*2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)*
***IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore***
*DOI: 10.7763/IPCSIT.2012.V51.94*

# Research on Interest Model of User Behavior

Yu Ye*, Guowen Wu and Xin Luo

School of Computer Science and Technology, Donghua University, Shanghai 201620, China,

**Abstract.** In searching for the websites and portal sites, user interest model is applied in describing the users' preferences. We summarized some popular methods of creating user interest model, such as creating models in explicit or implicit ways. On the analysis of users' browsing behaviors or content, we could extract the interests of users. An improved method is proposed to increase the number of web pages efficacious information containing, improve the accuracy of user interest model by introduce of search tails and so on.

**Keywords:** User model; hybrid interest model; VSM(Vector space model).

## 1. Introduction

Recently search engine rarely orders their results based on individual user interests. However different users have various interests and backgrounds, thus they should obtain the different feedback for the same query. For example: consider the same query "apple", the user has interest in life should obtain the related websites of the food, but the person who likes digital products, hopes to get the results of Apple digital products.Researching has they should obtain the different feedback for the same for user interest model construction have become an important part of personalized search engines. According to an analysis of report which extract from the website of hitwise(Table 1),the most common query length submitted to search engine(20.29%) was only two words long,65.86% of all queries were three words long or less and 10.57% of all queries were more than 5 words long .

The percentage of keyword which consists of four words long or more begin to rise. This suggests that short queries are often ambiguous. The percentage of keyword which consisted of four words long or more begin to rise. This suggests that short queries are often ambiguous. In order to solve the problems, user interest model is proposed which allows the user to obtain more precise information.

The contributions of this paper are as follows. At first, we make a general overview of the Keyword Search and the necessity of capturing user's interest. In section II a brief of the category of the collection for user interest is introduced. In section III, a general overview of the concept of user interest model, VSM, hierarchical model and tree structure model are introduced. At last we draw a conclusion in the last section.

## 2. User Interest Collection

User interest model is usually defined as a set of user's goals, plans, beliefs, knowledge and so on and as a description about man's understanding of the outside world and as a model which is used to capture user's needs, interest

TABLE I. PERCENTAGE OF CLICKS BY NUMBER OF KEYWORDS

| **Percentage of U.S. clicks by number of keywords** |
|---|

+ Corresponding author.
*E-mail address*:leafrain_001@hotmail.com

| *Subject (word)* | *Jan-08 (%)* | *Dec-08 (%)* | *JAN-09 (%)* | *Year-over-year Percent change* |
|---|---|---|---|---|
| 1 | 20.96 | 20.70 | 20.29 | -3% |
| 2 | 24.91 | 24.13 | 23.65 | -5% |
| 3 | 22.93 | 21.94 | 21.92 | 0% |
| 4 | 14.53 | 14.67 | 14.89 | 2% |
| 5 | 8.20 | 8.37 | 8.68 | 6% |
| 6 | 4.32 | 4.47 | 4.65 | 8% |
| 7 | 2.23 | 2.40 | 2.49 | 12% |
| 8+ | 2.81 | 3.31 | 3.43 | 22% |

and record or manage user's interest. Information of user's interest can be captured from relevance feedback. The relevance feedback is to take the results that are returned from user [1].

## 2.1. Explicit

Explicitly feedback is obtained from user's direct assessment. For example: asking for feedback such as ratings or appetite. One way to get user interest explicitly usually completed in the registration. In order to collect information we require user to fill in relevant interest when they login first time. Another way is that allow users to comment on results returning by search engine, such as "not interest", "somewhat interest", "interest", or "very interest". Then from the feedback we learn the users' current interests and update model of the user's for future information filtering.

## 2.2. Implicit

Recently, a great deal of effort in the research community has focused on improving user experience in web search through the incorporation of implicit user feedback [2], [3], and [4].

User interaction with Web search engines is ordinary, and complicated, in information-searching process. In the course of a search, users take various behaviors to obtain content which they are interested in. The fact that a user has browsed a website is actually that he is interested in the content of it. Obviously we can derive the user's interest by classifying information from user's browsing behavior such as which documents they select for viewing, the duration of time spent viewing a document, or page browsing or scrolling actions and so on.

User behavior can be summarized as the following aspects.

1）Search keywords

According to changes in the user's query input, infer whether there has change in user search behavior.

2）Browsing history

Site visited will be recorded in the server log, including the user's access time, user's time zone, the size of visited page and other information, by analyzing server log we can capture the user's search interest.

3）Bookmark

Typically, users will save their favorite web page for the next visit; therefore, we can obtain the user's interest by analyzing the bookmarks.

4）Mouse behavior

Users would do some operations on the link that they are interested in. For example: drag, click, suspended and so on. We can obtain information by analyzing these operations.

5）Using page dwell time

Dwelling on a page for a significant amount of time implies that the user is interested in it. There has more important information in the web page. In fact, previous studies have shown that a dwelling time of 30 seconds or more on a web page implies a kink of interest [5].

6）Search paths

Search paths are a series of pages starting with a query and ending with an event like closing the browser. Search results can only be used as the starting points for collection of user interests. A search path consists of an origin page, intermediate pages, and a destination page [6].

Origin: The first page in the paths of the SERP (search engine result page).

P2 is the origin in Fig.1.

Destination: The last page in the paths, before visited should browse a series of intermediate pages.

P5 is the destination in Fig.1.

Origin and destination is regarded as the core pages that can reveal user's interest.

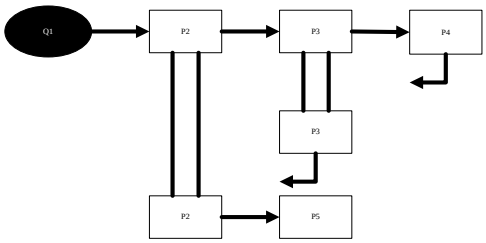


Fig. 1: Web behavior graph illustrating a search trail

## 2.3. User Interest Model Construction

User interest model is a collection of personal data associated to a specific user. It can also be considered as the computer representation of a user interest model. Traditional process of creating user interest model is Fig.2. User interest model has a short-term interest model and long-term interest model. The short-term interest model used to store user's recent interest and the stable interests stored in long-term interest model. The available data of user's interest which represented by the vector collected from registration information filled by user when his first login. Then we adjust and build short-term interest model by clustering and analyzing the information. Usually we use the vector space model to representing user interest model.

In addition, we can capture user's interest by analyzing log which include user behavior and browsing history. This is an implicit way to get user interest. We can conclude that the short-term interest should be marked as the long-term interest when the frequency of using the short-term interest reached a threshold.

# 3. USER INTEREST MODEL

There are mainly tow ways to describe the user interest model: Vector Space Model (VSM) and Hierarchy Model [7]. Salton's Magic Automatic Retriever of Text contains a similar vector space model, Inverse Document Frequency (IDF), the term frequency (TF), term discrimination values and related concepts such as feedback mechanisms [8].

## 3.1. VSM

Vector space model is a mathematical model which represents the feature vectors of a document. The basic of VSM is that a document can be represented by a vector. ($W_{1,1}$,$W_{2,2}$,$W_{3,3}$,..,$W_{i,i}$), $W_i$ is the weight of the i-item. There have some different ways to computing these values. One of the best methods is TF-IDF (term frequency-inverse document frequency) weighting. Generally we choose single words, words or longer phrases as the feature item of the document.

At first, Salton used a function of number of properties that are assigned to different documents to computing the similarity between both documents [8].

$W_i = (W_{i1}, Wi2\ldots\ldots W_{ik})$
$W_j = (W_{j1}, Wj2\ldots\ldots W_{jk})$

Following functions are necessary for similarity computing.

$$\sum_{k=1}^{n} W_{ik} \tag{1}$$

(1) is the sum of the weights.

$$\sum_{k=1}^{n} W_{ik} * W_{jk} \tag{2}$$

(2) is the sum of the corresponding term weights for $W_{ik}$ and $W_{jk}$

$$\sum_{k=1}^{n} \min(W_{ik}, W_{jk}) \tag{3}$$

(3) is the sum of the minimum weights.

And then Dice and Jaccard compute the similarity in documents by following coefficients [8].

$$sim_1(W_i, W_j) = \frac{2[\sum_{k=1}^{n} (W_{ik} * W_{jk})]}{\sum_{k=1}^{n} W_{ik} + \sum_{k=1}^{n} W_{jk}} \tag{4}$$

$$sim_2(W_i, W_j) = \frac{\sum_{k=1}^{n} \min(W_{ik}, W_{jk})}{\sum_{k=1}^{n} W_{ik}} \tag{5}$$

Relevancy rankings of documents can be calculated by calculating the similarity of the document. Usually we calculate the angle between vectors of different documents. Instead of the angle, we calculate the cosine of the angle between the vectors.

$$\cos\theta = \frac{v_1 * v_2}{\|v_1\| * \|v_2\|} \tag{6}$$

TF-IDF weights (term frequency-inverse document frequency) TF-IDF is the common weighting techniques in information retrieval. The model is known as term frequency-inverse document frequency model.

$$W_{t,d} = tf_t * \log \frac{|D|}{|\{t \in d\}|} \tag{7}$$

$tf_t$ is term frequency of term t in document d.

We have the term frequency, defined as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{8}$$

Where $n_{i,j}$ is the occurrence number of the term ($t_i$) in document $d_j$ and the denominator is the sum of number of occurrences of all terms in document $d_j$.

$\log \frac{|D|}{|\{t \in d\}|}$ is inverse document frequency.

| D | is the total number of documents in the corpus.

The denominator is the number of documents where the term t appears.

## 3.2. Tree structure

A tree structure is a way of representing the hierarchical nature of a structure in a graphical form.

Usually people use a tree structure model which with the nature of Hierarchy Model represents the user interest.

The general process of creating user interest is like Fig.2.

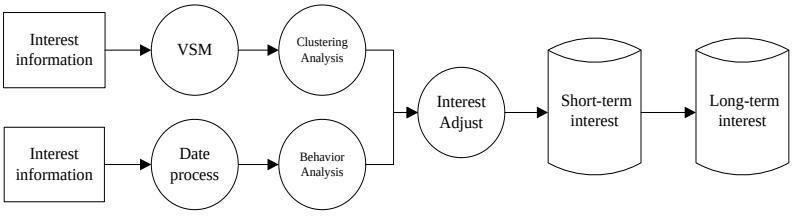


Fig. 2: Traditional process of creating user interest model

# 4. Summary and comments

Explicit building of user interesting model has several drawbacks. The user provides inconsistent or incorrect information, the model created is static whereas the user's interests may change over time, and the processing of creating user interest model may waste user's time and place a burden on the user [9], [10], [11].But in our view that we can complete the initialization of user interest model by the value that obtained explicitly and change with the drift of user's interest.

The vector space model has some limitations:

- Long documents are poorly represented because they have poor similarity values.
- Can not capture the items in the document fully for user's complex preference.
- The order in which the terms appear in the document can't be represented in the vector space.
- Not considered the synonym and polysemy of the words.

We suggested interest model should be expressed as a tree structure has 2 layers, the first layer describes the categories of user's interests and the second layer represents user's preferences. User ID used to distinguish different users.

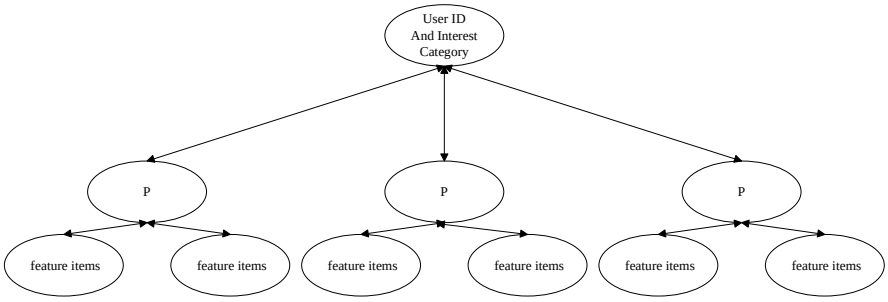


Fig. 3: Interest Tree

Each user has been created a tree structure to store interests.

The node in the first layer has property named p which is used to describe the frequency of user's interest in a certain period. In a period of time, the frequency of using the interest reaches a certain threshold and then the interest should be attributed to be the long-term interest.

A long-term interest have not been used in a long time should be forgotten. Nodes in the last layer store the feature items of user's interests .$T_j$ is the description of feature items, $W_j$ expressed the weight which gain by a certain method to the corresponding feature items.

The process of creating a user model is often neglected to collect user information explicitly, and not consider the order of browsing pages. And not selected important historical browsing information.

Therefore, we propose a method to create user interest model.

Step 1 Collecting and analyzing information from the set of user's behavior, and then selecting the valuable pages based on search paths. By filtering effective webs, we can low the amount of memory required to store the details

Step 2 Extracting feature words form the pages of Step1 and then creating VSM.

Step 3 Calculating the relevancy ranking of feature words and search key words is in order to remove noise.

Step 4 Initial formation of short-term interest model

Step 5 Adjusting to format the long-term interest model.

Decreasing the number of web pages required, increasing the number of web pages efficacious information containing, improving the accuracy of user interest model, and updating interest model as interesting moving, are merits of the method mentioned above.

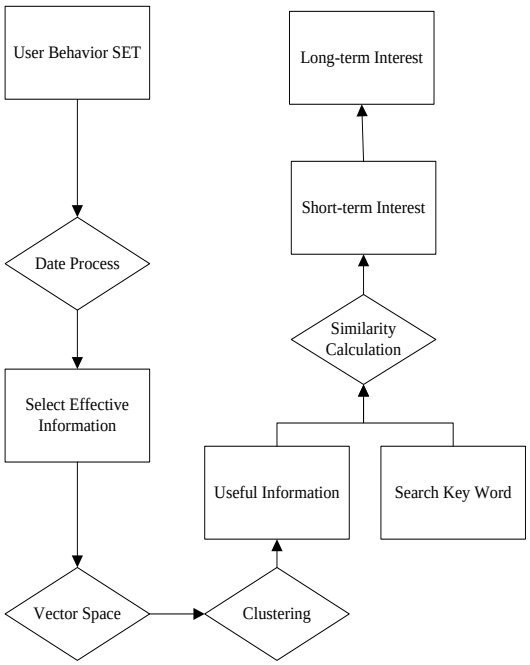


Fig. 4: Steps of creating user interest model

## 5. Acknowledgment

The authors are thankful to the experts who check and approve this paper, and to the classmates in No.149 lab of School of Computer Science and Technology in Donghua University.

## 6. References

[1] M. Claypool, P. Le, M. Wased and D. Brown, “Implicit interest indicators” In Proc of the IUI ‘01 New York: ACM Press, 2001:33- 40.

[2] J. Attenberg, S. Pandey and T. Suel, “Modeling and Predicting User Behavior” In Proc of the KDD ’09 New York: ACM Press, 2009:1067-1076.

[3] E. Agichtein, E. Brill and S. Dumais, “Improving web search ranking by incorporating user behavior information” In Proc of the SIGIR ’06 New York: ACM Press, 2006:19-26.

[4] S. Fox, K. Karnawat, M. Mydland, S. Dumals and T. White, “Evaluating implicit measures to improve web search” ACM Trans. Inf. Syst New York: ACM Press, 2005:147-168.

[5] R.W. White and J. Huang, “Accessing the scenic route: measuring the value of search trails in web logs” In Proc of the SIGIR ’10 New York: ACM Press, 2010:587-594.

[6] M. Bilenko and R.W. White, “Mining the search trails of surfing crowds:Identifying relevant websites from user activity” In Proc if the WWW2008 Beijing: 2008:51-60

[7] Y.H. Wu and Y.C. Chin, “Enabling personalized recommendation on the web based on user interests and behaviors” In Proc of the RIDE Heidelberg: IEEE Press, 2001:17-24.

[8] G.G. Chowdhury, “Introduction to Modern Information Retrieval” London Facet Press, 2004:200-210.

[9] M. Speretta and S. Gauch, “Personalized Search Based on User Search Histories” In Proc of the WI ’05 Los Alamitos: IEEE Press, 2005:622-628.

[10] D. Billsus and M.J. Pazzani, “A hybrid user model for news story classification” In Proc of the UM ’99 Banff: ACM Press,1999:99-108.

[11] C.C Chen, M.C Chen and Y. Sun “A self-adaptive personal view agent” In Proc of the KDD ’01 San Francisco:ACM Press, 2001:257-262.