Journal of Theoretical and Applied Information Technology
15th November 2012. Vol. 45 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A USER INTERESTS MODEL BASED ON ONTOLOGY

CUNCUN WEI

Faculty of Engineering, Zhejiang Business Technology Institute, Ningbo 315012, Zhejiang, China

ABSTRACT

In this paper, the personalized service on the basis of lack of semantic and the strengths and weaknesses of the existing user model update mechanism, combined with keyword-based user interest ontology semantic information-rich features, design a classification knowledge base, combined with keywords and user interest ontology concept representation. The model of the user's interest is divided into short-term interests and long-term interest, and the two interested in using different establishment and update mechanism. By constantly updating and optimization model, the model can accurately reflect the characteristics of the user's interest. Finally, the effectiveness of the method is verified by experiments, experiments show that the model can improve the performance of user interest modeling.

Keywords: *User Interests Model, Ontology, Semantic Concepts, Personalization, Long-Term Interest, Short-Term Interest*

1. INTRODUCTION

The personalized services such as information retrieving and information recommendation etc., will focus on describing the characters of user and resource according to the keywords and realizing the corresponding relationship between user and resource information by keywords matching. But there will be some problems such as resource overloading and information confusion due to the personal characters of the Internet such as openness, dynamics, in-organization and non-semantic. Therefore, it has become a new direction and researching the focus of searching technology about how to improve search precision and provide personalized service to users. In various personalized service system, the most important basic task is user profile mining and profile modeling. It is only through mining learner's interests, establishing reasonable models to describe and manage the use's interests, constantly updating and maintaining, and gradually optimizing the models to respond user's interest requirement accurately to provide personalized analysis data for the personalized resource service.

The paper combines the ontology and concept space, indicates the feature items of user interests with semantic concepts[1], calculates learner's interest-level to the topic through establishing the word frequency and utilize the suitable calculation methods, mining the concepts within the user's feedback files and the relationship between concepts, combines user's short-term interests and long-term interests to create user interests model with semantic concept hierarchy tree and embody the drifting of user profile and improves and completes the user interests model consistently on the related feedback mechanism.

2. USER INTERESTS ACQUISITION AND EXPRESSION

A. User Interests Acquisition

In order to realize the personalized service to various internet resources recognized by URI and establish the suitable user profile models, we must search and mining all the resources and information related to the user profiles as much as possible. There are many approaches such as analyzing Favorite or Bookmark, analyzing the browsed website history records and access times, analyzing the user's browsing operation like saving, printing, editing and copying etc. and analyzing the mouse moving, scrollbar moving and browsing time.

B. User Interests Mining Base On Semantic Concepts

The mining process of user's personalized information is the process to find out a user's interest sub-tree from the interest semantic concept tree, indicated from Figure 1. The aim of semantic concept tree [2] is to describe and store accurately the field knowledge points and the relationship between knowledge.

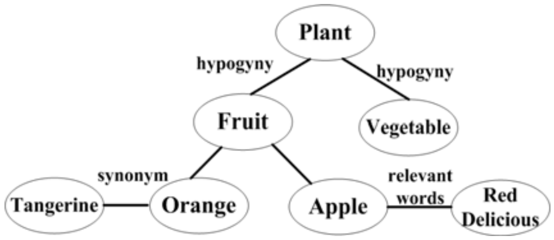

Figure 1: Semantic Concept Tree

The semantic concept tree is a classification tree [3]. The first level is the top level with the more general concept that indicates an independent area or topic. The above levels are gradually refined. For example, the plant is divided into fruit and vegetable. Except level relationship, there are also other various relationships between concepts. In order to indicate the relationship between concepts, lateral relationship will be added based on the tree-shape models to connect each independent concept such as 'orange' and 'tangerine brain' are the synonymous relation. These concept points connected with lateral relationship can be any joint on any level to create a semantic concept tree [4].

Word segmentation is necessary to extract the file's concept. The best reversed matching will be used when processing the word segmentation and then word selection will be processed according to part-of-speech rate selection or frequency selection. This model keeps nominal and calculates the weight of a word by TF algorithm [5].

As to the similarity level between file D and concept C, we may filter the words that do not appear in the characterized vector list [6]. If the interest character does not show on the file, the weight is 0. The file can be indicated as a weight vector D_k, $D_k=[W_1,W_2,W_3\ldots W_n]$, the similarity level between D_i and C_i can be calculated by the following format:

$$\text{Similarity}(D_k, C_i) = \frac{\sum_{i=1}^{n} nwd_{jk} * nwc_{ij}}{\sqrt{\sum_{j=1}^{n} nwd_{jk}^2 * \sum_{j=1}^{n} nwc_{ij}^2}}$$

(1)

nwd_{jk}, nwc_{ij} indicate the normalized weights.

The number of nonzero concept joints tends to be stable with the increase of the classified files. The interest model is gradually convergent and extraction process ends.

C. Importance Degree Calculation Of Semantic Interest Concept

We might get a concept set $\{C_1, C_2\ldots C_k\}$ which C_i is the word set $\{W_1, W_2, \ldots, W_n\}$ after combining and classifying all the listed words in the file by the above procedure [7]. This paper defines the appearance times of each concept in the awaiting treatment files as the appearance frequency of concepts. The appearance frequency F (C_i) of concept C_i is. F (W_n) indicates the appearance frequency of word W_n in the paper. The concept appearance frequency reflects the importance level of each concept in the file through the appearance frequency of words with the same meaning and different form. In order to express the semantic concept of the file accurately, new concepts set $\{n_1, n_2, \ldots, n_k\}$ will be composed by the concepts which are more important than thresholds and extracted from the concept sets through setting responsible thresholds.

3. CREATION AND REALIZATION OF USER INTERESTS MODEL

Memory can be divided into long-term memory and short-term memory. The change of the user profile is called amnesia that also can be divided into long-term and short-term.

Therefore the user profile model can be indicated by both short-term and long-term methods with the different drifting strategy

A. Creation Of Short-Term Interest Model

Short-term interest is created during the process of utilization of the personalized retrieval system. The short-term interest will be calculated by classifying the search conditions and a user's interest file, statistics the keyword frequency as the short-term interest level for the keyword and referencing the classified knowledge database.

B. Creation Of Long-Term Interest Model

The creation process is shown as below:

(1) Capture the related information of user's interest.

(2) Classify the words of information and statistics the frequency F_i of appearance of each keyword T_i.

(3) Add the keyword and frequency into the long-term interest table one by one, take the frequency as the learner's interest level and setup the interest creation time as the present time.

(4) Calculate the interest level V of each topic Node (B) by referencing the knowledge tree of classified knowledge table from down to up. K is the number of keyword joints belonging to this topic and F_i is the interest level of the keyword joints.

(5) Record the calculated topic interest level, belonged field and present time to the topic interest table.

C. The Drifting Strategy Of Short-Term Interest

Short-term interest is user's present interest. It is short and active. The update procedure is required to be responded rapidly by the sliding time window strategy. The calculation format is:

$$U_{t_k}^{cur} = \frac{1}{M + S_j} \sum_{j=1}^{n} \frac{1}{s_j} \sum_{i=1}^{S_j} \omega(t_k, p_i) \quad (2)$$

M is the size of time window, S_j is the browsed

website number in day j.

D. The Drifting Strategy Of Long-Term Interest

User's long-term interest is user's fixed interest. It is oppositely stable. The user's interest will be drifted when the user has the new one. Normally it won't affect the user's long-term interest. The user's long-term interest is calculated by giving the different weights to the user's short-term interest in the different period according to the time order. The format is:

$$F^j(k) = \frac{1}{S_j} \sum_{i-0}^{n} \gamma \rho_i(k, \tau) \quad (3)$$

$F^j(k)$ is long-term interest of C set; $\rho_i(k,\tau)$ is number I short-term interest of C_j set; γ is forgetting factor. User's shift speed will be various according to the forgetting factor a. User's interest shifts faster if a decayed faster and the affection of long-term interest by the short-term interest will be bigger.

4. UPDATE AND OPTIMIZATION OF USER INTERESTS MODEL

Author names and affiliations are to be centered beneath the title and printed in Times 12-point, non-boldface type. Multiple authors may be shown in a two- or three-column format, with their affiliations italicized and centered below their respective names. Include e-mail addresses if possible. Author information should be followed by two 12-point blank lines.

A. Short-Term Interest Model Update

Short-term interest is mainly the user's present interest. And it's update can be processed by interest combination method.

B. Long-Term Interest Model Update

This paper updates the user interests model by the gradual forgetting method. If the user's interest forgetting complies with the normal brain forgetting rule, which means user's interest will be graded reduced by time and the speed will be repaid first and then slow. User's present multi-access words can most represent the user's present interest. While those words without update for a long time can be filtered finally.

The paper introduced the concept of forgetting factor to forget the user's long-term interest gradually and accept the drift of user's interest.

The paper calculates the forgetting factor F (x) [8]. The format is:

$$F(x) = e^{\frac{in^2(pre-est)}{h1}} \quad (4)$$

pre indicates the present date, est indicates the date when the interest character word appear in the model first time, hl indicates the half life period that means the user's interest will be forgotten for half after hl days.

C. Short-Term And Long-Term Interest Transformation

The calculation format is shown as below:

- The system will inspect the short-term interest keyword periodically.
- The keyword backup will be inserted to the long-term keyword interest model one by one. If the keyword exists in the model, go to (3), otherwise go to (4)
- Set the new interest weight of long-term keyword by adding the keyword weight in the short-term interest model and the weight in the long-term interest model [9], go to (5).
- Add the keyword in the short-term interest model directly to the long-term interest model and set weight again.
- The related keyword will be cancelled in the short-term interest model.
- The interest level V of each topic Node (B) in the long-term interest model will be

calculated by format $V=\sum_{i=1}^{k} F_{i,}$, K indicates the number of keyword joints belong to this topic, F_i is the interest level of the keyword joint.

- Long-term topic interest table will be updated.

5. EXPERIMENT AND ANALYSIS

The aim of this experiment is to verify that the ontology-based representation method is more accurate to express user's interest than keyword based representation method. The experiment will draw a conclusion by comparing the results of retrieval information from the two methods.

The documents selected from xinhua.net are used as the testing database. The database contains 1000 documents concerning with 10 topics (athletics, economics and science and technology). Each topic includes 100 documents and will be tested as user's required field respectively. The content annotation and subject will be extracted as the character sets. Then IF/IDE algorithm is accepted to execute feature selection base on text.

The paper quantitatively analyzes the advantages and disadvantages of the traditional user's model with precision ratio and recall ratio. Precision ratio and recall ratio are evaluated by harmonic average F [10].

Recall ratio：R = a/c,

Precision ratio：P = a/b

harmonic average F=2*Precision*Recalll (Precision-Recall)

Letter a indicates the number of related documents located before d_i; b indicates the number of all documents located before d_i, c indicates the number of all related documents in the set.

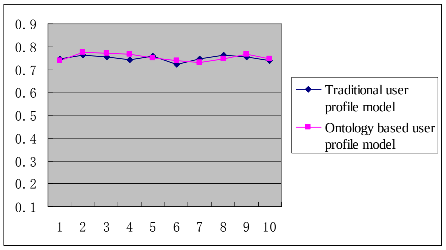

Figure 2: Comparing Precision Of Two Models

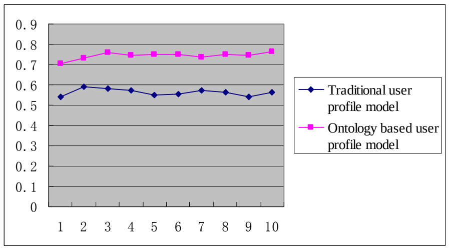

Figure 3: Comparing Recall Of Two Models

The experiment results showed in table 1 by comparing traditional user interests model and Ontology-based user interests model.

Figure 2 and figure 3 indicates that the precision ratio of the experimental result decreased while the recall ratio and the harmonic average increased significantly under the same user's conditions from the above data.

Table 1 : The Experiment Results

		Traditional user interests model			Ontology-based user interests model		
Id	Item	precision	recall	Harmonic average	precision	recall	harmonic average
1	Politic	0.748	0.543	0.629	0.737	0.703	0.720
2	Economy	0.761	0.592	0.666	0.775	0.734	0.754
3	Culture	0.753	0.584	0.658	0.769	0.757	0.763
4	IT	0.742	0.573	0.647	0.767	0.745	0.756
5	Education	0.759	0.548	0.636	0.752	0.75	0.751
6	Military	0.721	0.556	0.628	0.739	0.749	0.744
7	Medical	0.746	0.572	0.648	0.729	0.738	0.733
8	Agriculture	0.761	0.565	0.649	0.747	0.748	0.747
9	Law	0.756	0.542	0.631	0.768	0.746	0.757
10	Art	0.737	0.563	0.638	0.746	0.762	0.754

To the system of provision of personalized service, Recall ratio reflects the capacity of provisions related resource to users by the system as well as the personalized requirements. In this regards, it is more obvious superiority by ontology-based representative method comparing keyword based representative method. And it improves performance of user interests model creation by ontology-based user interests model comparing with the traditional methods.

6. CONCLUSION

Based on analyzing and comparing the existed various user interests model creation methods in the resource characterized searching system, the paper introduced a new user's interest expressing method combining the keyword and topic concept based on the classified knowledge database, and discuss the creation and update process of user interests model in detail. The model can reflect the user's interest characters accurately by the above creation and update mechanism. The experimental result indicates that this method improved significantly in efficiency comparing with the traditional keyword matching method. Besides, the method still needs to be perfected deeply under the following aspects: (1) How to utilize the machine learning method to capture the user's interest and related feedback and study the formation and application of user interests model from the view of syntax and semantics. (2) Automatic extraction of resource ontology description. (3) Set more completed regulation during the information provision.

ACKNOWLEDGEMENTS

This work was supported by Zhejiang province department of education scientific research achievements of the project in 2011 year (Y201121970) and Zhejiang province excellent young teachers of universities and colleges funding projects in 2010 year (The Zhejiang Department of Education (2010) No. 175, 304.); Ningbo educational planning issues of 2012 (YGH120).

REFRENCES:

[1] Philippe Ramadour, Myriam Fakhri, Corine Cauvet, "GO-SEM: a Goal-Oriented Method for Service Engineering", IJIPM: International Journal of Information Processing and Management, Vol. 2, No. 1, 2011, pp. 1-12.

[2] Kazem Alizadeh, Mir Ali Seyyedi, Mehran Mohsenzadeh, "A Service Identification Method Based on Enterprise Ontology In Service Oriented Architecture", IJIPM: International Journal of Information Processing and Management, Vol. 3, No. 2, 2012, pp. 65 – 77.

[3] Shuli Yuwen, Xiaoping Yang, "Standardizing the Medical Data in China", JCIT: Journal of Convergence Information Technology, Vol. 6, No. 8, 2011, pp. 107 -116.

[4] S Kiefer, J Rauch, R Albertoni, M Attene, "An Ontology-Driven Search Module for Accessing Chronic Pathology Literature", Lecture Notes in Computer Science, Vol.7046, 2011, pp.382-391

[5] Bo-Yeong Kang , Dae-Won Kim, Sang-Jo Lee, "Exploiting concept clusters for content-based information retrieval". Information Sciences, Vol. 170, No. 2, 2005, pp. 443-462.

[6] Castells. P, Fernández. M, Vallet, D. An, "Adaptation of the Vector-Space Model for Ontology-based Information Retrieval", IEEE Transactions on Knowledge and Data Engineering, Vol.19, No. 2, 2017, pp. 261-272.

[7] Vallet. D, Castells. P, Fernández. M, Mylonas. P, Avrithis. Y, "Personalised Content Retrieval in Context Using Ontological Knowledge", IEEE Transactions on Circuits and Systems for Video Technology, Vol.17, No. 3, 2011, pp. 336-346.

[8] Juan D Velásquez, Vasile Palade, "A Knowledge Base for the maintenance of knowledge extracted fromweb data", Knowledge Based Systems, Vol.10, No. 3, 2007, pp. 238-248.

[9] Zuo. L. D, Salvadores. M, Imtiaz. S. M. H, Darlington. J, Gibbins. N, Shadbolt. N. R and Dobree. J, "Supporting Multi-view User Ontology to Understand Company Value Chains". *SEMANTIC WEB - ISWC 2009, PROCEEDINGS*, vol.5823, 2009, pp. 925-940.

[10] Jinxi Xu, W Bruce, "Croft Improving the effectiveness of information retrieval with local analysis". ACM Transaction on information system, Vol.18, No. 1, 2000, pp. 79-112.