CONTENT RECOMMENDATION SYSTEM BASED ON PRIVATE DYNAMIC USER PROFILE

TING CHEN, WEI-LI HAN, HAI-DONG WANG, YI-XUN ZHOU, BIN XU, BIN-YU ZANG

Software School, Fudan University, China
E-MAIL: {052053009*, wlhan, 062053008, 0461057, 0361069, byzang}@fudan.edu.cn

Abstract:

As the amount of the accessible information in the Internet is overwhelming, personalized content recommendation system offers spam filtering service and suggests useful information to the end users. It is a hotspot in the research area of content management on WWW. Traditional recommendation systems do the data mining on web access logs, discover user's access patterns, and filter the information on behalf of the user at the server side. One critical limitation of traditional recommendation system is the lack of user's private daily data, such as schedules, favorite websites and personal emails. The reason for this limitation is the privacy leak issue when the server holds much more private user data. To solve this problem, this paper presents an agent-based personalized recommendation method called Content REcommendation System based on private Dynamic User Profile (CRESDUP). The system collects and mines the private data of user at the client side, discovers, stores and updates private Dynamic User Profile (DUP) at the client side. The system fetches preferred message from the content server according to DUP. An important usage of this technology is a personalized advertising system in the RSS (Rich Site Summary, or RDF Site Summary) reader application. Our experiment shows that the system can utilize DUP to identify the customers' potential preferences and deliver the more preferred messages, especially the advertisements, to people who are interested.

Keywords:

Content recommendation; Dynamic user profile; Privacy-Enhanced personalization; Data mining; RSS

1. Introduction

Nowadays as the information world is increasingly expanding, the problem of the overwhelming information flood is becoming more and more serious [6]. Users are overloaded by thousands of messages and a large quantity of information, most of which are spam [9]. What they need is a classification tool representing their preferences.

Content Recommendation (CR) technology [15] can help users get preferred information according to the users' predefined preferences [14] or some usage patterns mined from web access logs [4]. In general, more detailed private user information, including personal schedules, emails, recently visited websites, can help CR system to provide more accurate contents. However, this may cause the privacy leak issue.

In traditional content providing services [16], the personal data to improve the accuracy of the provided information only include web access logs, predefined static preference lists, part of the contact information. Users do not want to store much detailed personal information on the server. Besides, users must set their preference lists and contact information on each server to get the customized service from different service providers.

In this paper, we present a novel agent-based personalized recommendation system called Content REcommendation System based on private Dynamic User Profile (CRESDUP). This system is able to find out preferred messages on the Internet according to private user data and protect the privacy of user at the same time as these personal data are processed at the client side. A private Dynamic User Profile (DUP) is constructed through data mining on user's personal data at the client side. DUP is regularly updated according to the changes of the personal data during user's daily operations and the user's feedbacks. Furthermore, DUP can be reused for different content service providers because it is stored and utilized at the client side.

The rest of the paper is organized as follows. Section 2 introduces the background and surveys the related work; Section 3 presents our personalized CR method; Section 4 describes an application scenario: RSS Ad service with CRESDUP, discusses some main issues related to our

1-4244-0973-X/07/$25.00 ©2007 IEEE

method in the scenario and evaluates its performance; Section 5 discusses some remaining issues; Section 6 concludes the paper and discusses the future work.

2. Background

2.1. Content Recommendation

To provide more personalized service on web, Web Usage Mining (WUM) is proposed [16]. WUM typically extracts knowledge by analyzing historical data of users or servers. Web Personalized and recommender systems [16-17] are typical applications of WUM.

Recommendation systems emerged as an independent research area in the mid-1990s and are usually classified into three categories, based on how recommendations are made [2].

- *Content-based recommendations*: The user will be recommended items similar to the ones the user preferred in the past;
- *Collaborative recommendations:* The user will be recommended items that people with similar tastes and preferences liked in the past.
- *Hybrid approaches*: These methods combine content-based and collaborative methods.

CRESDUP focuses primarily on content-based recommendation methods since it need not communicate with other users.

2.2. Content Pushing Service and Anti-Spam

Content Pushing Service (CPS) can deliver subscribed messages to user. RSS is a popular application of CPS. To use RSS, users subscribe to sites and monitor the updates of these sites with an RSS reader. The RSS reader retrieves the RSS feed periodically from a server by downloading the RSS data file.

However, RSS may push some messages which are not preferred by subscriber. The preferences of subscriber are dynamically changing, but the subscribed feeds do not change so fast to fit his/her preferences. If these uninteresting messages are numerous, the subscriber may refuse to use the CPS. This problem does harm to content providers, because these messages could bind with some advertisements, which are the main source of income for most content providers. This is a typical security issue: anti-spam [10-12].

2.3. Motivation

This paper provides a novel method to recommend preferred message to end user based on his/her private data. We apply this method to advertisement recommendation during RSS content pushing.

2.4. Related Works

The content-based approach of recommendation has its roots in information retrieval [1], and information filtering [3] research. Content-based systems are designed mostly to recommend text-based items; the content in these systems is usually described with keywords. For example, a content-based component of the Fab system [2], which recommends web pages to users, represents web page content with the 100 most important words. Similarly, the Syskill & Webert system [8] represents documents with the 128 most informative words.

Besides the traditional heuristics that are based mostly on information retrieval methods, other techniques for content-based recommendation have also been used, such as Bayesian classifiers [7], and various machine learning techniques, including clustering, decision trees, and artificial neural networks [8]. In the area of user profile processing and evolution, Sparacino [13] presents a method based on Bayesian networks to construct a profile for providing customized services to museum visitors.

3. Architecture of Cresdup

3.1. Overview

To classify the messages according to the user's preferences, we have designed an intelligent personalized content recommendation system. We name this system CRESDUP (Content REcommendation System based on private Dynamic User Profile). In figure 1, the three-tier of the system is demonstrated. Details will be discussed in the following sections.

There are four core modules in our system:

User Raw Information Collection Agent; User Profile Analysis Agent; Content Recommendation Server; Content Recommendation Client Agent.

The system consists of three tiers. The first tier is the data layer, including the user raw information collection agent. The second tier is the logic layer, including the build-up process of DUP and the usage of DUP through different agents. The third tier is the presentation layer, including the customized UI (User Interface) on the PC clients or mobile devices.

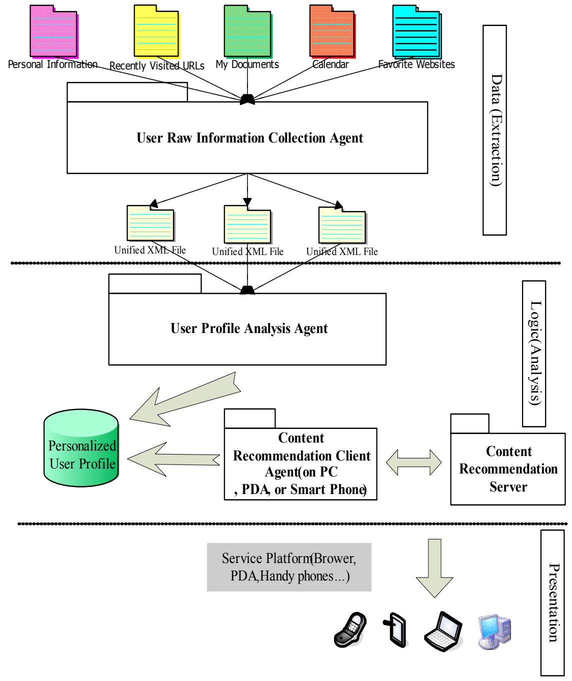

Personal Information
Recently Visited URLs
My Documents
Calendar
Favorite Websites
User Raw Information Collection Agent
Data (Extraction)
Unified XML File
Unified XML File
Unified XML File
User Profile Analysis Agent
Logic(Analysis)
Personalized User Profile
Content Recommendation Client Agent(on PC , PDA, or Smart Phone)
Content Recommendation Server
Service Platform(Brower, PDA,Handy phones...)
Presentation

Figure 1. The CRESDUP system as a three-tier application

3.2. User Raw Information Collection Agent

The user profile includes user's work pattern and life mode. It is created from the user raw information data. An agent is employed to do the data collection work. It collects the raw data from the user's personal data resources locally which include:

Basic Personal Information: Name, Age, Gender, Location, Company, Preference Lists, etc.

Recently Visited URLs: The websites the user has browsed recently.

Self-made documents: The documents which are designed, written and reviewed by the user, including emails, notes and work documents.

Calendar and Schedule: The schedule of a user, both that of work and life.

Favorite Websites: The websites user adds to the favorite list of the browser.

A feedback agent is used to collect the feedbacks as an important source for the improvement of DUP. So that DUP can be updated regularly and represent the user's dynamically-changing preferences.

Because of the different formats of the raw data, they are formatted into predefined XML files.

3.3. User Profile Analysis Agent

The user profile analysis agent makes a user profile from the formatted raw data. In this system, content-based recommendation algorithm is employed. The user will be recommended the data items similar to the ones they preferred in the past.

In content-based recommendation methods [2], utilities(*c, d*), which is the importance of a data item *d* for user *c* is estimated from the known utilities*(c, d_i)* assigned by user *c* to data items $d_i \in$ D which are similar to data item *d*.

After the raw information data are collected and formatted into predefined XML files, the user profile analysis agent uses the term frequency/inverse document frequency (TF-IDF) [5] measure for specifying the keyword weights of the raw data. A user profile, i.e., a set of attributes characterizing item, is made and used to determine the appropriateness of the information item for recommendation purposes.

TF-IDF is defined as follows: Assume that N is the total number of documents that are collected and that keyword k_j appears in n_i of them. Moreover, assume that $f_{i,j}$ is the number of times keyword k_i appears in document d_j. Then $TF_{i,j}$, the term frequency of keyword k_i in document d_j, is defined as

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}} \tag{1}$$

where the maximum is computed over the frequencies $f_{z,j}$ of all keywords k_z that appear in the document d_j. Since keywords that appear in many documents are not useful in distinguishing a relevant document from an irrelevant one. Therefore, the measure of inverse document frequency (IDF_i) is often used in combination with simple term frequency ($TF_{i,j}$). The inverse document frequency for keyword k_i is usually defined as

$$IDF_i = \log \frac{N}{n_i} \tag{2}$$

Then, the TF-IDF weight for keyword k_i in document d_j is defined as

$$w_{i,j} = TF_{i,j} \times IDF_i \tag{3}$$

And the content of document d_j is defined as

$$content(d_j) = (w_{1j,} \ldots_, w_{kj})$$

A document often contains different source types of keywords, such as subjects, contents, and other types of attributes. Hence, the final profile of the document is made up created from weighted keywords lists by the normalization of weighted keywords. Finally, the DUP is a collection of all the document profiles.

3.4. Content Recommendation Server

The content recommendation server manages the contents and sends them to the client according to the client request. The content recommendation server organizes the to-be-pushed contents into a category-tree (Figure 2). Each non-leaf node of the tree is a category while the leaf node is a piece of content. By TF-IDF algorithm, the keywords are extracted from the raw contents and the content profiles for each category are maintained.

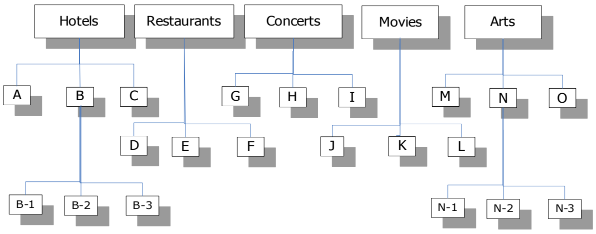


Figure 2. Category-tree on the server

The content profiles for each category will be used as the metadata of the contents. It is formatted into XML files. With this metadata, the server does not have to send the whole content to the client for the selection and filtering process. Thus, it saves time and network resource.

3.5. Content Recommendation Client Agent

The client is an agent to receive information data from the server based on user's profile. It can either be an independent module of the client-side software or an integrated part of the client-side software.

By receiving the content metadata profile from the server, the agent can compare it with the local DUP. Since both content metadata profile and DUP are represented as TF-IDF vectors w_c and w_u of keyword weights, the relation between them can be evaluated by some scoring heuristic, such as the cosine similarity measure.

$$u(c,u) = \cos(\vec{w}_{c,} \vec{w}_u) = \frac{\vec{w}_c \cdot \vec{w}_u}{\left\|\vec{w}_c\right\|_2 \times \left\|\vec{w}_u\right\|_2}$$

$$= \frac{\sum_{i=1}^{K} w_{i,c} w_{i,u}}{\sqrt{\sum_{i=1}^{K} w_{i,c}^2}\sqrt{\sum_{i=1}^{K} w_{i,u}^2}} \tag{4}$$

where K is the total number of keywords in the system. Consequently, the algorithm will assign higher values to those contents that have more similar features in the metadata file as DUP. Then the client selects these contents, creates a content request and sends it to the content recommendation server. The preferred contents will be delivered to the client by the server.

3.6. Main Workflow of CRESDUP

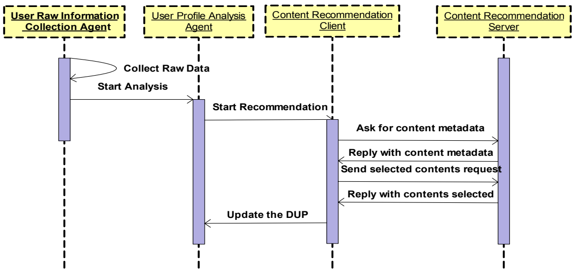

User Raw Information Collection Agent
User Profile Analysis Agent
Content Recommendation Client
Content Recommendation Server
Collect Raw Data
Start Analysis
Start Recommendation
Ask for content metadata
Reply with content metadata
Send selected contents request
Reply with contents selected
Update the DUP

Figure 3. Workflow of CRESDUP system

The main workflow sequence is illustrated in figure 3.

Firstly, the information collection agent collects the data from the user's daily usages on the computer. Secondly, with the user profile analysis agent, a DUP representing user's preferences is produced from these data. Thirdly, the content recommendation client agent uses DUP to select and filter the information received from the content recommendation server. The client can be deployed on the desktop PC/laptop PC, smart phones, PDAs and other smart client devices. The whole working process is conducted on the client side with no interference from public servers.

4. Application Scenario and Performance Evaluation

4.1. Application Scenario Description

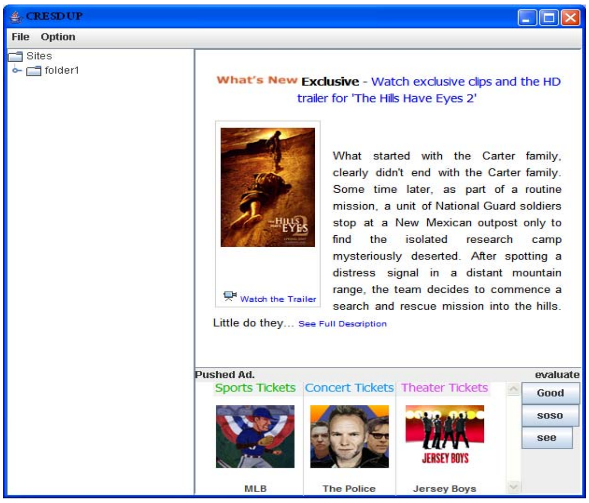
The RSS-based advertising system presented in this paper is an application of CRESDUP and is used to demonstrate the features of the technologies mentioned above.

Advertising on RSS readers based on user's preferences has large potential due to the very personal and intimate nature of the RSS application. The subscribed RSS feeds imply the interests of the user. Combining them with other personal information provides a precise personal profile of the user. Offering valuable and related ads when the user is reading the RSS data greatly improves the ad effectiveness. The advertising system is augmented with personalized profiles so that only relevant, targeted ads are pushed to the users.

The system is divided into two parts: an RSS reader (Figure 4) with CRESDUP support and an advertisement server.

The RSS reader is similar to the traditional RSS reader except for its advertising function. It contains an ad bar below the RSS view window. The ad contents provided in this banner are carefully chosen with respect to DUP. Therefore, the user is able to get the most interesting advertisements.

The ad server manages the ads, classifies them into organized categories and sends the ads to the ad agent clients according to their requests.



What's New Exclusive - Watch exclusive clips and the HD trailer for 'The Hills Have Eyes 2'

What started with the Carter family, clearly didn't end with the Carter family. Some time later, as part of a routine mission, a unit of National Guard soldiers stop at a New Mexican outpost only to find the isolated research camp mysteriously deserted. After spotting a distress signal in a distant mountain range, the team decides to commence a search and rescue mission into the hills. Little do they... See Full Description

Figure 4. The Client UI

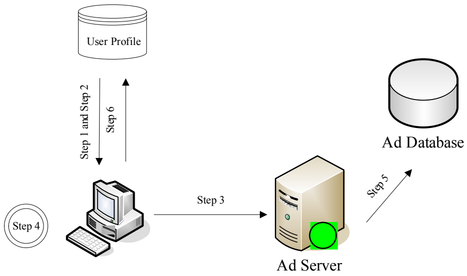


Figure 5. Main architecture and work flow sequence of this RSS advertising system

The main architecture and work flow sequence of this RSS advertising system are illustrated in Figure 5:

(1) User starts the RSS reader and subscribes some RSS feeds.

(2) The user raw information collection agent and the user profile analysis agent work together to build DUP.

(3) The RSS reader sends the ad category metadata request to the ad server and receives the reply.

(4) The client agent analyses the category metadata and computes the similarity between each node of the category tree and DUP.

(5) The compared results indicate the user's potential interests of the ads in this category, the higher the better. Depending on this result, the ads server sends the corresponding ads to the RSS reader client.

(6) When the user receives and reads the ads, an explicit feedback is made so that the system could also learn user's preferences by giving options like "I need more" and "No, thanks". The DUP is dynamically updated.

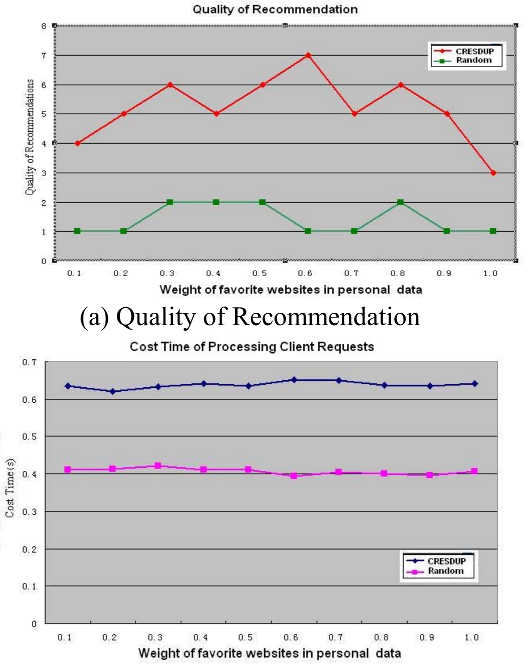
It is worth noting that the user's preferences are kept on the client side instead of the server. Thus the privacy is protected from being exposed to the server.

4.2. Performance Evaluation

In order to validate our application, several experiments have been performed. The ad server contains 200 pieces of ads divided into 10 categories, including movies, music, fashion, automobile, toys, sport, shops, jobs, health care and PC games. The DUP raw data resources include the user's subscribed RSS feeds, favorite websites and daily schedules. The weights of these resources can be changed and the sum of all the weights is 1.0. In the experiment, the weight of favorite websites is modified from 0 to 1.0 and the other two resources share the remaining weight value equally.

The first experiment shown in Figure 6a is the evaluation of the recommendation quality. The users are asked to give a score to the ads. The higher score indicates the more useful the ads are. It can be seen that most of the ads selected by CRESDUP meet the user's tastes.

Figure 6b is the cost time of processing client requests on the server. The overhead introduced is small. Therefore, CRESDUP is efficient.



(a) Quality of Recommendation

(b) Time Cost of Recommendation

Figure 6. The Performance Evaluation

5. Discussion

Dynamic updating of DUP is a remaining issue in CRESDUP. We use a simple offline solution to update DUP. A new DUP is periodically created. If the cost time of creating DUP is critical, the online way to update DUP is a better option. It can reduce the updating cost of DUP.

6. Conclusion and Future Work

Based on personalized content recommendation service, the information classification and filtering are done automatically. The intelligent information process is transparent and effective, which makes the user's life easy.

This recommendation system can be extended in several ways, which include improving the understanding of users and contents, incorporating the contextual information into the recommendation process, supporting multiple ratings, and providing more flexible and less intrusive types of recommendations.

We are planning to make this system take additional contextual information, such as time, place, and the job of the user, into consideration when recommending the related information.

References

[1] Baeza-Yates, R., Ribeiro-Beto, B.: Modern Information Retrieval. Addison-Wesley. (1999)

[2] Balabanovic, M., Shoham, Y.: Fab: Content-Based, Collaborative Recommendation. *Comm. ACM*, vol. 40, no. 3, (1997) 66-72.

[3] Belkin, N., Croft, B.: Information Filtering and Information Retrieval. *Comm. ACM*, vol. 35, no. 12, (1992) 29-37

[4] Ron, K.: Mining e-commerce data: the good, the bad, and the ugly. *Proceedings of the seventh ACM SIGKDD international conference*(2001)

[5] Jing, L., Huang, H., Shi, H.: Improved Feature Selection Approach TFIDF in Text Mining, *Proc. 1st Internet Conference on Machine Learning and Cybernetics*, Beijing (2002)

[6] Ali, F.F. and Don, H.D.: Managerial information overload. *Comm. ACM*, vol. 45, no. 10, (2002) 127 - 131

[7] Mooney, R.J., Bennett, P.N., Roy, L.: Book Recommending Using Text Categorization with Extracted Information. *Proc. Recommender Systems Papers from 1998 Workshop,* Technical Report WS-98-08. (1998)

[8] Pazzani, M., Billsus, D.: Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, vol.27, (1997) 313-331

[9] Thede, L.,Marshall, V.A.,Rick W.:An Economic Answer to Unsolicited Communication. *EC'04*. (2004)

[10] Goodman, J. and Rounthwaite, R.: Stopping outgoing spam. In *Proceedings of the ACM Conference on Electronic Commerce (EC'04)* (New York, May 17-20), ACM Press, New York, 2004, 30-39.

[11] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E.: A Bayesian approach to Filtering Junk e-mail. In Learning for Text Categorization – Papers form *the AAAI Workshop*. AAAI Technical Report WS-98-05 (Madison, WI, 1998).

[12] Goodman, J., Cormack, G. V., Hecherman, D.: Spam and the Ongoing Battle for the Inbox, *Comm. ACM*, Vol 50, 2 (Feb., 2007): 25-33.

[13] Sparacino, F.: Sto(ry)chastics: A Bayesian Network

Architecture for User Modeling and Computational Storytelling for Interactive Spaces. In *Proceedings of UBicomp 2003, Seattle, WA, USA,* Springer. (2003)

[14] Miller, B.N., Albert, I., Lam, S.K., Konstan, J.A. and Riedl, J.: MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System. *Proc. Int'l Conf. intelligent User Interfaces*, 2003

[15] Hill, W., Stead, l., Rosenstein, M., and Furnas, G.: Recommending and Evaluating Choices in a Virtual Community of Use. *Proc. Conf. Human Factors in Computing Systems*, 1995

[16] Baraglia R., Silvestri F.: Dynamic Personalization of Web Sites Without User Intervention. *Comm. ACM*, Vol. 50, 2 (Feb., 2007): 63-67.

[17] Eirinaki, M. and Vazirgiannis, M.: Web Mining for Web Personalization. *ACM Trans. On Internet Technology*, Vol. 3, 1 (Feb., 2003): 1-27.