A User Profiles Acquiring Approach Using Pseudo-Relevance Feedback

Xiaohui Tao and Yuefeng Li

Faculty of Science & Technology, Queensland University of Technology, Australia
{x.tao, y2.li}@qut.edu.au

Abstract. User profiles are important in personalized Web information gathering and recommendation systems. The current user profiles acquiring techniques however suffer from some problems and thus demand to improve. In this paper, a survey of the existing user profiles acquiring mechanisms is presented first, and a novel approach is introduced that uses pseudo-relevance feedback to acquire user profiles from the Web. The related evaluation result is promising, where the proposed approach is compared with a manual user profiles acquiring technique.

Key words: Pseudo-Relevance Feedback, User Profiles, User Information Needs, Personalized Web Information Gathering

1 Introduction

Definition 1. Let $\mathbb{S}$ be a set of subjects, an element $s \in \mathbb{S}$ is formalized as a 5-tuple $s := \langle label, type, neighbor, ancestor,$
$descendant \rangle$, where

- $label$ is the subject heading of s in the LCSH authorities, and $label(s) = \{t_1, t_2, \ldots, t_n\}$*;*
- $type$ is the topic type of s in the LCSH authorities, and $type(s) \in \{topical, geographic, corporate\}$*;*
- $neighbor$ is a function returning the subjects that have direct links to s in the LCSH thesaurus, and $neighbor(s) = \{s_i | s_i \neq s, s_i \in \mathbb{S}\}$ and $neighbor(s) \subseteq \mathbb{S}\}$*;*
- $ancestor$ is a function returning the subjects that have a higher level of abstraction than s and link to s directly or indirectly in the LCSH thesaurus, and $ancestor(s) = \{s_j | s_j \neq s, s_j \in \mathbb{S}\}$ and $ancestor(s) \subseteq \mathbb{S}\}$*;*
- $descendant$ is a function returning the subjects that are more specific than s and link to s directly or indirectly in the LCSH thesaurus, and $descendant(s) = \{s_k | s_k \neq s, s_k \in \mathbb{S}\}$ and $descendant(s) \subseteq \mathbb{S}\}$. □

In the past decades the information available on the Web has exploded rapidly. The Web information covers a wide range of topics and serves a broad spectrum of communities [1]. How to gather user needed information from the Web, however, becomes challenging. The Web information gathering and recommendation systems need to capture user information needs in order to deliver Web users useful and meaningful information. For this purpose, users profiles are used by many personalized Web information gathering and recommendation systems [6, 12, 13].

User profiles specify interesting topics and personal preferences of Web users, and are key in Web personalization to capture Web user information needs [13]. However, effectively acquiring user profiles is difficult. Some techniques acquire user profiles by interviewing users or requesting users to fill the questionnaires [17, 25]. Some techniques acquire user profiles by giving users a set of documents to read and feedback relevant/non-relevant for user information needs [19]. These mechanisms are inefficient. Some other techniques acquire user profiles from a collection of user desktop documents like browsing history [6, 20, 14]. However, their acquired user profiles contain noise and uncertainties. Therefore, the current user profile acquiring mechanisms demand to improve for their effectiveness and efficiency.

In this paper, a survey of the existing user profiles acquiring mechanisms is first performed, which categorizes the user profiles acquiring mechanisms into three groups of interviewing, semi-interviewing, and non-interviewing techniques. After that, a user profiles acquiring approach is proposed using the pseudo-relevance feedback technique. The proposed approach analyzes the semantic of topics, and uses the topic-related subjects to perform an initial search on the Web. The retrieved Web documents are filtered and assigned support values, based on the belief of their contents to the given topics. The documents are assumed to be the samples feedback by users, and their associated support values are more specific than only the binary values provided by users in real relevance feedback. The user profiles are then represented by these Web documents with support values. The evaluation result of the proposed approach is promising, where the approach is compared with a typical model implemented for the interviewing user profiles acquiring mechanisms. The proposed approach contributes to the personalized Web information gathering and recommendation systems that use user profiles.

The paper is organized as follows. Section 2 surveys the existing user profiles acquiring approaches and pseudo-relevance feedback methods, and Section 3 introduces the pseudo-relevance feedback user profile acquiring approach. The evaluation of the approach is discussed in Section 4. Finally, Section 5 makes the conclusions.

2 User Profiles Acquiring and Relevance Feedback

User profiles are used in Web information gathering for interpretation of query semantic meanings to capture information needs [6, 7, 13, 25]. User profiles may be represented by a set of documents that are interesting to the user [2], a set of terms [13], or a set of topics [6, 20] specifying the user interests and preferences. Kosala & Blockeel [9] pointed out that user profiles are important for the user modelling applications and personal assistants in Web information systems.

User profiles are defined by Li and Zhong [13] as the interesting topics of user information needs and the personal preferences of Web users. They also categorized user profiles into two diagrams: the data diagram and information diagram. The data diagram profiles are usually acquired by analyzing a database or a set of transactions [6, 13, 17, 20, 21]. The information diagram profiles are generated by using manual techniques such as questionnaires and interviews [17, 25], or by using information retrieval and machine-learning techniques [6, 18]. In order to acquire user profiles, Chirita *et al.* [4] and Teevan *et al.* [24] mined user interests from the collection of user desktop information e.g. text documents, emails, and cached Web pages. Makris *et al.* [16] comprised user profiles by a ranked local set of categories and then utilized Web pages to personalize search results for users. These works attempted to acquire user profiles by discovering user background knowledge first.

User profiles acquiring techniques can be categorized into three groups: the *interviewing*, *semi-interviewing*, and *non-interviewing* techniques. The interviewing user profiles are completely acquired using manual techniques; e.g. questionnaires, interviews, and user classifying training sets. One typical example is the TREC-11 Filtering Track training sets that are acquired manually by human-power effort [19]. Users read training documents and assigned positive or negative judgements to the documents against given topics. Based on the assumption that users know their interests and preference exactly, these training documents perfectly reflect user background knowledge. However, this kind of user profile acquiring mechanism is costly, as Web users have to invest a great deal of effort in reading the documents and providing their opinions and judgements. Aiming to reduce user involvement, semi-interviewing user profiles are acquired by semi-automated techniques. These techniques usually provide users with a list of categories, and explicitly ask users for their interested or non-interested categories. One typical example is the model developed by [23] that uses a world knowledge base to learn personalized ontologies, and acquires user profiles from user local instance repository. The limitation of semi-interviewing mechanism is that it largely relies on a knowledge base for user background knowledge specification. Non-interviewing techniques do not involve users directly but ascertain their interests instead. Such user profiles are usually acquired by observing and mining knowledge from users' activity and behavior [25]. Typical models are [6] and [20]'s ontological user profiles, and also models developed by [8, 14, 16]. They acquired user profiles adaptively based on the content of user queries and online browsing history. The non-interviewing mechanism however, is ineffective. Their user profiles usually contain noise and uncertainties. The current user profiles acquiring mechanisms demand to improve.

Pseudo-relevance feedback (also called blind feedback) techniques are widely used in information retrieval to improve the performance of search systems. The systems using pseudo-relevance feedback initialize a search first and assume that the top-k returned documents are relevant as that feedback by users manually. Characteristics of the top-k documents are learned and used to add new or adjust weights of old search terms. The systems then generate the final result set using these evaluated search terms [15]. Many developed systems using pseudo-relevance feedback have been reported having achieved significant improvements in Web information gathering performance [3, 5, 10, 22, 26]. Alternatively, Lee *et al.* [10] clustered the retrieved documents to find dominant documents in order to emphasize the core concepts in a topic. Instead of treating each top document as equally relevant, Collins-Thompson and Callan [5] re-sampled the top documents retrieved in the initial search according to the relevance values estimated by probabilities. As a result, a document is more relevant if it is higher in the ranking. However, many systems using pseudo-relevance feedback focus on expending query terms only, but not on describing user interests in user profiles. Thus, a research gap remains there to improve user profiles acquiring by using pseudo-relevance feedback.

3 Pseudo-Relevance Feedback User Profiles Acquiring

3.1 Semantic Analysis of Topics

User information needs are usually expressed by users using short phrases that contain only limited information. Users may use different query terms because of user perspectives, terminological habits and vocabulary. If the concepts and semantic content of information needs can be specified, information needs can be captured, and thus more useful and meaningful information can be delivered to Web users.

Aiming to capture a user information need, the concept space referred by the information need, namely a topic and denoted as $\mathcal{T}$, is identified. Let $\mathbb{S}$ be a set of concepts, in which each element s is a subject and $s \in \mathbb{S}$. The concept space referred by a topic $\mathcal{T}$ can be described by two sets of positive subjects S^+ and negative subjects S^-. The positive subjects refer to the concepts that $\mathcal{T}$ can be best described and discriminated from others. The negative subjects refer to the concepts that may cause paradoxical or ambiguous interpretation of $\mathcal{T}$. Identifying the concept space referred by $\mathcal{T}$ is thus to extract the S^+ and S^- of topic $\mathcal{T}$.

The positive and negative subjects are manually identified, based on the descriptions and the narratives provided by users for the given topic. Depending on the level of subjects supporting or against the given topic, the positive subjects and negative subjects are identified with a support value $sup(s, \mathcal{T})$, which is measured by:

$$sup(s, \mathcal{T}) = MB(\mathcal{T}|s) - MD(\mathcal{T}|s). \tag{1}$$

where $MB(\mathcal{T}|s)$ is the belief (how strong s is for $\mathcal{T}$) and $MD(\mathcal{T}|s)$ is the disbelief (how strong s is against $\mathcal{T}$) of subject s to topic $\mathcal{T}$. When $MB(\mathcal{T}|s)$ is greater than $MD(\mathcal{T}|s)$, s supports $\mathcal{T}$ and becomes a positive subject. In contrast, when $MB(\mathcal{T}|s)$ is smaller than $MD(\mathcal{T}|s)$, s is against $\mathcal{T}$ and becomes a negative subject. In the preliminary study, the $MB(\mathcal{T}|s)$ and $MD(\mathcal{T}|s)$ were specified by the user manually, and the range of $sup(s, \mathcal{T})$ values is [-1,1]. Based on these, the positive and negative subjects can be defined by:

$$\begin{cases} s \in S^+ \text{ if } sup(s, \mathcal{T}) > 0; \\ s \in S^- \text{ if } sup(s, \mathcal{T}) \leqslant 0. \end{cases} \tag{2}$$

Drawing a boundary line for the positive and negative subjects is difficult, because uncertainties may exist in these subject sets. The overlapping space between S^+ and S^- is considered negative, and the concept space referred by $\mathcal{T}$ can be defined as:

$$space(\mathcal{T}) = S^+ - (S^+ \cap S^-). \tag{3}$$

3.2 Acquiring User Profiles

User profiles in this paper are represented by training document sets, which is one of the common representations of user profiles in Web information gathering [13]. A training set usually consists of some positive and negative samples. Thus, the positive samples are the documents containing the topic relevant concepts, and the negative samples are those containing the paradoxical and ambiguous concepts of the topic.

The user profiles are acquired by using pseudo-relevance feedback technique. The initial search is performed by using a Web search agent to retrieve training documents from the Web. For a given topic, a set of queries can be generated based on the specified positive and negative subjects, where each s generates a query. The training documents are retrieved by using these $s \in S^+$ and $s \in S^-$, and assumed as the feedback by users.

The level of candidates supporting or against the given topic needs to be evaluated, as treating each top-k retrieved documents equally relevant is not adequate [5]. The level of training documents supporting or against the given topic may vary depending on (i) the performance of the search agent, (ii) the document's ranking in the returned list, and (iii) the support value of subject s that generates the query to retrieve the document. The documents with higher support values are more relevant to the topic.

The performance achieved by a Web search agent can be measured by using a training query and investigating the search results. Denoting a Web search agent's precision performance by $\wp$, the performance is measured by $\wp(\kappa) = \frac{|D_\kappa^+|}{\kappa}$, where $|D_\kappa^+|$ is the number of relevant documents in total κ number of documents retrieved, and $|D_\kappa^+| \leqslant \kappa$. The higher $\wp$ means the better ability of retrieving relevant documents.

The support values are also influenced by the documents' ranking positions in the list returned by the Web search agent. Although the retrieving algorithms used by Web search agents are in black box, the ranking position of returned documents is a solid evidence from the search agents for their relevance. The higher ranking documents are more likely to be relevant to the topic, and thus have better chance to be marked by users as "relevant" if in real user feedback [5].

Based on the previousely discussed three factors, using Eq. (1) and (2), the support value sup of a document d to $\mathcal{T}$ can be measured by:

$$sup(d, \mathcal{T}) = \sum_{s \in S^+ \cup S^-} sup(d, s) \times sup(s, \mathcal{T}); \tag{4}$$

where $sup(d, s)$ is the support value of d to s, which is calculated by:

$$sup(d, s) = \beta \times \wp(\kappa) \times (1 - \frac{r(d, D) mod(k)}{k}); \tag{5}$$

β has value $[0|1]$ for the occurrence of d in the document set D retrieved by using s. Thus, if $d \notin D$, $sup(d, s) = 0$. $r(d, D)$ is d's ranking in D determined by the Web search agent, and k is a constant number of documents in each cutoff in κ, e.g. $k = 10$.

According to Eq. (2), $s \in S^+$ gives positive $sup(s, \mathcal{T})$ values and $s \in S^-$ gives negative $sup(s, \mathcal{T})$ values, Eq. (4) finally assigns the training documents positive or negative values and classifies them into positive or negative sets:

$$\begin{cases} D^+ = \{d, |sup(d, \mathcal{T}) > 0\} \\ D^- = \{d, |sup(d, \mathcal{T}) \leqslant 0\} \end{cases} \tag{6}$$

4 Evaluation

4.1 Experiment Designs

The experiment design is as follows. The PREF model implemented for the proposed approach was compared with the TREC model acquiring user profiles manually. For a given topic, each model acquired a user profile by using their own approach. The user profiles were represented by training sets, each consisting of a set of positive documents D^+ and negative documents D^-. Each document d held a support value $sup(d, \mathcal{T})$ to the given topic. The different profiles were used by a common system to retrieve information from the testing set. The performance of the gathering system then relied on the profiles input by the PREF and TREC models. Based upon this, we could compare the quality of acquired user profiles and thus evaluate the proposed model.

The PREF model was the implementation of the approach proposed in this paper using pseudo-relevance feedback. The PREF model acquired user profiles from the Web using Google API [1]. As discussed in Section 3, for each experimental topic a set of positive and negative subjects were first specified manually. These subjects were then used to retrieve the candidate positive and negative documents via the Google API. The precision performance of Google API was investigated first and set as $\{0.9, 0.8, \dots, 0.0\}$ for cutoff level κ of top $\{10, 20, \dots, 100\}$ retrieved documents, where $k = 10$ as in Eq. (5). The $sup(d, \mathcal{T})$ values of candidates were calculated, and the retrieved documents were filtered and re-classified.

The TREC model was the implementation of interviewing user profiles acquiring mechanism, as discussed in Section 2. For each topic, the author of that topic in TREC[2] was given a set of documents in to read, and then to judge relevance or non-relevance to the topic. The combined set of judged documents was used as the training set for that topic [19]. The topics were created by the same authors who performed the relevance assessments for these topics as well. Thus, the TREC training sets reflect the users' interests in the topics, under the assumption that only users know their interests exactly. In the TREC model, these TREC training sets were used as the user profiles. Because users read and judged these documents manually and their decision making precess remained in black box, we valued the positive documents $sup(d, \mathcal{T}) = 1$ and negative documents $sup(d, \mathcal{T}) = -1$, as the full values.

The information gathering system was implemented based on Li & Zhong's model [13], which uses user profiles for information gathering. The model was chosen because not only it is verified better than the *Rocchio* and *Dempster-Shafer* models, but also it is extensible in using support values of training documents. The input support values, $sup(d, \mathcal{T})$, associated with documents d for topic $\mathcal{T}$ affects the system's performance sensitively. The technical details and the related justifications can be referred to [13].

The Reuters Corpus Volume 1 (RCV1) [11] was used as the testing set, which is a large XML document set (806,791 documents) with a great topic coverage. A set of topics were designed by the TREC linguists manually [19], in which each topic had title, description and narrative specified. The topics R101-115 were used in our experiments.

The performance of the experimental models was measured by the precision averages at eleven standard recall levels (11SPR) [19]. The 11SPR is the interpolated precision values against recall levels. The mechanism of interpolating precision at standard recall level $\lambda \in \{0.0, 0.1, 0.2, \dots, 1.0\}$ is to use the maximum precision obtained for each of the N topic for any actual recall level greater or equal to λ. The 11SPR is calculated by $\frac{\sum_{\lambda=1}^{N} precision_\lambda}{N}$.

[1] http://www.google.com

[2] Text REtrieval Conference, http://trec.nist.gov/.

4.2 Results and Discussions

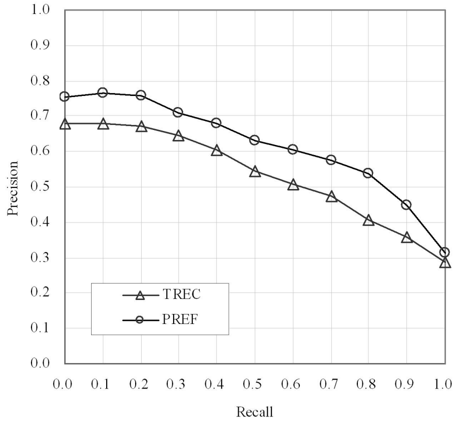

Fig. 1. Experimental 11SPR Results

As the experimental 11SPR results shown in Fig. 1, the PREF model outperformed the TREC, and the proposed user profiles acquiring approach is promising.

The user interests contained in the PREF user profiles had better coverage than that in the TREC profiles. In the TREC model, the user profiles were acquired manually by users reading and judging the training documents for the topics. This procedure ensured that the training documents were judged accurately, however, the coverage of user interests was weakened. Firstly, the number of documents retrieved from RCV1 and provided to the TREC linguists to read and judge was limited. In the experiments, on average there were about 70 training documents acquired for each topic in the TREC model, whereas in the PREF model, this average number was 200. Scondly, the training documents in the PREF model were acquired from the Web, and Web information covers a wide range of topics and serves a broad spectrum of communities [1]. Thus, comparing to the TREC model, the PREF user profiles had better user background knowledge coverage.

The PREF user profiles had more specific support values associated with the training documents. In the TREC model, only “positive” or “negative” could be chosen when the TREC linguists read a document. The top support value of 1 and 0 was assigned to the training documents. In case of that only a part of content in a document was relevant, useful information might be missing if the document was judged “negative”, and noisy information might be acquired if the document was judged “positive”. As a result, some user interests were missed and noisy information was obtained when acquiring user profiles. The PREF model, on the other hand, assigned the float support values to the training documents, depending on their specific relevance to the given topics. Therefore, comparing to the TREC model, the PREF user profiles had more specific support values associated. Moreover, the information gathering system commonly used in the experiments was sensitive to the input support values associated with the training documents. This leveraged the PREF model's performance as well.

5 Conclusions

In this paper, a survey of the existing user profiles acquiring techniques has been conducted. The current mechanisms are categorized into three groups of interviewing, semi-interviewing, and non-interviewing techniques. A novel user profiles acquiring approach has also been introduced in the paper, which analyzes the semantic of user information needs first and acquires user profiles from the Web using the pseudo-relevance feedback technique. The proposed approach was evaluated successfully in the experiments, by comparing with a typical model implemented for the interviewing mechanisms and acquiring user profiles manually. The proposed approach contributes to the personalized Web information gathering and recommendation systems.

References

1. G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. The MIT Press, 2004.
2. K. D. Bollacker, S. Lawrence, and C. L. Giles. A system for automatic personalized tracking of scientific literature on the Web. In *Proc. of DL'99*, pages 105–113, 1999.
3. G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. of SIGIR'08*, pages 243–250, 2008.
4. P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the Web. In *Proc. of SIGIR'07*, pages 7–14, 2007.
5. K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proc. of SIGIR'07*, pages 303–310, 2007.
6. S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1(3-4):219–234, 2003.
7. J. Han and K.C.-C. Chang. Data mining for Web intelligence. *Computer*, 35(11):64-70,2002.
8. J. D. King, Y. Li, X. Tao, and R. Nayak. Mining World Knowledge for Analysis of Search Engine Content. *Web Intelligence and Agent Systems*, 5(3):233–253, 2007.
9. R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, 2000.
10. K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proc. of SIGIR'08*, pages 235–242, 2008.
11. D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397, 2004.
12. Y. Li and N. Zhong. Web Mining Model and its Applications for Information Gathering. *Knowledge-Based Systems*, 17:207–217, 2004.
13. Y. Li and N. Zhong. Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.
14. F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40, 2004.
15. T. R. Lynam, C. Buckley, C. L. A. Clarke, and G. V. Cormack. A multi-system analysis of document and term selection for blind feedback. In *Proc. ofCIKM'04*, pages 261–269, 2004.

16. C. Makris, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis. Category ranking for personalized search. *Data & Knowledge Engineering*, 60(1):109–125, January 2007.
17. S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, 2004.
18. A-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proc. of HLT '05*, pages 339–346, Morristown, NJ, USA, 2005.
19. S. E. Robertson and I. Soboroff. The TREC 2002 filtering track report. In TREC, 2002.
20. A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proc. of CIKM'07*, pages 525–534, New York, NY, USA, 2007. ACM.
21. K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proc. of WWW'04*, pages 675–684, 2004.
22. T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proc. of SIGIR'06*, pages 162–169, 2006.
23. X. Tao, Y. Li, N. Zhong, and R. Nayak. Ontology mining for personalzied web information gathering. In *Proc. of WI'07*, pages 351–358, 2007.
24. J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proc. of SIGIR'05*, pages 449–456, 2005.
25. J. Trajkova and S. Gauch. Improving ontology-based user profiles. In *Proc. of RIAO'04*, pages 380–389, 2004.
26. S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proc. of WWW'03*, pages 11-18, 2003.