Ontology-based User Profile Model Used in Information Retrieval

Wei WANG, Kunhui LIN †

Software School, Xiamen University, Xiamen 361005, Fujian, China ;

Abstract
The information on Internet expands so fast, so how to provide effect and sufficient information to meet the customers’ satisfactory has become a frontier in Information Retrieval. Personalized information retrieval provides search results based on the unique customer’s characteristic. The key difficult in the long-term personalized information retrieval is how to depict the user’s characteristic. This article analyzes the user behavior characteristics on Internet and proves the feasibility for the user profile model built in search engine. Take the effect of Ontology adopted in the semantic search into consideration, the experiment combines the Ontology technology and the user long-term historical behavior data to produce the user profile model. It proposes an Ontology graph structure to identify the user profile instead of the traditional way of demonstrating the user interest by scattered keywords without relations. The result and the prospect of this method are satisfying.

Keywords: Ontology; User Profile; Information Retrieval

1. Introduction
Since the emergence of the Internet, its scale goes up everyday. As the report from CNNIC 2008 said, the number of Chinese web pages has reached 84.7 billion and annual growth rate has reached 89.4% in 2008. For the end users of Internet, how to find their needed information within the vast amount of information resources has become a focus of academia and industry, which but also brings enormous market to search engine flied.

The traditional information retrieval system will create indexes for as much Internet information as possible, and its query service is only responsible for receiving and explaining user queries, making searching in the index of library and returning the result set, according to simple matching strategy (like Simple Boolean model and Fuzzy Boolean model). In such a system, there exist some questions. Firstly, it requires that the users describe the intention of retrieval with several key words, which often cause expression ambiguous; secondly, for a given query of user, it is very difficult to realize the user requirements for high retrieval rate and high retrieval accuracy. High retrieval rate means to find out most documents correlated to topics that customers are interested in. High retrieval accuracy means not to include the documents whose theme are not correlated.
The basic reason why there exists lots of “information junk” is that the system does not consider problems from the users’ side. How to improve the information retrieval accuracy and recommendation intelligence and how to meet the needs of different users are the challenges of a new information service system [2]. Now as a study hotspot of academic, personalized search engine attempt to improve system performance from user side. Therefore how to create the model based on the user's behavior characteristic and interest has become the key of a system. On the "the Strait information search" platform, simulate the user behavior, and create user profile model correlated to user’s interest based on Ontology.

2. User Behavior Analysis

The architecture of a search engine includes three parts, which are page collecting, texture preprocessing and query service. The search service mainly involves the research fields including information abstracted, automatic abstraction and page ranking. User profile modeling is one part of the query service. It covers three aspects, which are collecting user behavior history data, abstraction and analysis of the user history data, and matching the result document with the user profile model.

The information that a user interacts with the server is stored in a log file. The information includes the user id and the user’s browsed file record. Through analyzing these log files, the information about the page that a user browsed and how long the user stays on this page can be discovered. In order to finish this analysis work, the steps are as follows. First, classify the log file according to user id to get individual browsing history data. The browsing history data includes the page URL, the visiting date and how long the user stays on this page. Order these records by the visiting data. The record can be described as the following tuple:

\[ t = < ip_i, uid_j, \{ ( l_{url_i}^1, t_{time_i}^1, l_{length_i}^1 ), \ldots, ( l_{url_m}^m, t_{time_m}^m, l_{length_m}^m ) \} > \]

Where \( ip_i \) represents the user’s ip, \( uid_j \) represents the user’s id, \( l_{url_i}^k \) represents the event that the user visited the kth page, \( l_{url_i}^k \) represents the page URL, \( t_{time_i}^k \) represents the visited date, \( l_{length_i}^k \) stands for how long user stayed on this kth page.

Based on the intuition, the interest of a user depends on one’s age, one’s hobbies, one’s education background and the living and working environment. These factors will not change in a long period of time. In other words, if it builds a user profile model on the search engine, this model will be steady without changing frequently. The following is proved this intuition.

Assuming the query word sequence of user \( U \) is \( S1 = \{ q_1, q_2, \ldots, q_n \} \), which includes m different key words. Order them by the query times in descending order, \( S2 = \{ Q_1, Q_2, \ldots, Q_m \} \). \( S3 = \{ C_1, C_2, \ldots, C_m \} \)
is the corresponding array where $C_i$ is the query time for the word $Q_i$. Analyze the proportion $Y$ between the query times of the first $X$ percent of the words $Q$ and the total query times. The result is as follows:

$$y = \frac{\sum_{i=1}^{m} c_i}{\sum_{i=1}^{m} c_i}$$

As the above diagram shows, the distribution of the query word of a user has locality. Most of the query words are in a small range of words. In other words, the query words are much centered. Like the rule in commercial environment, which says 80 percent of the trade comes from 20 percent of customers. The 80/20 rule is applicable here too, which says 80 percent of the query times comes from 20 percent of the query words.

According to the analysis of the maze of Peking University of the user history record between April 15th 1999 and June 10th 1999[4], the decay degree of the similarity of the query words is low. In other words, in a long period of time, the user query is steady. Also based on the analysis of the deviation among the $n$ adjacent query words, as the deviation is steady, the query words are not only steady in a long period of time, but also related with each other.

In [4], the self-similarity of a user’s query action is proved by the researchers in Peking University. The self-similarity [7] is a criterion of describing a sequence with the structure similarity in a long time. The characteristic of the self-similarity is the long-term dependency, not like the short-term dependency that Poisson distribution does.

Based on the above theory, the query for a user is steady in a long period of time without frequent changes. So user profile model is feasible in search engine, not only the domain specific search engine, but also the general search engine. The user profile model stands for the user interest, also called user interest model, which stands for one’s interest or focus in a long time.

3. Architecture of User Profile Model

Traditionally the description of user profile model is based on the key word vector, which includes a scattering query word as the user profile. User interest is defined as $I = \{(t_k, w_k) | t_k \in T_i \}$, where $T_i$ represents the user query word set that is abstracted from the user history log file. $w_k$ is the weight for the $k$th query word, reflecting the user interest degree of the query word. The disadvantage of this kind of method is that it ignores the words’ relationship. If we take the advantage that Ontology plays in semantic search field into the user profile model, the user profile could have the semantic meaning. So the format of a user profile is changed from a traditional word vector into an Ontology graph. With the relationship among the words having been considered, the efficiency of the user interest model will be improved.
3.1. Ontology

Ontology is a philosophy concept, which means explaining and illustrating nature of things. In the field of Information Retrieval, Ontology refers to the specification of conceptual model. It gives the word set for a certain field and the relationship between each word as well. In this way, Ontology brings semantic information and improves the performance in turn.

Definitions of concepts are usually a frame of a name, a set of relationships with other concepts, and descriptions of it in natural language. There are four basic kinds of relationships, namely is-a, part-of, instance-of and attribute-of. Relationships can also be defined according to the specific circumstances of certain domain.

Alexander Pretschner from Germany and Susan Gauch from America have done researches on personalized search based on Ontology[9]. A user’s personalized search model is a function of the times that he has browsed on the Web, a weighted Hierarchical Network of Concepts consisting of about 4300 nodes.

Our lab has performed enhancement experiment on the topic crawler and has gained experience on using Ontology.

3.2. System architecture

The query service based on user profile model includes three parts, i.e.

1. Collecting the user behavior history data, which provides the source data for analyzing the user interest.
2. Analyzing the user behavior and training a user profile model, which is based on Ontology technology.
3. Matching the user interested data according to the user profile model from the indexed text data.

The core operation is step 2. In order to train a user profile, there should be a domain material source, which is used to train the domain Ontology. The experiment focuses on ‘the information about The Strait’. The domain material is collected from internet like sina.com, sohu.com, which is all about The Strait. Also the former research about Ontology in domain crawling[8], also provides some source material. In the domain Ontology, there are many concept entities, which are represented as a weighted vector called domain concept vector. Combined with the clustering technology[13], like KNN, abstract user profile model from the user history log file. The steps is as follows:

4. The User Profile Modeling Steps

4.1. Domain ontology

How to build the domain Ontology is one of the keys in building user profile model. Guarin[10] had proposed the five requirements of Ontology, which are definite and objective, complete, consistent, the maximum monotonous extendibility and the minimum commitment. Nowadays the method to build an Ontology entity is still not unified, however there is one common that is building an Ontology entity is required a domain expert and a linguist to be involved. How to build a Ontology automatically, is still a question too. In the experiment, based on[11], we build the Ontology about The Strait Information. The Strait Information Ontology is built based on the domain material source automatically and helped by
hand-annotation too. If there is sufficient domain source texture data, abstract the domain words and hand-annotate the words relationship. This way of building an Ontology entity automatically is feasible. However where to get the sufficient domain source data is yet a challenge too. In the experiment, the domain source data is limited that is collected from some big web sites like sina.com and sohu.com.

The Strait Information Ontology is built as follows:

1. Segment the words in the domain source data. Filter the stop words according to the stop words table. And normalize the abstracted words. The formula is as follows:

   \[ w_{i,k} = \frac{f_i}{\sum_{j=1}^{m} f_{j,k}} \]  

   Where \( w_{i,k} \) represents the weight of the \( i \)-th domain word in the \( k \)-th domain document. \( f_{i,k} \) represents the time frequency of the \( i \)-th domain word in the \( k \)-th domain document. \( m \) is the total number of the domain words.

2. Build the concept matrix

   \[ H = \begin{bmatrix} w_{1,1} & \cdots & w_{1,k} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{m,k} \end{bmatrix} \]  

   Where each row in \( H \) represents the weights of a domain key word \( w_i \) in different domain documents. While each column in \( H \) represents a word vector for a domain document \( k \).

3. SVD decomposition

   \[ H = u_{m \times m} * s_{m \times k} * v_{k \times k} \]  

   Where \( u_{m \times m} \) is the domain term matrix, which describes a set of domain concept and a vector of domain words to describe the concept. In experiment, the concept about the Strait information is hand-annotated.

4. Cluster the concepts

   Taking different clustering methods could have different effect. Build Ontology entity based on the result clustering concepts. The word relationship in Ontology is hand-annotated. So as is the above depicted, the domain concept vector should be trained. Traditionally, to get the concept vector is based on training from
enough marked documents; however, it is hard to find the sufficient domain marked documents. So to get a domain concept vector $C$, based on the Ontology entity built before, we let the node $C$ and the child nodes in the Ontology graph to be the concept vector of the concept node $C$. For example:

![Fig.3 The example of concept entity in Ontology network](image)

In the diagram above the domain concept vector of Concept $C$ consists of itself and the child nodes under it. $C = \{C, C-1, C-2, C-3, C-1-1, C-1-2\}$. Next step is how to figure the weight of each node. The weight of each node depends on the distance $d$ in Ontology graph and the contribution degree. So the formula is as follows:

$$w = \frac{1}{\text{node parent}} \times d, \quad d = \begin{cases} 0.5 & \text{level} = 1 \\ 0.2 & \text{level} = 2 \end{cases}$$

(6)

Where $\text{node parent}$ represents the parent node of node $N$. If $N$ has $K$ different parent nodes it can be considered that the probability that $N$ belongs to a class of the parent classes is $1/K$. $d$ is the distance factor.

If a node level is 1, which means it is the child node of $N$. If the node level is larger than 2, which means it is the child node of the grand-child node of $N$, and the distance factor is 0. So based on the steps above, the domain concept vector can be trained from the Ontology entity that is built before. The domain concept vector is a weighted vector.

### 4.2. User profile abstraction

#### 4.2.1. User behavior vector

To demonstrate the user behavior through the user history log files, the method is to abstract the feature items and its contribution or importance from the documents. The representation model is various. Here the VSM is taken into consideration. The detail steps are:

1. Get the domain concept vector $T = \{T_1, T_2, \cdots, T_m\}$, which is built during the process of building the Ontology.

2. Calculate the weight $W_i$ of $T_i$ in user historical document $F$, and represent the user historical browsed file $F$ into $F = (W_1, W_2, \cdots, W_n)$, which is a weighted vector of the domain concept. The $W_i$ depends on
the term frequency in the document and the document frequency that has the term, which is known as the famous TF-IDF formula. Also considering the position of the term, if the term appears in the document title, the weight is equal to 1. The formula is as follows[12]:

$$W_k = \frac{(0.5 + 0.5 \frac{tf(i)}{tf_{max}})(\log \frac{n}{df(i)})}{\sqrt{\sum_{j \in F} ((0.5 + 0.5 \frac{tf(j)}{tf_{max}})^2(\log \frac{n}{df(j)})^2)}}$$

(7)

Where $tf(i)$ is the term frequency in document $F$, $df(i)$ is the document frequency that includes the term, $n$ is the total document number, $tf_{max}$ is the maximum term frequency in document $F$.

4.2.2. User profile model

Calculate the similarity between a domain concept vector $C$ and a weighted vector $F$ for a user browsed historical page. Firstly make two vector aligned. If the term $t$ appeared in $C$, but not appeared in $F$, add term $t$ into $F$ and the weight for $t$ is 0. After vector alignment, $F = \{w_1, w_2, \cdots, w_m\}$. Secondly, calculate the similarity using the following formula:

$$Sim(F, C_i) = \frac{\sum_{j=1}^{m} (w_j \cdot c_{ji})}{\sqrt{\sum_{j=1}^{m} w_j^2 \cdot \sum_{j=1}^{m} c_{ji}^2}}$$

(8)

Calculate the similarity for each domain concept vector. Choose the top three concepts $C_i$ with the highest similarity with document $F$ as the class of the document $F$. Thirdly, calculate the class of each user historical browsed document. Count the different classes among all the document classes and accumulate the weight of the same classes. As the document number increases, the total different class number is convergent[11]. When the steps are finished, the user profile model can be described by all the classes of user browsed documents, which is $I = \{C_1, C_2, \cdots, C_s\}$ with the weight to control the user interest degree about this concept.

5. Experiment

The experiment focuses on the weapon information about The Strait. The Ontology used in this experiment is based on our lab’s former experiment[8]. Also combined with the hand-annotation of the concept and the word relationship between terms, the Ontology network has nearly 150 concept nodes. The following is a part of the weapon Ontology:

In the above Ontology, the domain concept vector of the concept ‘plane’ is \{plane,type,missile\} with the weight \{1,0.5,0.5\}. For lack of sufficient user log data, the data is the stimulated user historical data. Take a document set as the user historical data, whose interest is stimulated for plane information. Build the user profile model following the above steps. The most top concept node in the User profile Ontology model is ‘plane’.
However the experiment has some limitations. Firstly, it lacks of enough domain materials, so the weapon Ontology is not completed. Secondly, because it lacks of user historical data the data are stimulated. But based on the Ontology effect used in semantic search and the lab’s former experiment[8], the prospect of the method for describing user interest as an Ontology graph is satisfying. This experiment proves the method’s feasibility.

6. Summary

In order to make the personalized search engine, which provides query results from the customer side, the user profile model is the key. This article built a domain Ontology. Calculate the similarity between the domain concept vector and the user historical document vector. Describe the user’s interest in the Ontology graph with the weight to demonstrate the user interest degree about the concept. Based on user profile model, the search engine can provide some personalized services, like pushing automatically search engine, etc. There could be some improvement, including modifying the user model according to the user interaction with the query results. The limitation lies in insufficient and immature Ontology technology and difficulty in getting complete and real user behavior data for experiments. However, based on effect that Ontology brings for semantic search, this model method has satisfying prospect.

Acknowledgement

This work was supported by the National Hi-Tech Research and Development Program (863) of China (No. 2006AA01Z129), the National Natural Science Foundation of P.R. China (No. 60672018) and the 985 Innovation Project on Information Technique of Xiamen University (0000-X07204).

Reference


