A Context-Aware based Dynamic User Preference Profile Construction Method

JeeHyun Kim
Department of Computer Software
Seoil University
Seoul-si, Korea
jkhim@seoil.ac.kr

Qian Gao
School of information
Shandong Polytechnic University
Jinan, Shandong Province, China
gq@spu.edu.cn

Young Im Cho
College of Information Technology
University of Suwon
Hwaseong-si, Gyeonggi-do, Korea
ycho@suwon.ac.kr

Abstract—This paper proposed a new horizon for the personalized information retrieval which can use image input instead of text query to express the user’s intention. Furthermore, this paper proposed a novel dynamic user preference profile construction method which can comprehensively track the users’ local device operation behavior and browsing behavior by using the context information (interactive historical information and user information related with the retrieval) which are stored and adopted in all of the smart devices owned by the same user (such as documents, E-mail, picture and so on) to grasp the relevance between the different documents more accurately.

Simulation used the dataset of a branching factor with a depth of four levels in the hierarchy of the Open Directory, and the simulation results show that the proposed method can meet the user’s intention much better and achieve higher precision ratio and recall ratio than other knowledge or context based personalized information retrieval systems.

Index Terms—Context-Aware, Dynamic, Behavior, Weighted

I. INTRODUCTION

We live in an era loaded with the enormous amount of information, and it is shown by statistics from authoritative organizations that there are more than 9 billions of publicly indexed Web pages, as in [1]. The Internet has become the main source for people to obtain and exchange information. The key issue of the nowadays information retrieval system is to take the user’s own situations or circumstances into account. Put another way, it should provide personalized retrieved information according to various demands of different users.

But nowadays it is difficult to effectively gather information and retrieve the most relevant documents on the topic of interest from the Web because of the large amount of information in all types of formats. Researches devote themselves to modeling a dynamic user profile to record the user’s preference, but in most cases they focus on the tracking of the user’s browsing behavior, instead of capturing the user’s local computer operation behavior. Beside, traditional ontology based personalized information retrieval system normally classify the concept on the basis of the general domain ontology, which cannot indicate the difference between individuals. Furthermore, sometimes, the traditional text query cannot express the exact intention of the user, even an expert may not use limited text query to express their meaning, while the query that is prone to ambiguity can be elaborated clearly by some unstructured data, such as pictures, video. Therefore, some methods that can analyze and understand other kinds of data need to be researched.

Aiming at those problems, in this paper, we proposed a novel dynamic user preference profile construction method which can comprehensively track the users’ local device operation behavior and browsing behavior by making use of the context information (interactive historical information and user information related with the retrieval) which are stored and used in all of the smart devices owned by the same user (such as documents, E-mail, picture and so on) to grasp the relevance between different documents more accurately.

The rest of this paper is organized as follows: Section 2 gives a short review of the related work. Section 3 describes the framework and the realization of our system. Section 4 gives the simulation based on our proposed method and we arrive at the conclusion in Section5.

II. RELATED STUDY

One of the most promising methods of realizing the personalized information retrieval is to create the user preference profile. User profiles may include demographic information such as name, age, country, education level, etc, and may also represent the interests or preferences of either a group of users or an individual. Personalization of Web portals, for example, may focus on individual users.

In order to build a user profile, some source of information about the user must be collected through direct user intervention, or implicitly, through agents that monitor user activity. Although profiles are typically built only from topics of interest to the user, some projects have pursued the method to include information about non-relevant topics in the profile, as in [2] [3]. In these approaches, the system is able to use both kinds of topics to identify relevant documents and discard non-relevant documents at the same time. Hence in general, a good user preference profile should comprise the results of lexical analysis, the input query, the documents clicked by the user, the queries by the user in the past, and some weight values. However, a user preference profile that includes incorrect user preferences only brings about trouble.
to users. Incorrect user preferences are generally obtained by the static profile approach. In this static profile approach, preferences or weight values are static and do not change once the user preference profile is created. In contrast to static profiles that maintain the same information over time, dynamic profiles can be modified or augmented, which takes time into consideration and can differentiate between short-term and long-term interests, as in [4] [5]. Short-term profiles represent the user’s current interests whereas long-term profiles indicate interests that are not subject to frequent changes over time.

User profiles are generally represented as sets of weighted keywords, semantic networks, or weighted concepts, or association rules. Keyword profiles are the simplest to build, but because fundamentally they have to capture and represent all (or most) words by which interests may be discussed in future documents, they require a large amount of user feedback in order to learn the terminology by which a topic might be discussed. Semantic user profiles have an advantage over keyword-based profiles because they can explicitly model the relationship between particular words and higher-level concepts. Thus, they can deal more effectively with the inherent ambiguity and synonymy of natural language.

### III. PROPOSED METHOD

#### A. Framework of the Context-Aware based Dynamic User Preference Profile Construction Model

This paper contains two main functions. The first one is the image annotation and sentences generation function which is for the purpose of analyzing the meaning of the image so as to realize the personalized information retrieval according to the users input image query. The second one is the dynamic user preference profile construction function for the purpose of collecting the user’s local access behaviors and browsing behaviors considering both the long term behavior and short term behavior to track the frequently changing user interest.

The framework and workflow of the proposed method are shown in Fig.1. When one uses the proposed information retrieval system, he/she can choose two kinds of query: one is image query, and the other one is text query. If the user inputs an image query, the image query will be processed by the content based image annotation agent, which can analyze and convert the image into text annotation to understand the meaning of the image and utilize the created annotation as the keyword to search the related information. The user’s profile, which was constructed based on the domain ontology, combines local information context aware method, browsing behavior information context aware method to track the user’s daily behaviors, then realizes the dynamic updating of the user preference profile. Furthermore, in order to analyze the content of the image stored in the image database, this system adopted a sentence generation agent to extract the position information among the images insomuch as to generate the natural language automatically. Based on these natural languages, the system can find which images stored in the image database are most relevant with the input images query.

![Diagram](image.png)

**Fig.1 Framework of the context-aware based dynamic user preference profile construction model**

#### B. Client Agent

A person often owns more than one intelligent device, such as a desktop computer, notebook, smart phone, and tablet. It is not enough to judge an individual’s preferences merely on the basis of the personal information stored on a single computer; instead, a user preference profile strategy should incorporate all the devices of a user in order to determine his preferences.

Hence Client Agent is used to verify user identities by establishing a union user account and to monitor whether a device is prepared to implement a user preference profile creation task.

Every day, the device tracker runs a simple loop that periodically (for example every five minutes) sends heartbeat...
method calls to every device belonging to user j. The heartbeat obtained from every device informs the device tracker that a device is at work (the heartbeat also doubles as a channel for messages). If the device is at work, its Active Degree (in the form of “UserID.ADj,” where j is the index of the device) will be set to “+1,” which is used to judge the activeness of the device. As part of the heartbeat, a device tracker will tell whether a device is ready to implement a user preference profile creation task, and if it is, the device tracker will tell the User Preference Profile Construction Agent

(1) Method for User’s Identity

In this paper, two option methods are adopted to identify the users: one is cookies, and the other one is logins. The former one is for current sessions, and the latter is an option for users who choose to register with a site. If a user has only one computer or he doesn’t want to set up a union account number, then when his browser client first connects to the system, a new user id is created. This id is stored in a cookie on the user’s computer. When one revisits the same site from the same computer, he uses the same user id. This places no burden on the user at all. However, if the user uses more than one computer, each location will have a separate cookie, hence a separate user profile. Also, if the computer is used by more than one user, and all users share the same local user id, they will all share the same, inaccurate profile. Finally, if the user clears their cookies, they will lose their profile altogether, and if users have cookies turned off on their computer, identification or tracking is not possible.

This thesis adopts an optional method based on logins. If the user has more than one computer, it is better to create an account via the registration page, and login and logout every time he visits the site. Once the user identifies himself during login, the identification is generally accurate, and the user can use the same profile from a variety of physical locations.

(2) Method for Local Data Collection

The metadata includes file name, creating time, modification time and access time which can be obtained by the win32 API offered by the windows system. The Local user behavior includes duration information and application program windows switching information which can be obtained by using open source tools “User Activity Logger” as in [6], to record the users’ daily working behavior on the personal computer. Here, we adopt the “Windows Hooks” mechanism and the kernel mode driven method to acquire the interactive information between users and the application programs. User Activity Logger employs log file to record the users’ interactive information obtained by the “Windows Hooks” mechanism and kernel mode-driven method.

(3) Automatic Image Annotation Agent

In this paper, we adopt the approach proposed by Xi and Cho, as in [7] which generated sentential annotations for general images. This method is feature weighted according to the statistical distribution of the features when clustering the image region so as to avoid the clustering algorithm being dominated by weakly relevant features, which improves the clustering accuracy.

(4) Natural Language Sentences Generation Agent

After obtained the annotated images, we use the statistical generative model to generate sentences to describe image content based on the annotated images, as in [8]. Given a training set of images with annotations, we parse the image getting position information, then, we use Machine Learning to get the probabilities of combinations between labels and prepositions, obtain the data to text set, create triple <keyword1, pre position, keyword2>. Finally we generate sentences from the xml report.

(5) User Preference Profile Construction Agent

(1) Static User Preference Profile Construction Agent

In the beginning, when a user using his browser client to first connect to our proposed system or creating an account via the registration page, the statistic user preference profile will be initiated by the weight a rating scale between 0 and 1 which was gathered explicitly during signup stage. Sum of the weight of all categories within a particular domain will be 1.

(2) Dynamic Update Subagent

Dynamic user profile is used to address users’ frequently changed interest which is based on the context information according to the user device documents (e.g. reading duration information, file content and metadata, file content and metadata, application programming switching information, the time spent on a document stored in the users’ device, switching frequency and their combination), which has a strong positive relationship with users’ interest.

In this paper, we use three stages to update the user preference profile dynamically. First, we initialize the user preference profile according to the use’s inputted interred information which was obtained from the login information. Then we update the weight of each term in the user preference profile based on the domain document profile, the context information about the user device document, the feature words, the users’ local behavior and browsing behavior.

a. Local Behavior based Updating

Two kinds of context information—metadata and local user behavior were used to weigh the user device document and each concept in it.

We use (1) to calculate the time weight of a user device document:

\[
CRT(Di) = \frac{1}{\log(1.6 + \frac{\text{CurrentTime} - \text{CreatingTime}}{3600})} + \beta \frac{1}{\log(1.5 + \frac{\text{CurrentTime} - \text{AccessTime}}{3600})} + \gamma \frac{1}{\log(1.4 + \frac{\text{CurrentTime} - \text{ModificationTime}}{3600})}
\]

Here, \(\alpha, \beta, \gamma\) are three constant parameters, which was satisfied \(\alpha + \beta + \gamma = 1\), and the value of \(\alpha, \beta, \gamma\) will be obtained by the simulation.

Local user behavior includes duration information and application program windows switching information. The reading time refers to the time spent on browsing the local document during this period, existing click and stroking of the mouse and the keyboard. The switching frequency refers to
the total times of switching from the local document windows to other application program windows (local or webpage) sequentially.

When a user browses a document for a period of time, if he sequentially switches to another same window for a few times, it means that the content of the two windows is closely correlated. Furthermore, the total number of a window has been switched, which displays the significance of this document for its user. The more documents have been switched, the more attention the document can get from the user. Since the duration time is also an important affecting factor, we use (2) to define the switching frequency of document $d_i$.

$$SF(d_i, CN_i) = \sum_{k=1}^{n} dsw(d_i, window_k, t)$$

Where $dsw(d_i, window_k, t)$ indicates the switching frequency of document $d_i$ directly switching to the windows $k$ in the period of $t$, $t$ is a constant that can be decided by the user, and $SF(d_i, CN_i)$ indicates the switch frequency of document $d_i$ directly switching to all of the other application program windows.

Then we make use of the switch frequency to calculate the similarity of two documents, as in (3). Based on the equation, we calculate the behavior weight of the document, which can reflect how important the document will be, as in (4):

$$Sim(d_i, d_j) = \frac{dsw(d_i, d_j)}{RT(d_i, d_j)}$$

Where $RT(d_i, d_j, t)$ indicates the total reading time before the $d_i$ was switched to $d_j$ in the period of $t$.

$$BWweight(d_i) = \frac{1}{n} \sum_{k=1}^{n} Sim(d_i, d_k) \times SF(d_k, CN_k)$$

Where $n$ indicates the total number of child nodes, $d_k$ belongs to the child nodes.

Then we combine (1) and (4) to get the local behavior based weight of the document $i$, as in (5):

$$LBW(d_i) = \phi \times BWweight(d_i) + \theta \times CRT(d_i)$$

Where $\phi$ and $\theta$ are two constants satisfying $\phi + \theta = 1$, and the value of $\phi$ and $\theta$ will be obtained by the simulation.

Since the activity of the smart device can also affect the importance of the document (That is to say, if the smart device is rarely used, then the documents stored on this device is less important than the documents stored on the other frequently used smart devices), finally, we combine the Active Degree obtained by the Client Agent to refine the weight of each concept in document $i$, as in (6):

$$RLW(d_i) = \sum_{term} Active Degree \times LBWt(d_i)$$

Here $RLW(d_i)$ is a vector space model that represents the weight of each concept in document $i$. $LBWt(d_i)$ is the weight of the each term that belongs to concept $i$ which is calculated according to the context information obtained in $i^{th}$ smart device.

b. Browsing Behavior based Updating

First of all, the stemmed words were extracted from the web site by performing Words Stemming, Part of Speech (POS) [SNLPG] and Stop Words Removal. Then for each stemmed word we adopt the “Semantic Similarity Matching Algorithm” as in [9] to look up the most similar concept between the stemmed words set $V(wo_1, wo_2, \ldots, wo_n)$ and the corresponding domain ontology. For each stemmed word, if it has no similar concept or illustration term in the corresponding domain ontology, then discard this word from the feature word set, otherwise choose the most similar concept from the corresponding domain ontology, and insert the top M related terms $(t_1, t_2, \ldots, t_m)$ into the feature words set.

When a user clicks a web page, it means that the user may have interest in this web page. However the extent of a user’s interest to this web page is dependent on the series of behaviors after opening the web pages. We comprehensively calculate the weight of the feature item according to the term weight $\omega_i$ calculated by the local behavior based updating subagent, Term Frequency (TF), and browsing behavior, as in (7):

$$\omega(t_i, wp_j) = \gamma \times \frac{TF(t_i, wp_j) \times \log(1 + \max(TF(t_i, wp_j)) \times \text{average}(RLW) \times \delta}{\sum_{t_j \neq wp_j} \sum_{t_j \neq wp_j} \log(M) \times M}$$

Where, $TF(t_i, wp_j)$ is the frequency of feature concept $t_i$ that appears in the web page $wp_j$, $WP$ is the collection of all the web pages, $M$ is the number of all the web pages, $M_t$ is the number of web pages including concept $t_i$, $\text{average}(RLW)$ is the average weight of concept $i$ in all the documents that include concept $i$. $\gamma$ is a constant modifier used to make the value of $\omega(t_i, wp_j) < 1$, $\delta$ is the interest degree calculated by the users’ browsing behavior, as in (8):

$$\delta = \begin{cases} 
0 & \text{if } \frac{T_{\text{reading}}}{N_{\text{total}}} < \text{Thr} \\
0.5 & \text{if } \frac{T_{\text{reading}}}{N_{\text{total}}} \geq \text{Thr} \\
1 & \text{if user save or print or collect}
\end{cases}$$

Where $T_{\text{reading}}$ is the time used to read the web page $p_i$, $N_{\text{total}}$ is the total number of the words in the web page, and Thr is the threshold which can be obtained by the average reading time and simulation.

For each web page with the term $t_i$, we use (7) to calculate the weight $\omega(t_i, wp_j)$, and then choose the maximum one as the weight of the feature word $t_i$ as in (9):

$$\omega_h(t_i, wp_j) = \max_{wp_i \in WP} (\omega(t_i, wp_j))$$

IV. SIMULATION RESULTS

We constructed an HDFS Cluster to simulate our
proposed method (a single computer with an Intel Core4 CPU, 4 GB RAM as NameNode, and 5 computers with Intel Core2 CPUs and 4 GB RAM as DateNode). We used Java as the programming language and the Eclipse integrated development environment. In this study, we analyzed and operated the ontology files using an open-source framework “Jena.”

We use a branching factor of double with a depth of four levels in the hierarchy. The experimental data set contained 268 concepts in the hierarchy and a total of 2,829 documents indexed under various concepts.

The indexed documents were pre-processed and divided into two separate sets including a training set, and a test set. The training set was utilized for the representation of the personalized ontology profile, and the test set was utilized as the document collection for searching.

The training set was consisted of 1514 documents used for the one-time learning of the personalized ontology profile. The concept terms and corresponding term weights were computed using the formulas described in section 4.5. A total of 1315 documents were included in the test set (including 1150 documents and 165 pictures), which were used as the document collection for performing our search experiments.

The user data is collected and analyzed using the above proposed method on a daily basis for one month. The test set documents were originally indexed under a specific domain, and all of its concepts were treated as relevance documents for that domain whereas all other test set documents were treated as non-relevance.

The entire process consisted of determining profile stability first and then building 2 profiles: an initial profile, a profile built after 20 days of browsing after reaching stability (profile for several category combined together), as shown in Fig. 2 and Fig.3.

In the initial, since the user preference profile is not perfect, it cannot reflect the user’s intention very accurately. As a result, if the user’s input query term has ambiguity, some irrelevant websites may be retrieved, as shown in Fig. 4(The red one is the relevance returned results, the green one is the irrelevance returned results). We can see that the returned results include the animal domain and the computer domain, which cannot reflect the user’s daily interest.
But along with the time, more local document reading behavior and browsing behavior were collected and analyzed, so that the user’s long term preference profiles were trending towards stability. Therefore, even if the text query is somewhat ambiguous, the searched information is closer to the user’s real intention, and much more relevant information will be fed back, as shown in Fig.5. As can be seen that the returned results show high interest of the user on the user’s interest domain.

We use the evaluation method “the precision averages at 11 standard recall levels (11SPR)” to evaluate the performance of the proposed system, and the experimental 11SPR results are plotted, as shown in Fig. 6. The 11SPR curves show that my retrieval system based on the user preference profile is the best, followed by the Ontology model as in [10], TREC model, the web model, and the Category model.

![Fig.6 The 11SPR results of the proposed method versus other 4 approaches.](image)

The overall results show that our approach is better than the other three approaches in terms of the “precision averages at 11 standard recall levels”.

V. CONCLUSION

This paper investigated the realization of personalized information retrieval for text information and image information by constructing context information based dynamic user preference profile, and was simulated using the dataset of a branching factor with a depth of four levels in the hierarchy of the Open Directory.

First, we proposed a new horizon for the personalized information retrieval which uses image input instead of text query to express the user’s intention. If the user cannot use text query to express their intention more clearly, it can use image query instead of text query. In order to realize this purpose, this thesis proposes a novel system which generates the sentential annotations for general images.

Then the proposed system constructed a dynamic user preference profile based on the domain ontology and context information in local device document and browsed website pages, which can comprehensively track the users’ long-term behaviors and short-term behaviors.

Simulation shows that the proposed algorithm can increase the precision and flexibility as well as reduce the calculation amount compared with other knowledge structured based user preference profile construction method and context-aware based user preference profile construction method.
REFERENCES


First J. Kim (02-27-1956) received a Doctor of Computer Science from Dankook University in Seoul, Korea in 2004, her M.B.A degree in the Information Management from Dankook University in 1994, her B.S degree in Mathematics from Ewha Womans University in 1978.


Second Q. Gao (07-26-1978) is a PhD. candidate in the Department of Computer Science at the University of Suwon in Suwon city, Gyeonggi-do of Korea, she received her M.S. degree in computer application technology from Shandong University in Jinan city, Shandong province of China in 2008, and her B.S. degree in computer science and technology from Shandong University of Science and Technology in Taian city, Shandong Province of China in 2001.


Third Y. Cho (08-21-1963) is received a post-doc. degree at the University of Massachusetts in USA in 2000, her Ph.D. degree in Computer Science from Korea University in 1994, her B.S. degree in Computer Science from Korea University in 1988, and her M.S. degree in Computer Science from Korea University in 1990.


She has received many awards from International Society like Control Robot and System Society, Korea Intelligent System Society etc. Also, she got award from the Korea government as a Korea the man of merit particularly in Informatization area in 2014.