing more descriptive, attribute-based information. In fact, directories are parts of our daily lives. The most familiar examples of directory are phone books and yellow pages that help users to look up the related information conveniently. We refer to these printed directories, alphabetical or classified lists of resource containing names, locations and identifying information, as everyday directories, or sometimes offline directories [6]. And in the world of computer network, DNS (Domain Name System) is a practical example of directory.

To apply the concept of directory service in computer network applications, there are many researches and protocols [8, 10, 11] proposed. By using compliant protocols, directory services in computer network can be integrated and well constructed, which may be more efficient than printed directories. These directory services in computer network are generally called online directories.

2.4 LDAP (Lightweight Directory Access Protocol)

Under the aegis of the IETF OSI-DS (The Internet Engineering Task Force Open System Interconnection-Directory Services) working group, LDAP (lightweight DAP), which is also known as X.500 Lite, was developed in 1989. It arose from initial experience with deploying X.500 directory in Internet. LDAP was formerly used exclusively as a front-end to the X.500 directory and its goal gives simple lightweight access to an X.500 directory, facilitates the development of X.500 DUAs, and uses X.500 for a wide variety of applications [1, 4, 8].

In LDAP, all cooperative directory entries are arranged and shredded in a hierarchical structure to reflect the political, geographic and organizational boundaries. In this hierarchical structure, there is a root on the top. Under the root, the immediate entries represent countries. And then entries of states, companies, or national organizations follow by the entries of countries. Finally, the entries represent individuals, which might be organization units, people, and shared resources like printers, documents, and something else [5].

3. Collaborative Environment for Rating Information Collecting

As we mentioned in Section 2.2, the source of rating information of the rating process based upon PICS can mainly be partitioned into self-labeling and third-party labeling. However, to obtain the rating information of all web pages, sometimes the above two methods are too ideal to apply. The followings are three possible reasons.

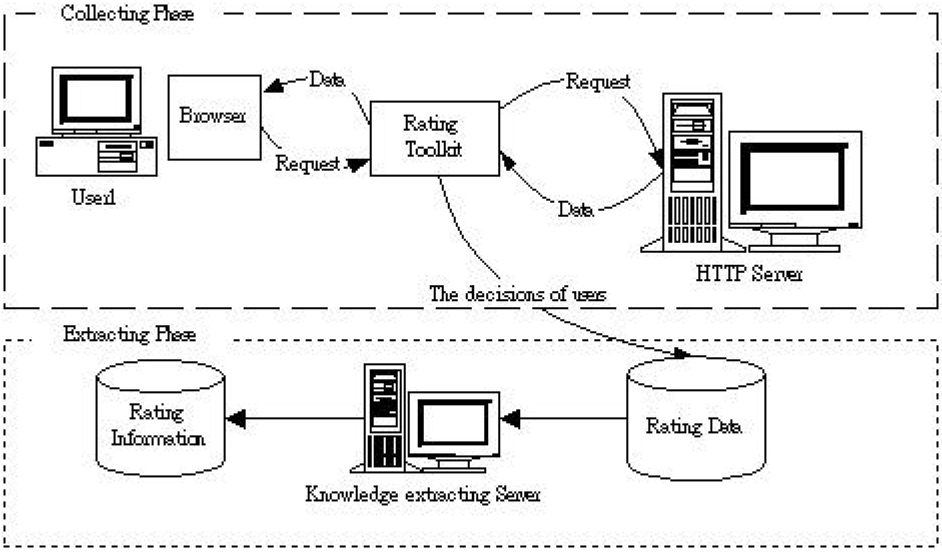


Figure 1: The architecture of the collaborative environment.

- ✍ There is no obligation for content providers to provide the rating.
- ✍ It seems almost impossible to rate all documents by voluntary or non-profit organizations due to the extremely large amount of documents.
- ✍ The acceptable automatic rating system is hard to design.

To solve the above problems, we have designed a collaborative environment for rating information collecting which first collects the ratings made by the huge amount of users on Internet and then extract the rating information or label of each page from the rating data obtained. The architecture of the collaborative environment can be presented in Figure 1:

3.1 Collecting phase

To assist volunteer to rate web pages, an rating software, named *Rating Toolkit* (*RT* in short) is built. By using *RT*, the users must firstly choose a rating category as the basis of their rating process.

3.1.1 Rating Category

Among the contents on WWW, there are a variety of ways to divide these contents into several categories. And levels will be defined in each category.

3.1.2 Provide rating using Rating toolkit (RT)

As shown in Figure 1, when an HTTP request is sent to Internet, RT running on client and bridging the browser and HTTP server (or proxy), records the requested URL information and then sends the request to the HTTP server.

3.1.3 Structuralized Rating Data Collecting

When rating data collecting server receives rating data from *RT*, it must first store the rating data to a temporary storage space before useful rating information is extracted. In order to store rating data properly for extraction phase, an indexed structure is defined. The structure uses terms in URL as index to build a tree. Each node of the tree consists of a table, recording the rating data about the web page whose URL ended at that node in the tree.

3.2 Extracting phase

As we know, the purpose of extracting phase is to extract the rating information or label from rating data provided by users. The extrating phase includes a Voting algorithm and a Rating Information Extracting Module.

3.2.1 Voting Algorithm

To simplify our discussions in the rest of this paper, assume one rating category is used. Besides, we have to find a method to compromise users' opinions. Since the users can easily make their decision by the help of the description of each rating level, a majority voting is used in our collaborative environment.

A weighted vote will be assigned to each participant. The weighted vote of a user represents his/her influence to the voting result; the adjustment of weighted vote will be detailedly described in next section. In this voting algorithm, the candidates are the rating levels of a rating category. For each web page, a voting will be held to consider its rating level.

When sufficient rating data is collected, the extracting phase will be iteratively executed. In the beginning of first iteration, for each page, the level with the maximum weighted vote will be treated as the actual rating level of this page, where the weighted vote of a level is obtained by summing all the weights of users who vote this level, and the weight represents the influence of a user's decisions. In the subsequent iteration, the new weighted vote of a level can be obtained by the following formula:

$$S_i = S_i' + \sum W_j,$$

where S_i' is the weighted vote of level i in previous iteration, and W_j is the weight of user j who rates the page as the level i. It should be noted that once the web page is modified between two consecutive iterations, S_i' will be ignored by resetting as 0. And it is clear that the actual rating level of this web page is the one with maximum S_i.

3.2.2 Rating Information Extracting Module

As mentioned above, the weight of each voting participant represents his/her influence, which can be set to be 1 if all the participants are treated equally. Otherwise the following *Weight Adjusting Algorithm* is proposed to reduce the influence of the participant in each iteration when whose rating data is much different from the others.

Weight Adjusting Algorithm:

Notations:

s : represents the sum of all votes for a given page P.

l : represents the actual rating level of P.

N_i : is equal to the number of votes for rating level i of P.

U_i : represents the set of users who rate P as level i.

t : is the threshold for low ratio.

User name	Total#	Noise#	Weight

is the part of profiles for all users, where *Total#* is the number of pages this user has voted and *Noise#* is the number of votes belonging to the levels with ratio < t.

Step 1: Evaluate the ratio $r_i = N_i / s$ for each rating level $i \neq l$.

Step 2: If $r_i < t$, then increase *Noise#* of u by 1, for all u belongs to U_i.

Step 3: Increase *Total#* of u by 1, for all u belongs to U_i.

By repeatedly executing the above algorithm until all rated pages have been considered, the new weight for each participant can be recalculated using the formula below,

$$weight = (Total\,\# - Noise\,\#) / Total\,\#.$$

4 LDRS (LDAP-based Distributed Rating Service)

On the other hand, there are two main methods mentioned in Section 2.2 to distribute the rating information in PICS protocol, including HTML META tag and label bureau. But there are some problems may occur in the META tag approach, as listed in the following.

1. Content providers does not necessarily provide the rating label in her/his web page.
2. No one can force the authors of web pages to add rating labels.
3. If the rating information in web pages is totally decided by the authors, it seems too subjective.

As for the method using label bureau to distribute the label, we can avoid problems occurring as mentioned above easily. All rating information is kept in label bureau, and users can select rating system they preferred by choosing the label bureau following certain rating system. No one needs to change the source content and worry about the authors' opinions of web pages. By this method, users can decide whether or not to use rating information and decide which rating systems they like to use.

However, the following three issues should be considered first before we design a label bureau for rating information distribution. These issues includes the storage to store extrmely large amount of rating information, performance of rating serveice when the usage of raing service increases, and the integrity of rating information when they stored in distributed servers.

In this work, an LDAP-based rating service, named LDRS (LDAP-based Distributed Rating Service), is proposed to construct a distributed environment for PICS compliant rating service, and solve the issues in storage,

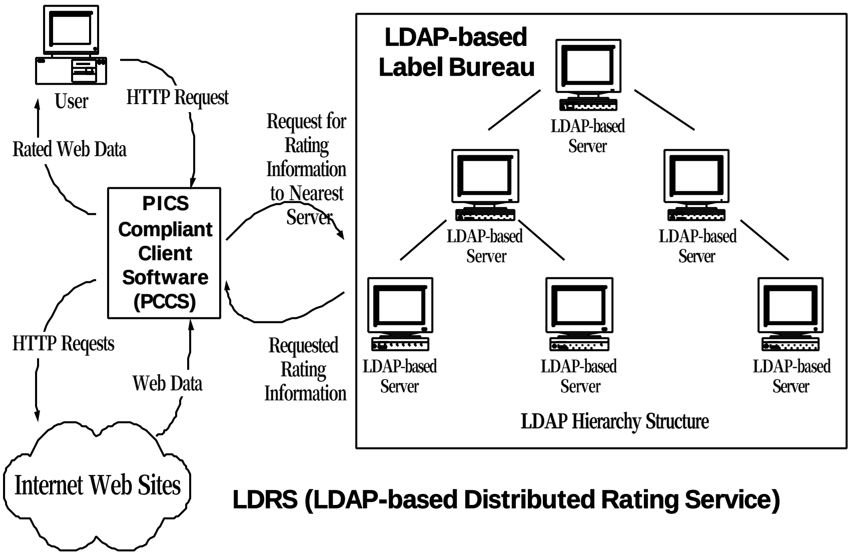


Figure 2: The architecture of LDRS.

performance, and integrity of rating information. The architecture of LDRS is shown in Figure 2.

LDRS consists of server and client modules. An LDAP-based Label Bureau is designed as the rating information distribution server. Besides, the client software, named PCCS (PICS Compliant Client Software) is designed as a gatekeeper to connect label bureau and control the access rights by filtering web pages.

4.1 The Hierarchy Concept in LDRS

All information in directory is represented as data entries and every entry has its own *DN* (distinguish name), which consists of hierarchical attributes. Hierarchical attribute is the information to distinguish data entry in each branch of the directory hierarchy tree.

Besides, data attributes are another important information in directory data entries. They are the requested information as we allocated the data entry in directory tree using hierarchical attributes.

By associating the data attributes with the hierarchical attributes, the information can be stored in different servers, which are arranged in hierarchical structure. When we request information from directory, no matter which server is connected, the directory structure is logically integrated into one complete rating service.

As to rating service, the rating information of web pages is allocated by URLs (Uniform Resource Locators), which are naturally hierarchical arranged and stored. Because the query from client software is based on URL, the rating information in label bureau may be arranged into tree-like hierarchy. This conforms to characteristic of hierarchical structure in LDAP. In fact, the architecture of proposed rating service is similar to DNS.

4.2 LDAP-based Label Bureau

The functionality of label bureau is that when label bureau received a query from client software, it must respond with the rating information of target web page. And the initial directory database for label bureau can be built up by constructing LDAP service.

4.2.1 Design of Label Bureau Schema

After understanding the concept of LDIF, whiech is used to describe a directory and its entries in text format., needed schema can be designed. The contents of web pages on Internet are concerned in our work, and the web pages will be rated according to their URLs. We need to analyze URLs to know what kinds of files the browsers get. This information is important for building up label bureau schema. Because all searching and updating rely on URL, so a DNS similar schema can be designed.

4.2.2 Object Class and Attribute

All data entries in directory are known as object classes. The data property of entry can be described using an object class (*ObjectClass*) attribute, and there are many data attributes in an object class.

There are a lot of object classes in LDAP specifications. How to choose appropriate one for a given system is an important issue. Based on hierarchy concept in DNS, we choose and create corresponding object classes for representing information in label bureau.

4.3 PCCS (PICS Compliant Client Software)

The client software, named PCCS (PICS Compliant Client Software) supports content filtering and access controlling functionalities according to the rating information from label bureau. When PCCS received a request of user, the software would take responsibility to filter web pages according to the returned rating information, unless nothing is returned when no corresponding rating information exists. The design of PCCS in our architecture is shown in Figure 3.

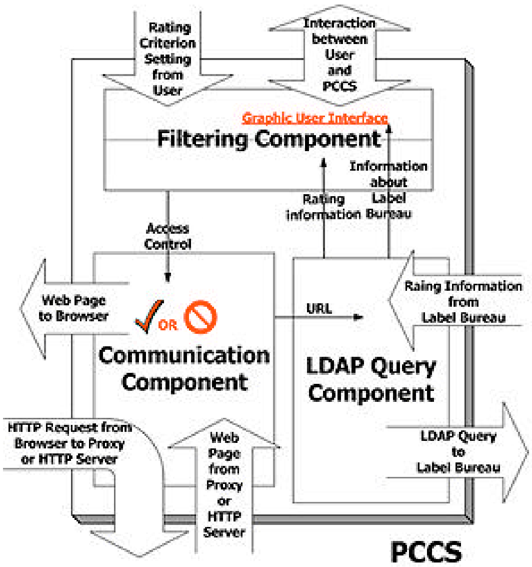

Rating Criterion Setting from User
Interaction between User and PCCS
Graphic User Interface
Filtering Component
Information about Label Bureau
Rating information
Access Control
Web Page to Browser
✓ OR
Communication Component
URL
Raing Information from Label Bureau
LDAP Query Component
HTTP Request from Browser to Proxy or HTTP Server
Web Page from Proxy or HTTP Server
LDAP Query to Label Bureau
PCCS

Figure 3: Three components in PCCS.

There are three components in this PCCS architecture including Communication Component, LDAP Query Component and Filtering Component, which will be discussed in following sub-sections.

4.3.1 The Communication Component

The Communication Component in PCCS is responsible for the connections between browser, proxy, and Internet. When user tries to get a web page, browser will send an HTTP request with URL of that page to the corresponding server. At the same time, Communication Component will catch the request in parallel and bridge the server and client browser. All received requests in Communication Component will be transformed to corresponding formats before sent to proxy or HTTP servers. At the same time, the parsed URL in each HTTP request will be used in LDAP Query Component.

After the corresponding server or proxy returns the data content, Communication Component will take responsibility for sending the data back or not according to the access control rights made by Filtering Component, which will be discussed later.

4.3.2 The LDAP Query Component

The LDAP Query Component is responsible for communication with the LDAP–based label bureau. When Communication Component parses the URL information from user's request, LDAP Query Component will generate a query according to the syntax of LDAP protocol with URL to get the rating information in label bureau. After LDAP Query Component receives the response from label bureau, it will parse the corresponding result of rating information by the syntax to get the values of rating categories of web page. These values then will be used in Filtering Component, and evaluated with the rating criteria set by user to determine the access right.

4.3.3 The Filtering Component

Filtering Component is responsible for the access control evaluation in PCCS. Using Filtering Component that we developed, users can have interactions with PCCS. First, this component gets the rating criteria set by user through Graphic User Interface and generates corresponding criteria ranges. After that, when LDAP Query Component gets the rating information from label bureau, Filtering Component will be triggered to evaluate the access right using the parsed values in rating information with the rating criterion ranges.

As the result of the function returns true, Filtering Component will permit Communication Component to send the requested web page to browser. Otherwise, Communication Component will be acknowledged to send a warning to user.

5. Experiments and Implementation

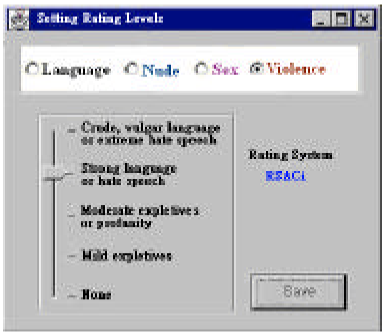
In this section, the development of Internet in Taiwan will be first introduced, and a questionnaire is also be designed to understand the opinions concerning the Internet content selection of users, and some results will be shown in the experiment. Besides, LDAP-based Distributed Rating Service is implemented and some performance evaluations are also shown in the experiment.

5.1 Opinions about content rating in Taiwan

In this work, a questionnaire has been designed and used to collect the opinions about the content selection in Taiwan. According to the statistics of 600 answers of our designed questionnaire, up to 80% answers agree the promoting of Internet content selection, but about 65% answerers didn' t know the method of PICS protocol. Among the answerers agreeing to promote Internet content selection, up to 90% would support the collaborative platform for Internet content selection, and about 70% answerers are willing to be the rating volunteers to assistant the work of Internet content selection.

5.2 Implementation of LDRS

For the distribution of rating information, LDRS is implemented for users to use rating service. In addition to the LDAP based label bureau, corresponding PCCS is also implemented in our experiment.



Setting Rating Levels

○ Language ○ Nude ○ Sex ● Violence

Crude, vulgar language or extreme hate speech

Strong language or hate speech

Moderate expletives or profanity

Mild expletives

None

Rating System: RSACi

Figure 4: Setting the rating criteria.

When rating criteria is set and PCCS is started, PCCS will listen to browser connections and sends corresponding requests to both WWW servers and label bureau. Then PCCS will decide the access rights of web pages for users according to the rating information from label bureau and user-set criteria.

After the access rights of web pages is decided, the web page content will be sent to browser if the web page is allowed to browse, or a warning will be sent and displayed in user' s browser.

5.3 Experiment

As mentioned in Section 4.3, PCCS in LDRS bridges user and Internet content to provide rating service. In our experiment, PCCS is integrated with proxy of LAN (Local Area Network), which is naturally the gateway for LAN users to connect to Internet. By integrating proxy server with PCCS, large amount of rating information requests and higher flow rate of web pages can be obtained to evaluate the performance of LDRS architecture. An enhanced proxy following the architecture of LDRS is implemented.

In our experiment, the LDAP-based label bureau with 5,090 records is used during the construction of rating service. Each record stores rating information of a single web page, and the rating information is transferred and stored in LDAP-based label bureau.

In the experiment, it has found that as the amount of rating information grows, the average response time of a request will also increase. If we observe the performance of the LDAP system and proxy system used, it can be easily seen that the bottleneck of the enhanced proxy system is the LDAP system. But actually, comparing to general database system, LDAP system is specifically designed for larger amount of data with hierarchical attributes. The response time of general database will be more than that of LDAP system. So more proper hierarchical arrangement using or newer version essentially in LDAP system will improve the response time for enhanced proxy.

Also, the HTTP request and LDAP query request were parallel sent; it means the HTTP data receiving process will not be blocked, but it seems likely to be buffered. The throughput of HTTP data will be the same for both the enhanced proxy server or normal proxy. In Figure 5 the average throughputs of proxies with different amount of rating information are shown; it seems no obvious difference between these proxies exists.

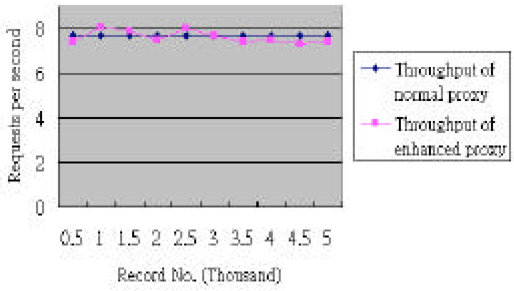


Figure 5: Average throughputs of proxies.

As a conclusion, using the enhanced proxy server to provide rating service is effective if the amount of requests is not huge. It seems the enhanced proxy sever may be suitable for small-scale LAN and provide a management of the content of network an over all network management. But as the amount of network traffic increases, personalized PCCS may be needed for better performance and may provide user specific filtering strategy.

6. Conclusion

In order to solve the problems of rating information collecting and distributing in a rating system, a collaborative environment for rating information collecting and LDAP-based Distributed Rating System for rating information distribution are proposed. In a collaborative environment for rating information collecting, users' opinions about web content will be collected by *Rating Toolkit*, and then usable rating information can be extracted by using *Weight Adjusting Algorithm* and *Voting Algorithm*.

An *LDAP-based Distributed Rating System* is also proposed in this paper, rating information can be efficiently stored and distributed by this mechanism. *LDAP-based Distributed Rating System* consists of two parts, including *LDAP-based Label Bureau* and *PICS Compliant Client Software*. By taking advantage of the benefits of LDAP, the issues in storage, performance and integrity of rating information in a rating system can be solved.

Since users' tendencies and assistance are very important in collaborative environment, we also design a questionnaire to understand the opinions concerning the Internet content selection of users. Besides, *LDAP-based Distributed Rating Service* is implemented and the performance of the rating service is shown in the experiment.

This work was supported by Ministry of Education and National Science Council of the Republic of China under Grand No. 89-E-FA04-1-4, High Confidence Information Systems.

Reference

[1] Gordon Benett, "LDAP: A Next Generation Directory Protocol," August 15, 1996, http://idm.internet.com/foundation/ldap.shtml.

[2] R. C. Ellickson, *Order without law: how neighbors settle disputes*, Harvard University Press, 1991.

[3] Georia Tech., GVU's 6th WWW survey, *http://www.cc.gatech.edu/gvu/user_sruveys/survey-10-1996/*, 1996.

[4] GroupLens, http://www.cs.umn.edu/Research/GroupLens/grouplens.html

[5] Timothy A. Howes and Mark C Smith, "A Scalable, Deployable, Directory Service Framework for the Internet," CITI Technical Report 95-7, April 28, 1995.

[6] Timothy A. Howes, "The Lightweight Directory Access Protocol: X.500 Lite," CITI Technical Report 95-8, July 27, 1995.

[7] Timothy A Howes, "LDAP: Use as Directed-The co-author of LDAP sets the record straight on what the protocol can and can't do," Data Communication, February 1999.

[8] Timothy A. Howes, Mark C. Smith, and Gordon S. Good, "Understanding and Deploying LDAP Directory Services," Netscape Communications Corporation, first edition, 1999.

[9] Mon-Fong Jiang, Shian-Shyong Tseng, and Yao-Tsung Lin, "Collaborative Rating System for Web Page Labeling," World Conference of the WWW and Internet, Honolulu, Hawaii, 1999.

[10] Steve Kille, "LDAP and X.500 Article," Messaging Magazine, September 1996, http://www.messagingdirect.com/publications/IC-6033.html.

[11] "Administrator's Guide – Netscape Directory Server Version 4.1," Netscape Communications Corporation, 1999.

[12] "Deployment Guide - Netscape Directory Server Version 4.0," Netscape Communications Corporation, 1998.

[13] "Installation Guide – Netscape Directory Server Version 4.1," Netscape Communications Corporation, 1999.

[14] "Managing Servers with Netscape Console – Version 4.1," Netscape Communications Corporation, 1999.

[15] H. A. Peng, "How Countries Are Regulating Internet Content," *INET'97 proceeding*, 1997.

[16] PICS, Platform for Internet Content Selection, *http://www.w3.org/PICS*

[17] RSAC, Recreational Software Advisory Council, *http://www.rsac.org*

[18] W3C, World Wide Web Consortium, *http://www.w3.org*