Integration of Information Filtering and Data Mining Process for Web Information Retrieval

Xujuan Zhou, Yuefeng Li, Peter Bruza, Yue Xu

Faculty of Information Technology
Queensland University of Technology
QLD 4000 Australia

{x.zhou, y2.li, p.bruza, yue.xu}@qut.edu.au

Abstract *This paper examines a new approach to Web information retrieval, and proposes a new two stage scheme. The aim of the first stage is to quickly filter irrelevant information based on the user profiles. The proposed user profiles learning algorithm are very efficient and effective within a relevance feedback framework. The aim of the second stage is to apply data mining techniques to rationalize the data relevance on the reduced data set. Our experiments on RCV1 (Reuters Corpus Volume 1) data collection which is used by TREC in 2002 for filtering track show that more effective and efficient access Web information has been achieved by combining the strength of information filtering and data mining method.*

Keywords Information filtering, User profiles, Data mining, Pattern taxonomic model

1 Introduction

Web search engines are designed based on traditional information retrieval techniques to return a set of potential relevant documents that match the user's direct query. The queries submitted to search engines by Web users are generally very short containing only two or three words [2]. Although such simple keywords approach works very well if the short query is an unambiguous term with fairly high discriminating power (e.g. using "Sweatshop" as query term) in general, these short queries can not clearly describe a user's true information search intent and they will open up the problem of vocabulary mismatch [1]. In deed, search engines provide a "one size fits all" solution to all users. This solution often leads to information overload problem.

Relevance feedback (RF) is a well known method to help users conduct searches iteratively and it has been shown that RF can significantly improve retrieval performance [7]. In this paper, we propose to construct a user profile through user's interactive feedback. In stead of requiring the user to explicitly express and specify their information needs beforehand, we alleviate the user's cognitive burden by only asking her to indicate whether a small set of documents are relevant or not. This set of user feedback is then used as training data set and a learning method will be developed to learn a user profile from training data. In this project, 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.

The main objective of the research work presented in this paper is to develop a novel Web information retrieval system which integrates information filtering and data mining strategies to provide more precise results for the Web search. The remainder of the paper is organized as follows. Section 2 highlights previous researches in the related area. The proposed two-stage method of filtering and data mining will be illustrated in Section 3 and Section 4. The empirical testing results will be reported in Section 5. Section 6 describes the findings of the experiments and discusses the results. The concluding remarks and future researches are given in section 7.

2 Related works

In dealing with Web information overload issues, classical methodologies/techniques from information retrieval/filtering (IR/IF) and data mining have been applied separately with various success. IF systems learn user profiles from their interaction with systems and then use this information to analyze new documents. The profiles can be constructed using a variety of learning techniques including the vector space model, genetic algorithm, and the probabilistic model or clustering. Recently, a number of ontology-based user profiles models have been developed, e.g., [3, 10].

Data mining is the process of automatically extracting useful knowledge from large data sets. Web mining is concerned with data mining on the Web. Many Web data mining methods have been developed to underpin IF system. For example,Web usage mining provide an excellent way to learn about users' interest [8]. The authors of [9] have developed a pattern taxonomy model (PTM) for Web information gathering. Many up-to-

Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia, December 10, 2007. Copyright for this article remains with the authors.

date Web mining techniques (e.g., sequential association rules, closed-pattern based non-redundant association rules and rough association rules) have been intergraded into this method.

The idea of integrating IF and data mining for Web information retrieval has evolved from these two well established, but largely disparate fields. This proposed method intends to exploit the advantages of IF and data mining within the one system.

3 Learning user profiles for IF

There are two phases in combined system. The first phase comprises IF and second phase relies on data mining. The most challenging issues in filtering is to develop a method to learn user profiles efficiently and effectively from very limited user intervention and to implement a way to "filter" information from a huge collection of documents. This section presents a user profile learning method based on rough association rule mining whereby only positive feedback is required [4, 5].

3.1 User profile construction

User profiles will be represented by an ontology. Syntactically we assume that the ontology consists of primitive classes and compound classes. The primitive class is constructed from terms in Θ. The primitive classes are the smallest concepts that cannot be assembled from other classes. However, they may be inherited by derived concepts or their children. Then a set of primitive objects (terms) can be selected from the set of keywords by using the existing background knowledge.

Let D be a training set, which includes a non-empty set of positive documents D^+ and a set of negative documents D^-. Let $\Theta = \{t_1, t_2, \ldots, t_k\}$ be a set of selected terms (or primitive classes).

A set of terms is referred to as a $termset$. Given a document d (or a paragraph) and a term t, $tf(d,t)$ is defined as the number of occurrences of t in d. A set of term frequency pairs,

$$P = \{(t,f) | t \in T, f = tf(t,d) > 0\}$$

is referred to as a *pattern* in this paper.

Let $termset(P) = \{t | (t,f) \in P\}$ be the termset of P, pattern P_1 equals to pattern P_2 if and only if $termset(P_1) = termset(P_2)$. A pattern is uniquely determined by its termset. Two patterns should be composed if they have the same termset (or they are in a same category). To compose two patterns which have same termset, the composition operation, $\oplus$, that defined in [4, 5] will be used to generate new patterns.

The compound classes are constructed from a set of primitive classes: $\Omega = \{p_1, p_2, \ldots, p_k\}$. There are "is-a" and "part-of"relations between these objects. A document is irrelevant if its any part-of section does not include any pattern.

Let $p =< termset(p), wd(p) >$, we can also view it as a rough association rule which has the form of

$$< termset(p), wd(p) > \rightarrow positive,$$

where $termset$ is a set of selected terms, and wd is a weight distribution of these terms in the rule.

Rough association rules can be discovered from a set of positive documents (or paragraphs) by extracting patterns of term frequency pairs, composing patterns with the same termsets, and normalizing the weight distributions (see [4] or [5]).

Let $O = \{(p_1, N_1), (p_2, N_2), \ldots, (p_n, N_n)\}$ be a set of compound objects (discorded patterns), where p_i are patterns $(1 \leq i \leq n)$ and N_i denote numbers of patterns that composed together. A support function can be attained from O, which satisfies:

$$support(p_i) = \frac{N_i}{\sum_{(p_j,N_j)\in O} N_j} \qquad (1)$$

for all $(p_i, N_i) \in O$.

To describe the semantic relations between compound classes (discovered patterns), we use a common hypothesis space Θ. We then can map the discovered patterns onto the common hypothesis. The following mapping is designed for this purpose:

$$\xi : DP \rightarrow 2^{\Theta} - \{\emptyset\}, \text{ such that}$$

$$\xi(p_i) = termset(p_i) \qquad (2)$$

where $DP = \{p_i | (p_i, N_i) \in O\}$.

Finally, we can obtain a probability functions pr_ξ on the set of terms to represent the discovered in the discovered patterns, which satisfies:

$$pr_\xi(t) = \sum_{p \in DP, t \in \xi(p)} \frac{support(p)}{|termset(p)|} \qquad (3)$$

for all $t \in \Theta$.

3.2 Filtering

The objective of filtering phase is to filter our non-relevant incoming documents. To determine a reasonable threshold, in this paper, we discuss how to classify incoming documents into three regions: relevant, boundary, and irrelevant documents region according to the above discovery.

Let p be a pattern and d be a new incoming document. Our basic assumption is that d should be relevant if $termset(p) \subseteq d$. The set of incoming documents that satisfy $termset(p) \subseteq d$ is called the *covering set* of p and denoted as $[p]$. The **positive region** (*POS*) is the union of all covering sets for all $p \in DP$.

The set of incoming documents that satisfy $\exists p \in DP \Rightarrow termset(p) \cap d \neq \emptyset$ is called the **boundary region** (*BND*). Also, the set of incoming documents that satisfy $\forall p \in DP \Rightarrow termset(p) \cap d = \emptyset$ is called the **negative region** (*NEG*). Given an incoming document d, the decision rules can be determined naturally as follows:

$$\frac{\exists p \in DP \Rightarrow termset(p) \subseteq d \neq \emptyset}{d \in POS}$$

$$\frac{\exists p \in DP \Rightarrow termset(p) \cap d \neq \emptyset}{d \in BND}, \text{ and}$$

$$\frac{\forall p \in DP \Rightarrow termset(p) \cap d = \emptyset}{d \in NEG}.$$

The probability function pr_ξ on Θ (see Equation 3) has the following property:

$$\sum_{t \in d} pr_\xi(t) \geq \min_{p \in DP}\{\sum_{t \in \xi(p)} pr_\xi(t)\} \qquad (4)$$

for all $d \in POS$.

We use $\min_{p \in DP}\{\sum_{t \in \xi(p)} pr_\xi(t)\} + \alpha$ as the threshold. A very important conclusion we can draw from the above analysis is that the above threshold can retain all POS incoming documents and part of BND incoming documents, where α is an experimental coefficient that is used for obtaining the part of BND incoming documents.

By incorporating filtering into the web search, likely irrelevant data will be filtered out quickly at the beginning. The remaining data will be comprised of relevant, boundary or maybe a few irrelevant documents. The size of the remaining dataset is dramatically reduced.

4 Data mining

After filtering, a data mining process based on the pattern taxonomy model (PTM) [9] will be carried out on the residual data set. The following is a brief introduction to PTM.

A sequence $s =< x_1, \ldots, x_m >$ ($x_i \in T$ is a termset) is an ordered list. A sequence $\alpha =< a_1, \ldots, a_m >$ is a sub-sequence of another sequence $\beta =< b_1, \ldots, b_n >$, denoted by $\alpha \subseteq \beta$, if and only if $\exists i_1, \ldots, i_m$ such that $1 \leq i_1 < i_2 \ldots < i_m \leq n$ and $\alpha_1 \subseteq \beta_{i_1}, \alpha_2 \subseteq \beta_{i_2}, \ldots, \alpha_m \subseteq \beta_{i_m}$. A sequential pattern s is a *very closed sequential pattern* of s' if $s \subseteq s'$ and $support(s) - support(s') < \lambda \times support(s')$, where λ is a small positive decimal.

The above definitions can be used to create a pattern taxonomy as depicted in Figure 1, where a, b, c, and d are terms, the arrows are "is-a" relation, e.g., phrase $< (a)(b) >$ is a sub-sequence of $< (a)(b)(c) >$.

If the frequency is used to define the *support* function for all patterns, then $support(< (a)(b) >) \geq support(< (a)(b)(c) >)$. In general, 3 subsequence patterns of $< (a)(b)(c) >$ can be obtained. They are $< (a)(b) >$, $< (a)(c) >$ and $< (b)(c) >$. If patterns have supports which are very closed to their parents supports then these patterns are called non-closed patterns. The not very closed sequential patterns will be removed. e.g.,$< (a)(c) >$ in Fig. 1 has been pruned.

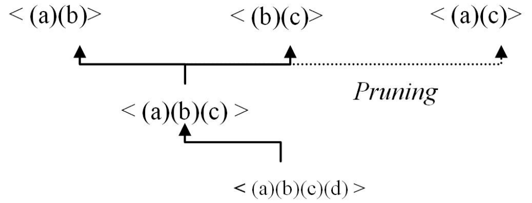

Figure 1: Pattern taxonomy.

After a pattern taxonomy has been extracted from a training set, it is utilized to calculate $pr(d)$ which is the relevance degree of each new incoming document d for a given topic. The following is the procedure of making decisions to return relevant document to the user:

1. Find all longest patterns in document d;
   e.g., ($< (a)(b)(c) >$) is a longest pattern
   if ($< (a)(b)(c)(d) >$) does not appear in d.

2. Determine $pr(d)$ according to the taxonomy.
   e.g., $pr(d) = support(< (a)(b)(c) >) + support(< (a)(b) >) + support(< (b)(c) >)$.

5 Experiments

To evaluate the effectiveness of the filtering and retrieval function used by our proposed system, several experiments have been conducted.

5.1 Dataset

The standard TREC test collections RCV1 was used to test the effectiveness of the proposed model. TREC has developed and provided 100 topics for the filtering track aiming at building a robust filtering system. Each topic is divided into two sets: training set and testing set. Our experiments use the Split of TREC-10/2000. Document relevance judgments have been supplies for each topic. The set of one hundred 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 an officially judged relevant document. The documents have been pre-processed by removing stop-words and stemming terms before they are used in all experiments.

5.2 Baseline methods

Two baseline models are used: BM25 model and a PTM-based model. BM25 [6] is one of sate-of-the-art retrieval function used in document retrieval, such as Web search. In this paper, the following weighting function is used: given a query Q, containing keywords $q_1, q_2 \ldots, q_n$, the BM25 score of a document D is:

$$\begin{aligned} score(D, Q) &= \sum_{i=1}^{n} \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \\ &* \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|D|}{avgdl})} \end{aligned}$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i. where $f(q_i, D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D (number of words), and $avgdl$ is the average document length in the text collection from which documents are drawn. Parameter settings in the experiments were $k_1 = 1.2$ and $b = 0.75$.

The PTM-based method used the pattern-based taxonomy rather than single words to represent documents. The authors of [9] have conducted experiments on TREC collection (RVC1 corpus) and have compared the performance of their model with keyword based models such as Rocchio and traditional Probabilistic model. They concluded that their method outperforms the keyword based methods.

5.3 Results

Effectiveness was measured by three means: The F_β measure ($\beta = 1$ is used in our experiments), Mean Average Precision (MAP) and the break-even (B/E) point.

The results reported in here are the average scores of B/E point, MAP and F_1 on all 100 TREC topics for all methods.

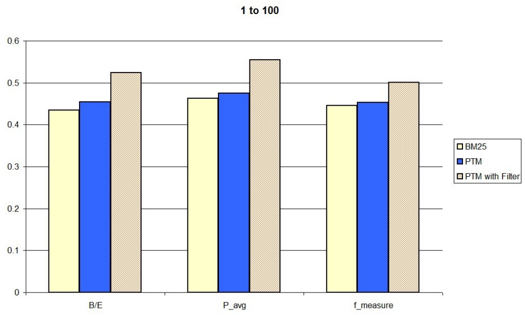

Figure 2: Results on topics 1-100 for all three methods

6 Discussion

In the filtering phase, by using the rough association rule method to build ontology-based user profiles, the only positive documents are needed. An ontology is able to provide rich semantic relationship between patterns. Therefore, the ontology-based user profiles can perhaps express user information needs and searching goals more accurately and comprehensively. Based on these profiles, most irrelevant documents are able to be filtered out and consequently the chance of generating noisy patterns is reduced significantly. In the data mining phase, a pattern taxonomy is built on a data set which has less noise hence promotes precision without negatively impacting recall.

In short, the experiment results provide evidence that the combination of filtering and data mining can improve information access significantly.

7 Conclusions

This paper illustrates a new model which integrates an ontology-based user profile filtering and pattern based data mining technology together to alleviate Web information overload and mismatch problems. The proposed method has been evaluated using standard TREC data collection with encouraging results.

Compared with the orthodox data mining method PTM the experiments based on the new method demonstrated that the performance of information retrieval can be significantly improved. The improvement of the new method is mainly due to the success of irrelevant information removal by the filtering process.

Acknowledgements This paper was partially supported by Grant DP0556455 from Australian Research Council.

References

[1] Peter Bruza, Robert McArthur and Simon Dennis. Interactive internet search: keyword, directory and query reformulation mechanisms compared. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 280–287, New York, NY, USA, 2000. ACM.

[2] Bernard J. Jansen, Amanda Spink and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, Volume 36, Number 2, pages 207–227, 2000.

[3] Y. Li and N. Zhong. Ontology-based web mining model. In *IEEE/WIC International Conference on Web Intelligence*, 2003.

[4] Yuefeng Li and Ning Zhong. Rough association rule mining in text documents for acquiring web user information needs. In *Web Intelligence*, pages 226–232, 2006.

[5] Yuefeng Li and Ning Zhong. Mining rough association from text documents for web information gathering. *T. Rough Sets*, Volume 7, pages 103–119, 2007.

[6] Stephen E. Robertson, Steve Walker and Micheline Hancock-Beaulieu. Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive. In *TREC*, pages 199–210, 1998.

[7] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, Volume 41(4), pages 288–97, 1990.

[8] Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, Volume 1, Number 2, pages 12–23, 2000.

[9] S.T.Wu, Y.Li, Y. Xu, B. Pham and P.Chen. Automatic pattern-taxonomy extraction for web mining. In *the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 242 – 248, China, 2004.

[10] X. Zhou, Y. Li, Y. Xu and R. Lau. Relevance assessment of topic ontology. In *The Fourth International Conference on Active Media Technology*, Relevance Assessment of Topic Ontology, 2006.