

An Improved Ontology-Based User Interest Model

Zhu Liang¹, Yan Jun¹, Ling Haifeng¹ & Qian Haibo¹

¹ Engineering Institute of Engineering Corps, PLA University of Science and Technology, Nanjing, China

Correspondence: Zhu Liang, Engineering Institute of Engineering Corps, PLA University of Science and Technology, Nanjing 210007, China. E-mail: javis@qq.com

Received: April 1, 2012

Accepted: April 26, 2012

Online Published: May 21, 2012

doi:10.5539/mas.v6n6p39

URL: <http://dx.doi.org/10.5539/mas.v6n6p39>

Abstract

In the personalized information retrieval, the design of user interest model is a key problem. Through analyzing the Ontology-based User Interest Model, propose a new hybrid model that contains both long-term and short-term model, and the long-term model updated from Vector Space Model by transform algorithm. Experiments showed that the new model tracked user's interests more accurately, and greatly avoided the Cold Start problem.

Keywords: user interest model, ontology, forgetting function, ontology project

1. Introduction

User interest model is not a general description of the individual user interest, but a method-oriented formal description with specific data structures (Huttenlocher D. P., Klanderman G. A., & Rucklidge W. J., 1993). User interest model is the basis and core of a personalized search engine and personalized recommendation system. A good user interest model can express interest of individual users, to improve search quality, and help the personalized recommendation system to provide the more accurate referral service. Therefore, the building of appropriate user interest model has become the key problem of personalization information service.

At present, the study of user interest model already transited from the traditional vector space model to ontology-based user interest model. For example the Quickstep system (Middleton S. E., Shadbolt N., & Roure D., 2004) uses a scholarly research subject ontology to express the field of interest to the user, the aceMedia system (Castells P., Fernandez M., & Vallet D., 2005) described the user interest characteristic through a ontology concept vector. The study on ontology-based user interest model has three shortcomings: (1) It does not distinguish between stochastic short-term interest and long-term subject. (2) It does not take into account the decay of the interest with the passage of time. (3) Cold start problem.

Because of the above shortcomings, this article designed an improved ontology-based user interest model (IOBUIM), and through experimental verification new model validity.

2. Ontology-Based User Interest Model

Ontology-based user interest model can be formally expressed as follows (Guan Q. Z. & Zhou Z. R., 2007; Pretschner A., 1999):

$$OBUIM = (PersonalI, PersonalO, InterestD) \quad (1)$$

Where $PersonalI = \{name, sex, birth, profession, hobby\}$ express the user basic individual information.

$PersonalO = \{C, R, H^c, Rel, A^0\}$ express the user personalization ontology, where C is concepts set, R is relational set, $H^c \subseteq C \times C$ express concept hierarchy, $Rel: R \rightarrow C \times C$ express non-classified relations between concepts, A^0 express axiom collection which uses some logical language.

$InterestD = \{ \langle concept, Degree \rangle \mid concept \in PersonalO.C, Degree \text{ are real numbers between 0 and 1} \}$ express the set of each concept in $PersonalO$ and it's interest degree.

OBUIM has strong adaptability, can reflect the current semantic concept hierarchy as well as user preferences, thus adapts to the changing environment, expresses the user's preferences accurately.

The insufficiency lies in:

(1) It does not distinguish between stochastic short-term interest and long-term subject. Human's memory divides into the long-term interest and the short-term interest. Regarding short-term interest, because the capacity is very limited, when the information cannot reappear quickly, it will be forgotten very quickly. Long-term interest will be forgotten gradually only because of long-term interest information not to use for a long time when the environment or the situation changes.

(2) It does not take into account the interest will decay with the passage of time. Forgotten on the things of interest is a gradual process, the importance is highest when the interest appears, with the lapse of time, the importance and the interest drops gradually.

(3) Cold start problem. The *PersonalO* is comes from the domain ontology partial mappings, but domain ontology does not have weight information, so the *PersonalO* relies on the individual model learning algorithm to express the user individuality interest accurately, but the existing learning algorithm has a long start-up period, in the start period, the precision is very low.

3. IOBUIM

This article makes three improvements in the OBUIM foundation:

(1) The user sustained long-term interest reflects the preferences of the user features, short-term interest corresponds to the user's random, temporary needs.

(2) User interest will decay over time, therefore, this paper applied the forgetting function in the model updating algorithm.

(3) To avoid the "cold start" problem, this article designed an upgrade algorithm which can transform a mature user interest model based on VSM to IOBUIM.

3.1 Structure of IOBUIM

As shown in Figure 1, this paper uses the hybrid model which composed together by the long-term model and the short-term model, the short-term model represent the user's short-term interest which obtained by user's short-term behavior observation; The long-term model represents the user's long-term preferences, obtained from development accumulation for a long time.

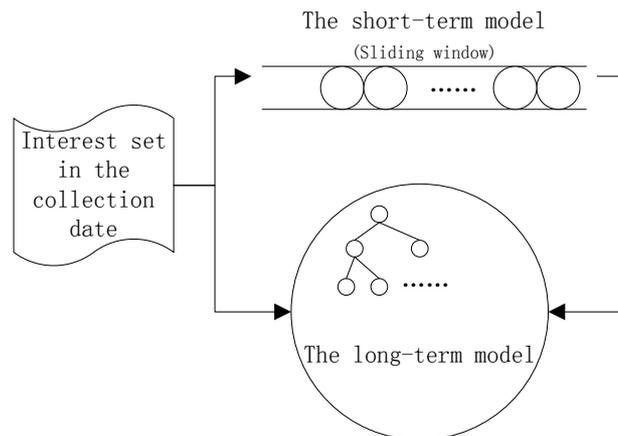


Figure 1. Structure of IOBUIM

The formal description is as follows:

$$IOBUIM = \left\{ \begin{array}{l} (PersonalI, PersonalO, InterestD, Time) \\ \{(t_1, num_1), (t_2, num_2), \dots, (t_n, num_n)\} \end{array} \right. \quad (2)$$

In the long-term model, the definition of *PersonalI*, *PersonalO* and *InterestD* are same to Eq. 1, $Time = \{ \langle concept, t_{last} \rangle \}$ was added to the long-term model to express the set of each concept in *PersonalO* and it's last time t_{last} which is paid attention by the user, preparing data for the calculation of the time decay function.

In the short-term model, $t_i (1 \leq i \leq n)$ express the i th short-term interest key word, num_i express the user browsing number of times of the i th short-term interest key word in a specific period.

3.2 Interest Update Algorithm

In order to enable the automatic control of the user interest characteristic information, and the auto-adapted dynamic change along with the user interest change, interest update algorithm needs to adjust the weight of user interest automatically (Qiu Xiaojun & Liu Fasheng, 2010). The short-term interest update algorithm uses the sliding window algorithm, and the long-term interest update algorithm uses the forgetting function.

The short-term algorithm flow is as follows:

Step 1: The personalized system obtains user interest vector $\{(k_1, \omega_1), (k_2, \omega_2), \dots, (k_n, \omega_n)\}$ using the TF-IDF technology.

Step 2: For $i=1$ to n , try to match the key words k_i in the long-term interest model, if successful, then turn to step 7.

Step 3: Matches the key words $k_i (1 \leq i \leq n)$ in the short-term interest model, if match, then num_i plus 1, and stop algorithm.

Step 4: If interest window is not full, new interest k_i added to the front of window, and stop algorithm. If the interest window is full, turn to step 5.

Step 5: Remove the interest t_n in the end of interest window. If num_n is not less than a certain threshold, turn step6, otherwise discard t_n . New interest k_i added to the front of window.

Step 6: Projects the new interest with the domain ontology to the *PersonalO*, calculates the weight, and stops algorithm.

Step 7: Updates the weight of interest and semantic close interest weight in the long-term model, and stop algorithm.

User interest will change with the lapse of time (Li Feng et al., 2008). Literature (Song Lizhe, 2006) proposed one kind of progressive linear forgetting function, but according to the psychology knowledge, human's forgetting rule is not the complete linear variation. Therefore this paper proposes the misalignment forgetting function based on the normal distribution density function, as below:

$$f(t) = \frac{1}{\alpha\sigma\sqrt{2\pi}} \frac{e^{-\frac{(t-t_{last})^2}{2\beta^2 t_{sec}^2}}}{e^{2(\beta\sigma)^2 t_{sec}^2}} \quad (3)$$

Where t is the current time, t_{last} is the last time of which is paid attention by the user. t_{sec} is the attenuation coefficient. Therefore, the larger t_{sec} , the more gentle decay curve. α, β are the adjustment factors, in this paper, we set $\alpha=0.52$, $\beta=2$.

3.3 Transform Algorithm

Vector Space Model (VSM) is by far the most popular user model representation and its formal description is:

$$VSM = \{(t_1, \omega_1), (t_2, \omega_2), \dots, (t_n, \omega_n)\} \quad (4)$$

This is an n-dimensional feature vector, each dimension vector is composed of a keyword t_i and its weight ω_i . The weight may take a boolean value or a real value, expressing the degree of some concept which the user is interested. Many systems used this kind of vector space model representation, such as Amalthea, IGIMA, Letizia, Personal Web Watcher, Syskill & Webert, Webmate, Web Watcher, Fab, Krakatoa Chronicle, MovieLens, WebCobra(Wu Lihua & Liu Lu, 2006).

In order to reduce the individual model cold start time, to avoid waste of resources available VSM model, this paper proposes a transform algorithm for upgrade existing VSM model to ontology-based model. The algorithm flow is as follows:

Step 1: Determines the location of each concept t_i in the *DomO*.

Step 2: Projects the concept correspondence's *DomO* subset to the *PersonalO*. The projection not only need project the key words to the *PersonalO*, moreover must project the semantic similar concept collection to the *PersonalO*. The semantic similar scope cannot be too big, so a concept proliferation threshold value is needed. As shown in Figure 2, the circle with oblique line represents with the match key words, the circle with grid represents the semantics proliferation key words through semantics similarity computation and the concept proliferation threshold value.

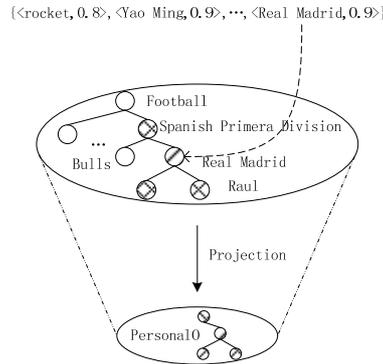


Figure 2. Match, concept proliferation and projection

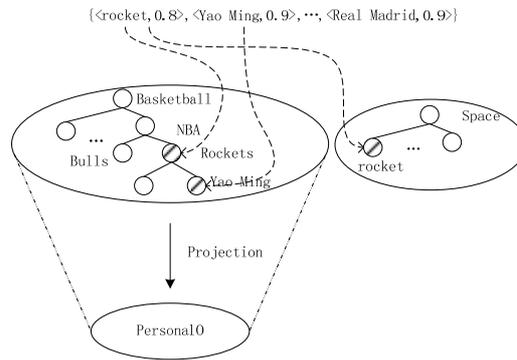


Figure 3. Solution of polysemy

Step 3: If a concept t_i corresponds to a number of $DomO$, such as the keyword "rocket" correspond to $DomO$ "Basketball", $DomO$ "Space" and so on. Because People's interest cannot be isolated, so the most similarity concept t_x can be found out in VSM through semantics similarity computation, and matched in $DomO$. The $DomO$ which contains t_i and t_x is the appropriate $DomO$ to project. For instance, there are keywords such as "Yao Ming" in VSM, then the $DomO$ "Basketball" will be projected at last, as shown in Figure 3.

Step4: Calculates the weight of concept in the $PersonalO$. $PersonalO$ not only contains the key words in VSM, but also contains many semantic proliferation key words which needs to give a corresponding weight. to each semantic proliferation key words, the computation of $Degree_j$ in accordance with the following formula:

$$Degree_j = \sum_{i=1}^n \omega_i \alpha_i^j \tag{5}$$

Where ω_i is come from VSM, α_i^j express the semantic similar degree between t_j in $PersonalO$ and t_i in VSM, $\alpha_i^j \in [0, 1]$.

4. Experimental Results

Laboratory tool is a self-built search engine by the Machine Learning and Pattern Recognition Group (MLPRG) of university. The experiment is divided into three parts, the first part applies IOBUIM, and the long-term interest model data transformed from original VSM, the second part applies OBUIM, and the model data transformed from original VSM, the third part with the new OBUIM. The experiment time length is one month, let attenuation coefficient $t_{sec} = 86400$ s, namely user interest decay according to the number of days.

Appraisal data not only has traditional recall and precision ratio, but also has joined the model size. Experimental results are shown in Figure 4, Figure 5 and Figure 6:

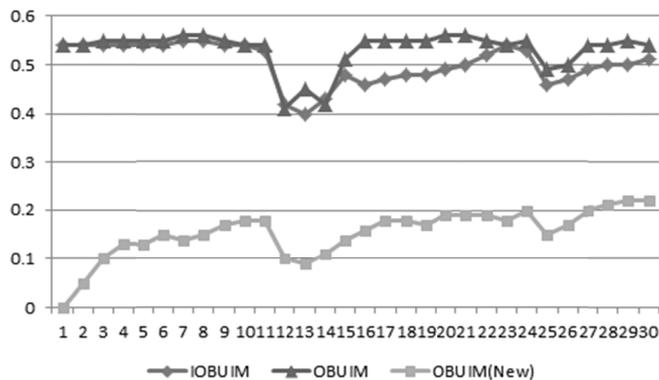


Figure 4. Recall ratio

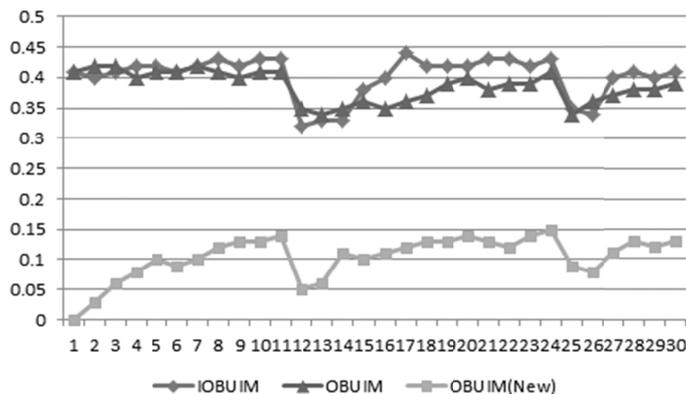


Figure 5. Precision ratio

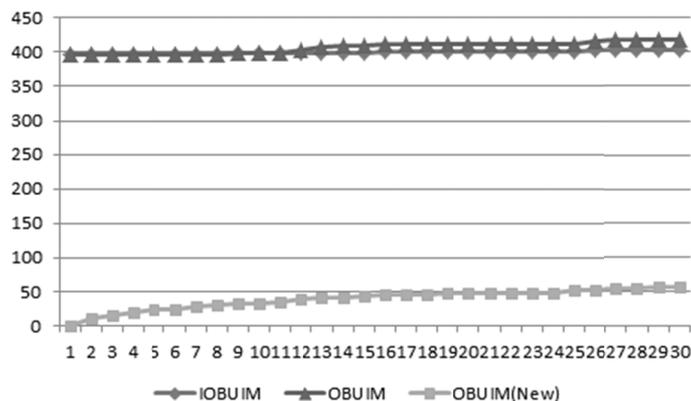


Figure 6. Model size

It can be seen from these three figures that after a month of study, on these three appraisal data of OBUIM (New) model in the third part have not achieved the extent of the first two parts, showing that the "cold start" will consume a long time.

Then mainly comparing first part model IOBUIM and second part model OBUIM:

(1) According to Figure 4, the first part and second part recall starting from a starting point, as time goes by, the second part of recall slowly was higher than the first part recall. When the user short-term interest change is quite specially quick, the recall disparity expands quicker between second part model and the first part model, when user short-term interest changes are quite steady, the first part model recall then slowly move closer to the second part model recall. This is mainly because of when the user short-term interest changes, the first part model is only puts this new concept in the short-term interest model, and no semantic concepts into the model, while the second part model directly projects the semantic related concept into the model to ensure the recall. After a period of time,

the new concepts in short-term interest model of first part will be projected into the long-term interest model through the *DomO*, therefore the first part model's recall can move closer to the first part model's.

(2) According to Figure 5, the precision of first part and second part model starting from a starting point, the precision of the first part model maintains stable, while the precision rate of the second part model gradually declined. This is mainly because the traditional OBUIM's update algorithm has not forgotten function, and affected the accuracy of the model. From the figure we can see that when the short-term interest dramatic changes, the first part model's precision ratio more quickly than the second part to adjust back to normal levels.

(3) It can be seen from Figure 6, the size of the first model grows slowly, the second part model's size grows quicker, moreover when the interest change is quicker, the growth is quicker.

General experimental conditions, the IOBUIM contrasts the OBUIM has some losses in the recall ratio, but the disparity is not big, moreover the IOBUIM fitting user interest change situation, and the precision ratio is higher. At the same time, because of the promotion algorithm, the IOBUIM has avoided a long time of the "cold start" problem. On the whole, the improved model IOBUIM has manifested certain superiority.

5. Summary

This paper has analyzed traditional OBUIM, pointed out that the traditional OBUIM existence insufficiency in the model structure and the update algorithm, pointed out the cold start problem. To solve these problems, IOBUIM has designed. The experimental result indicated that IOBUIM has manifested certain superiority.

References

- Castells, P., Fernandez, M., & Vallet, D. (2005). *Self-tuning personalized information retrieval in an ontology-based framework*. Proceedings of the 1st International Workshop on Web Semantics. New York: Springer Verlag.
- Chen, S. Y., & Wu, J. H. (2008). Ontology-based Concept Semantic Similarity Computation and its Application. *Micro Electronics & Computer*, 25(12), 96-99.
- Gauch, S., Chaffee, J., & Pretschner, A. (2003). Ontology based personalized search and browsing. *Web Intelligence and Agent System*, 1(3-4), 219-234.
- Guan, Q. Z., & Zhou, Z. R. (2007). Research of Ontology-based user model. *Computer Applications*, 27(10), 2504-2507.
- Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(13), 850-863.
- Li, F., Pei, J., & You, Z. Y. (2008). Adaptive user interest model based on the implicit feedback. *Computer Engineering and Applications*, 44(9), 76-79.
- Li, Y. H., Bandar, Z., & Mclean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4), 871-882.
- Middleton, S. E., Shadbolt, N., & Roure, D. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information System*, 22(1), 54-88.
- Pazzani, M., & Billsus, D. (1997). Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, 27, 313-331.
- Pretschner, A. (1999). *Ontology Based Personalized Search*. Lawrence, KS: The University of Kansas.
- Qiu, X. J., & Liu, F. S. (2010). Research on User Interest Model Based on Hierarchical Vector Space Model. *Modern Computer*, 16-19.
- Song, L. Z., Niu, Z. D., Yu, Z. T., Song, H. T., & Dong, X. J. (2006). A Method of Drifting User's Interests Based on Hybrid Model. *Computer Engineering*, 32(1), 4-6.
- Wu, L. H., & Liu, L. (2006). User Profiling for Personalized Recommending Systems-A Review. *Journal of the China Society for Scientific and Technical Information*, 25(1), 55-62.