Utilizing Search Intent in Topic Ontology-based User Profile for Web Mining

Xujuan Zhou, Sheng-Tang Wu, Yuefeng Li, Yue Xu, *Raymond Y.K. Lau, Peter D. Bruza
School of Software Engineering and Data Communications
Queensland University of Technology, QLD 4000, Australia
x.zhou@student.qut.edu.au, {s.wu, y2.li, yue.xu, p.bruza}@qut.edu.au

**Department of Information Systems, City University of Hong Kong*
Tat Chee Avenue, Kowloon, Hong Kong SAR,
raylau@cityu.edu.hk

Abstract

It is well known that taking the Web user profiles into account can enhance the effectiveness of Web mining systems. However, due to the dynamic and complex nature of Web users, automatically acquiring worthwhile user profiles was found to be very challenging. Ontology-based user profile can possess more accurate user information. This research emphasizes on acquiring search intentions information. This paper presents a new approach of developing user profile for Web searching. The model considers the user's search intentions by the process of PTM (Pattern-Taxonomy Model). Initial experiments show that the user profile based on search intention is more useful than the generic PTM user profile. Developing user profile that contains user search intentions is essential for effective Web search and retrieval.

1. Introduction

Even though Web searching techniques and hardware have become very sophisticated, they are still far from perfect. Most search engines rely entirely on user queries to perform searching tasks. A user query refers to a list of keywords (plus operators), and the query response refers to the list of pages ranked depending on their similarity to the query [2]. The keyword-matching-only Web searching approach often leads to information overload.

To alleviate the information overload problems, Web mining has emerged as an important research area in the field of Web Intelligence [5]. Web mining was developed from data mining techniques to automatically discover and extract information from web documents and services [1]. Web mining techniques include finding and extracting relevant information that is hidden in Web-related data. Improving the web mining techniques is critical in dealing with the information overload. .

Currently, Web searching does not consider the user's search intension. Most of the search engines simply disregard the Web user's profile. User profiles are an important source of metadata for Information Retrieval (IR) processes. To improve precision and increase information access efficiency, the Web search process has to evolve further with the ability to incorporate user's search intension. However, valuable Web user profiles are difficult to acquire without manual intervention.

Ontological user profiles can facilitate the search engines to perform more intelligent search and retrieval tasks. Recently, some ontology-based user profiles models have been developed [3, 4, 5, 9, 11]. In these researches, user profiles focusing on various aspects have been integrated into the Web searching. Unlike these recent developments, this research focus on building a user profile based on user's information search intention. To emphasize the topics that may catch the user's attention, this method is called Topic Ontology-based-user-profile Model (TOM).

This paper includes 5 sections: Section 2 briefly describes user search intent in user profiles. The proposed method for learning user search intention is presented in section 3 followed by results of some experiments. Conclusions and future works are given in Section 5.

2. User intention profiles

One of the key issues in developing an effective Web mining system is to construct accurate and comprehensive user profiles that can describe the user information needs and information searching goals. Li and Zhong [5] classified Web user profiles into two diagrams: the data diagram and information diagram. The data diagram is the discovery of interest registration data and customer portfolios. The information diagram is the discovery of interest topics for Web user information needs. Compared with the data diagram, information diagram has two distinguished characters: firstly, there are more duplicates in the data; and secondly, the meaning of data values (terms) is ambiguous due to the existent of "synonymy" and "hyponymy" relations between terms. It is a more challenging to build information diagram type user profiles. In this research, Web user profile mining is

Proceedings of the 2006 IEEE/WIC/ACM International Conference
on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE

referred to mining Web user profile in the information diagram, i.e. mining user interested topics.

2.1. User profile represented by topic ontology

In this research, the user profile is constructed from the topics of a user's interest i.e., search intent. The topic in a particular document comprises the terms which represent the subjects. By using the ontological approach, the user profile also includes the topic's semantic relationship. Hence, this type of user profile is called topic ontology. The topic ontology is constructed from primitive objects (e.g., terms). They consist of primitive classes and compound classes. The primitive classes are the smallest concepts that cannot be assembled from other classes. They may be inherited by derived concepts or their children. The compound classes are constructed from a set of primitive classes.

The process of mining ontology from Web documents was developed by Li and Zhong [3, 4]. In this mining procedure, the base backbone and the top backbone are employed to connect patterns with each other. The base backbone is used for the linkage between primitive classes while the top backbone for the linkages between compound classes. The process of building the topic ontology requires both the base backbone and the top backbone constructions.

The following definitions describes the basis process of topic ontology mining.

***Definition* 2.1.** Let $T = \{t_1, t_2, \ldots, t_k\}$ be a set of keyword (or terms). Let D be a training set of documents, which consists of a set of positive documents, D^+; and a set of negative documents, D^- . Let each document is a set of terms (may include duplicate terms).

***Definition* 2.2.** A set of terms is referred to as a *termset*. *Term frequency tf* (d,t) is defined as the number of occurrences of t in d if given a document d and a term t. A set of term frequency pairs, $P = \{(t,f) \mid t \in T, f = tf(t,d) > 0\}$, is referred to as a *pattern.* A pattern is uniquely determined by its termset.

***Definition* 2.3.** Let $termset(P) = \{t \mid (t,f) \in P\}$ be the *termset* of P. Given a pattern $P = \{(t_1, f_1), (t_2, f_2), \ldots, (t_r, f_r)\}$, its normal form $\{(t_1, w_1), (t_2, w_2), \ldots, (t_r, w_r)\}$ can be determined by equations:

$$w_i = \frac{f_i}{\sum_{j=1}^{r} f_j} \quad \text{for all } i \le r \text{ and } i \ge 1 \qquad (3.1)$$

***Definition* 2.4.** *support*(P) is used to describe the extent to which the pattern is discussed in the training set: the greater the support is, the more important the pattern is.

***Definition* 2.5.** Topic ontology is represented by O. It consists of a set of patters. $O = \{P_1, \ldots, P_n\}$. There are some relations between patterns: if P_1 is subset of P_2 then "part-of" relationship holds by these two patterns; if $P_1 \cap P_2 \neq \emptyset$ then intersect relationship exists between P_1 and P_2; if $P_1 = P_2$ then "is-a" relationship exists between P_1 and P_2, and these two patterns should be composed to generate a new patterns, $P_1 \oplus P_2$. $support(P_1 \oplus P_2) = support(P_1) + support(P_2)$, where $\oplus$ is a composition operator. The *support* can be normalized by:

$$support: O \rightarrow [0,1], \text{ such that}$$

$$support\ (P) = \frac{support\ (P)}{\sum_{Pj \in O} support(P_j)} \qquad (3.2)$$

Using some existing algorithms on the bottom up approach [6], the hierarchy of all keywords in T can be obtained. Here, T consists of a set of clusters, Θ, where each cluster in Θ is represented as a term. $\Theta \subseteq T$ is called the set of primitive objects. Then all compound classes can be constructed from some primitive ones using *OntoMining* algorithm [4].

***Definition* 2.6.** The correlation in the top backbone of the ontology is represented by an *association set* $<support, \beta>$ from O to Θ according to Li and Zhong [12], where β is a mapping which satisfies:

$$\beta : O \rightarrow 2^{\Theta \times [0,1]} - \{\emptyset\} \text{ such that}$$

$$\beta(P) = \{(t_1, w_1), (t_2, w_2), \ldots, (t_r, w_r)\} \subseteq \Theta \times [0,1],$$

and $\beta(P)$ is P's normal form. An association set maps a pattern to a termset and provides a term weight distribution for the terms in the termset.

2.2. Interpretation of a user's search intention

Web search intentions can be studied in two means. On one hand, a user may be interested in more focused information and his/her search goal is to find exact information related to the key word queried. On the other hand, a user may wish to find more general information. Understanding the goals and behavior of information seekers would be useful for building more accurate Web user profiles. In mathematical terms, Web user search intentions can be generalized as specificity and exhaustivity intent. Specificity (*spe*) describes the extent of the pattern (or topic) i.e., user's interests have a narrow and focusing goal, whereas exhaustivity (*exh*) describes a different extent of the searching pattern (or topic) i.e., general/wider scope of user interests.

The definitions of *specificity* and *exhaustivity* are discribed in the following. They are inspired by the Dempster-Shafer (D-S) theory. The numeral functions for measuring specificity and exhaustivity are:

$$spe: 2^{\Theta} \rightarrow [0, 1]; \text{ such that}$$

$$spe\ (A) = \sum_{P \in O, termset(P) \subseteq A} support\ (P) \qquad (3.3)$$

$$exh: 2^{\Theta} \rightarrow [0, 1]; \text{ such that}$$

$$exh\ (A) = \sum_{P \in O, termset(P) \cap A \neq \phi} support\ (P) \qquad (3.4)$$

Proceedings of the 2006 IEEE/WIC/ACM International Conference
on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE

for all $A \subseteq \Theta$.

According to Shafer, the D-S theory is based on two ideas of obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence. There are three important functions in D-S theory: the *basic probability assignment* function (bpa or m), the *Belief* function (*Bel*), and the *Plausibility* function (*Pl*). Specificity of pattern P is related to a belief function and exhaustivity of pattern P is related to a plausibility function, respectively. According to Equation (3.3) and (3.4), the specificity of pattern P is expressed by all its sub-patterns and its exhaustivity is expressed by all patterns that overlap with it. A probability function from a given association set ($<support, \beta>$) is

$$Pr(t) = \sum_{P \in O,(t,w) \in \beta(P)} support(P) \times w \qquad (3.5)$$

for all $t \in T$.

Depending on the user's search goal, the system may choose one of the following two methods to assess relevance of the Web information for them. If users intent to search specific topic, patterns in the user profile ontology have higher value of $spe(P_i)$. This means that Web user needs have more details information. Hence, Web information that has higher relevant patterns is more suitable for them. The relevance function will be:

$$relevance_{spe}(P_i) = spe(P_i) \sum_{t \in Pi} Pr(t) \qquad (3.6)$$

In other cases, users may wish to find more general information, the user profile ontology have greater value of $exh(P_i)$. Web documents with broad-spectrum content will be more suitable for them. The relevance function should be:

$$relevance_{exh}(P_i) = exh(P_i) \sum_{t \in Pi} Pr(t) \qquad (3.7)$$

The method for assessment of relevance of topic ontology has been developed and was presented in previous work [11]. The relevance assessment will proceed during Web search depending on user's search intent. That is, after finding out the user's search intent, the system will decide using which relevance function to assess whether the topic is relevant topic or not.

3. Proposed method of learning of search intention

While learning the topic ontology that contains user search intention, the term frequencies are important for document as a whole. However, in a sentence or paragraph, the frequent sequential patterns are more important. The proposed method adapts Pattern Taxonomy Model (PTM) to distinguish user intent by analyzing the user feedback. PTM was developed by Wu *et. al.* [10]. This method is able to derive rich semantic information underlying the user's Web searching history/behavior. The Web user feedbacks were obtained implicitly. Whether a document is relevant can be judged by user's positive or negative feedback.

Pattern taxonomy is a tree-like structure that illustrates the relationship between closed patterns extracted from a text collection. An example of pattern taxonomy is shown in Figure 1. The arrow in this figure indicates the sub-sequence relation between patterns. The pattern $\langle t_1, t_2 \rangle$ is a sub-sequence of pattern $\langle t_1, t_2, t_3 \rangle$, and pattern $\langle t_2 \rangle$ is a sub-sequence of pattern $\langle t_2, t_3 \rangle$. The root of the tree in the bottom level represents one of the largest patterns. Once the tree is constructed, the relationship between patterns can be quantified.

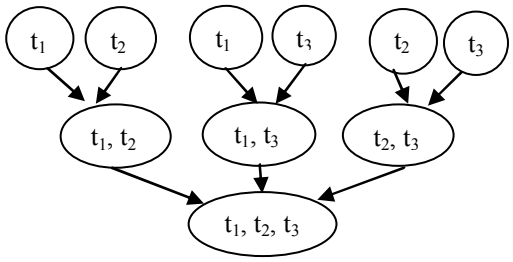

Figure 1. A example of PTM

Depending on the underlying relationship, different PTM tree structure has very different characteristic. Some PTM trees have many branches and a few levels whereas some PTM trees have many levels and a few branches. The pattern which has more levels indicates that the Web user has a clear direction on what he/she is looking for. The value of specificity is greater than exhaustivity and such pattern represents the "specificity" intent. Whereas the pattern has more branches shows that the user's search focus is not well defined. The Web users may have a wide range of interests. The value of exhaustivity is larger than specificity and such patterns represent the exhaustivity intent.

The length of sequential pattern P, denoted as $len(P)$, indicates the number of words (or terms) contained in P. A sequential pattern which contains n terms can be denoted in short as ***nTerms* pattern**. For instance, given pattern $P = \langle t_1, t_2 \rangle$, we have $len(P) = 2$, and P is a *2Terms* pattern. The number of patterns with different length can affect the pattern taxonomy structure. It was found that the number of *2Terms* and *3Terms* patterns can be major factors for determining if a topic is specificity or not. Hence, the factor F is computed by the following formula:

$$F_{spe} = \frac{Num_{2TP} + Num_{3TP}}{Num_{1TP}}$$

where Num_{1TP}, Num_{2TP}, Num_{3TP} are the number of *1Terms* patterns, the number of *2Terms* patterns, the number of *3Terms* patterns, respectively. If F_{spe} is greater than a threshold which is an experimental coefficient then the topic is a specific topic.

Proceedings of the 2006 IEEE/WIC/ACM International Conference
on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE

4. Experimental evaluation

Experimental tests has been conducted based on the Reuters RCV1 (Reuters Corpus Volume 1) corpus. It is used by TREC (Text REtrieval Conference) in recent years for the adaptive filtering track. TREC has developed and provided 100 topics for the filtering track aiming at building a robust filtering system [7]. Each topic is divided into two sets: training and test set.

The relevance judgments have been given for each topic. The set of 100 TREC topics is used to represent the diverse Web user's information needs. The experiments simulated user feedback by assuming that the user would recognize as relevant the chosen some documents that were officially judged as relevant from a set of given documents. 39 topics are recognized as the specificity topics by the system from the 100 topics. 100 topics are used as input for PTM and these 39 topics are used as input for TOM. In this study, the experiments are carried out by learning specificity search intention.

Two traditional factors of measuring effectiveness are *Recall* and *Precision*. The recall is the proportion of relevant documents that are retrieved while the precision is the proportion of the retrieved documents that are relevant. A number of other measures are derived from them. F-measure, average precision measures and precision/recall breakeven point are used in our evaluation procedure.

Wu *et. al*. [10] used the pattern-based taxonomy rather than single words to represent documents. They have conducted experiments on TREC collections and have compared the performance of their model with keyword based models. They concluded that their method outperforms the keyword based method. Therefore, the PTM method will be the baseline for this study. Figure 2 illustrates the P/R breakeven point, average precision and F-measure for TOM and PTM. It demonstrated that the performances of TOM method are better and more consistent than the original PTM method.

5. Conclusions

Integrating ontology-based user profiles into the processing can be very beneficial for improving the efficiency of Web information search and retrieval. Considering the Web user's search intention will assist building a more useful user profile. PTM is able to provide rich semantic relationship between the patterns. With the Web user's search intention derived from PTM, the new method shows a considerable improvement in terms of search effectiveness. This demonstrates that the user profiles based on the search intention can improve the information retrieval performance.

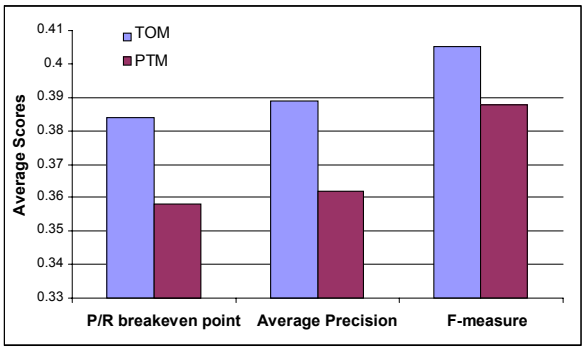

Figure 2. The P/R breakeven point, average precision and F-measure for TOM and PTM

Acknowledgments

This work was partially supported by Grant DP0556455 from the Australian Research Council.

References

[1] O.Etzioni, "The World Wide Web: quagmire or gold mine?," *Communication of the ACM*, vol. 39, pp. 65-68, 1996.

[2] R.Garofalakis, M.N.Rastogi, S.Seshadri, and K.Shim, "Data Mining and the Web: Past, Present and Future," presented at Workshop on Web Information and Data Management, 1999.

[3] Y.Li, N.Zhong, "Ontology-based Web mining model," presented at IEEE/WIC International Conference on Web Intelligence, Halifax, Canada, 2003.

[4] Y.Li, N.Zhong, "Web mining model and its applications for information gathering," *Knowledge-Based Systems*, vol. 17, pp. 207-217, 2004.

[5] Y.Li, N.Zhong, "Mining Ontolgoy for automatically acquiring Web user information needs," *IEEE Transactions on Knowldeg and Data Engineering*, vol. 18, pp. 554-568, 2006.

[6] A.Maedche, *Ontology learning for the semantic Web*: Kluwer Academic Publishers, 2003.

[7] T.Rose, M.Stevenson, and M.Whitehead, "The Reuters Corpus Volume 1 - From yesterday's news to today's language resources," presented at the 3th International Conference on Language Resources and Evaluation, Las Palmas, Spain, 2002.

[8]] G.Shafer, *A Mathematical Theory of Evidence*: Princeton University Press, 1976.

[9] J.Trajkova, S.Gauch, "Improving Ontology-Based User Profiles," presented at RIAO 2004, Vaucluse, France, 2004.

[10] S. T.Wu, Y.Li, Y.Xu, B.Pham, and P.Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining," presented at the IEEE/WIC/ACM International Conference on Web Intelligence, Beijin, China, 2004.

[11] X. Zhou, Y. Li, Y.Xu, and R.Lau, "Relevance Assessment of Topic Ontology," presented at The Fourth International Conference on Active Media Technology, Brisbane, Australia, 2006.