

Adopting Relevance Feature to Learn Personalized Ontologies

Yan Shen, Yuefeng Li, and Yue Xu

Electrical Engineering, Computer Science, Science and Engineering Faculty,
Queensland University of Technology, Australia
{y12.shen,y2.li,yue.xu}@qut.edu.au

Abstract. Relevance feature and ontology are two core components to learn personalized ontologies for concept-based retrievals. However, how to associate user native information with common knowledge is an urgent issue. This paper proposes a sound solution by matching relevance feature mined from local instances with concepts existing in a global knowledge base. The matched concepts and their relations are used to learn personalized ontologies. The proposed method is evaluated elaborately by comparing it against three benchmark models. The evaluation demonstrates the matching is successful by achieving remarkable improvements in information filtering measurements.

Keywords: Relevance Feature, Specificity Term, Ontology, Local Instance, Global Knowledge Base, Concept Matching.

1 Introduction

Information overload is continuously a hard problem for Information Retrieval (IR). Current retrieval methods acquire a precise query upfront to express search intents, which is fairly difficult for any user who has no background knowledge or past experience. Relevance Feature (RF) has been alternatively taken into account to indicate user needs in data mining methods [1] [2]. Generally, it comprises of patterns, terms and their weights while the weights are used to measure relevance. The feature extractions are mainly based on two aspects: 1) relevance feedback, where a user judges the results from previous retrieval and tells the system which documents are relevant or not [3]; and 2) pattern mining, which is an efficient technique against previous retrievals to filter out noisy by studying frequent, closed, and closed sequential patterns. Note that the extracted features are useful to point out user preferences.

Ontology-based technique has been recognized as a critical part of advanced search over last decade. It assists to refine search intents within specific domains and access new knowledge by tracking semantic relations. Recently, some researchers [4] have attempted to build ontological user profiles according to discovered user background knowledge. Some of the knowledge is integrated by both global and local analyses [5]. The global employs a global knowledge base (ontologies, thesauruses, or online knowledge bases) that mirrors the content of the

world wide web for common knowledge representation, whereas the local investigates native information or observes user behaviours from local instances [6]. A same motivation is to produce personalized ontologies so as to better understand user needs by concepts.

A key challenge here is how to accurately match local information and global knowledge. The basic idea of existing method [6] is to use a conditional probability to determine relevant concepts for describing a theme of local instances. However, this technique usually obtains a low performance because of the mismatch problem [4]. Mismatch means some relevant specific concepts have been omitted. The problem of mismatch occurs when the popular features (useful or high frequent features) match many relevant but general concepts (usually appear in the top of the ontology); but do not match some relevant specific concepts (usually appear at the bottom of the ontology). This paper presents an alternative method to address the mismatch problem. The solution aims to learn personalized ontologies for user knowledge understanding. A pattern mining method is developed to discover RF from local instances automatically. All extracted terms in the RF are classified into three groups: Positive specific (SPE), General, and Negative specific respectively. Library of Congress Subject Headings (LCSH) is learned as a global knowledge base. The SPE terms are selected to match subject headings in the LCSH. The attempt-to-proposed approach is evaluated by comparing against three benchmark models with a standard Reuters Corpus Volume 1 (RCV1) testing set. The experimental results show the proposed matching is successful and achieve significant improvements in information filtering measurements. This research contributes to ontological user profiling and knowledge engineering. The related outputs are critical when systems expect to return proper search results and provide personalized services.

The rest of the paper is organized as below. Section 2, significant related work is involved; Section 3 outlines an overview of proposed approach, and requisite definitions are provided; Section 4 describes the extraction of RF; Section 5 introduces concept matching method; Section 6 shows conducted experiments and results; The paper is concluded by Section 7.

2 Related Work

2.1 Relevance Feature Extraction

Feature extraction is a fundamental stage for the majority of IR models. One of the most popular methods is TF*IDF. It uses keywords as elements in the vector of the feature space. The bag of words can be obtained by diverse term weighting approaches. However, these approaches struggle against the problem of selecting appropriate number of features among an enormous set of terms to guarantee the retrieval efficiency [7], or said over-fitting. In addition, the corresponding approaches are restricted on term statistics in the entire collection, but do not take relevance information (e.g. user feedback) into account. As an extension of term-based methods, pattern mining techniques are investigated by data mining communities for many years. These techniques extract useful patterns from large

data collections instead of terms. Some studies categorize patterns into frequent, closed, and closed sequential. Simultaneously, they measure the specificity of patterns explicitly according to relevance feedback from human beings. Pattern taxonomy model (PTM) was first introduced by Wu et al. [1]. It improves search performance by using both frequent and closed patterns. Later on, it has been further researched by [2] to develop a two-stage model for irrelevant information filtering. Li et al. [3] build the Relevance Feature Discovery (RFD) model by mining patterns from positive and negative documents respectively. The related output shows potentials for enhancing information filtering and user profiling.

2.2 Ontology-Based Techniques

Since ontology-based techniques can bring more discriminative and arguably capabilities by carrying “semantics”, people commenced to build user profiles by accessing common knowledge. Li and Zhong [8] develop a term-based ontology leaning method to automatically discover ontologies from data sets in order to understand user information needs by concepts. Gauch et al. [9] use ontology references based on the categorization of online portals and propose to learn personalized ontologies for users. Developed by King et al. [4], IntelliOnto is built based on the Dewey Decimal Classification (DDC) system and attempt to describe the background knowledge. Sieg et al. [5] utilize ontological user profile on the basis of the user’s interaction with a concept hierarchy which aims to extract the domain knowledge. More recently, Tao et al. [6] propose an ontology-based knowledge retrieval framework, namely ONTO model, to capture user information needs by analyzing general knowledge and local instance repository. However, they disregard the work of matching and assume the local information can be ideally referred to the proper concepts in a global knowledge base.

Of all related work, the process in [6] seems the most similar to ours but two differences are: 1) our study adopts relevance feedback as user information needs rather than asking users to specify their needs manually, and 2) provides a sound solution to cope with local information and global knowledge mismatch.

3 Design and Definitions

Figure 1 illustrates a design of the proposed approach. Local instances as input consist of two parts: relevance feedback and a set of training documents. The relevance feedback is initialized as positive and negative. The positive indicates the document is relevant to user interests, whereas the negative is irrelevant. By applying the following method discussed in subsection 4.1, the extracted relevance features can be categorized into three groups: SPE, General, and Negative. Here, the approach only selects terms in the SPE group as candidates because they contain more topic-related interests rather than other groups [3]. LCSH, which is a thesaurus of subject headings, is chosen to be the global knowledge representation. Each subject heading denotes a concept among the knowledge base. The concept is a short phrase where contains one or couple of terms. We match

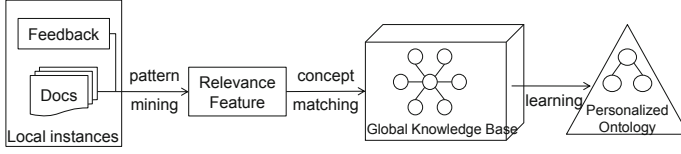


Fig. 1. Overview for the Entire Approach

the SPE terms to the subject headings, and personalized ontologies learning will eventually take advantage of the matched concepts and their original relations in the LCSH.

3.1 Definitions about Patterns and Closed Patterns

Relevance feature discovery is to find useful features, including patterns, terms and their weights, in a training set D , which consists of a set of positive documents, D^+ , and a set of negative documents, D^- . In this paper, we assume that all documents are split into paragraphs. So a given document d yields a set of paragraphs $PS(d)$.

Let $T = \{t_1, t_2, \dots, t_m\}$ be a set of terms which are extracted from D^+ . Given a *termset* X , a set of terms, in document d , $coverset(X) = \{dp | dp \in PS(d), X \subseteq dp\}$. Its *absolute support*

$$sup_a(X) = |coverset(X)|;$$

and its *relative support*

$$sup_r(X) = \frac{|coverset(X)|}{|PS(d)|}.$$

A termset X is called *frequent pattern* if its sup_a (or sup_r) $\geq min_sup$, a minimum support. Given a set of paragraphs $Y \subseteq PS(d)$, we can define its *termset*, which satisfies

$$termset(Y) = \{t | \forall dp \in Y \Rightarrow t \in dp\}.$$

Let $Cls(X) = termset(coverset(X))$ be the closure of X . We call X *closed* if and only if $X = Cls(X)$. Let X be a closed pattern. We have

$$sup_a(X_1) < sup_a(X) \tag{1}$$

for all pattern $X_1 \supset X$.

Closed Sequential Patterns. A sequential pattern $s = \langle t_1, \dots, t_r \rangle$ ($t_i \in T$) is an ordered list of terms. A sequence $s_1 = \langle x_1, \dots, x_i \rangle$ is a sub-sequence of another sequence $s_2 = \langle y_1, \dots, y_j \rangle$, denoted by $s_1 \sqsubseteq s_2$, iff $\exists j_1, \dots, j_i$ such that $1 \leq j_1 < j_2 < \dots < j_i \leq j$ and $x_1 = y_{j_1}, x_2 = y_{j_2}, \dots, x_i = y_{j_i}$. Given $s_1 \sqsubseteq s_2$, we usually say s_1 is a sub-pattern of s_2 , and s_2 is a super-pattern of s_1 . In the following, we refer to sequential patterns as patterns.

Given a sequential pattern X in document d , $\text{coverset}(X)$ is still used to denote the covering set of X , which includes all paragraphs $ps \in PS(d)$ such that $X \subseteq ps$, i.e., $\text{coverset}(X) = \{ps | ps \in PS(d), X \subseteq ps\}$. Its *absolute support* and *relative support* are defined as the same as for the normal patterns.

A sequential pattern X is called a *frequent pattern* if its relative support $\geq \text{min_sup}$, a minimum support. The property of closed patterns (see Eq. (1)) can be used to define closed sequential patterns. A frequent sequential pattern X is called *closed* if not $\exists$ any super-pattern X_1 of X such that $\text{sup}_a(X_1) = \text{sup}_a(X)$.

3.2 Global Knowledge Base: LCSH

Global knowledge is the common-sense knowledge acquired by people based on their experience and education. LCSH is an ideal global knowledge representation because of a rich vocabulary is used to cover all subject areas. In the LCSH, subject headings are basic elements to convey knowledge in the format of concept, they have three main types of references: *Broader Term* (BT), *Narrower Term* (NT) and *Related Term* (RT). Related definitions are clarified as follows. Definition of Subject headings: Let C denote a set of subject headings in LCSH, a subject $s \in C$ is formalized as a pair $(\text{label}, \text{reference})$, where

- *label* is the heading of s in LCSH thesaurus;
- *reference* is a statement regarding all references that the subject c has.

The LCSH world knowledge base can be formalized as:

Definition of World Knowledge Base: A world knowledge base ontology is a directed acyclic graph structure defined as a pair $\Theta := (C, R)$, consisting of

- C is a set of subjects in LCSH $C = \{s_1, s_2, \dots, s_n\}$;
- R is the semantic relations $R = \{\text{ref}_1, \text{ref}_2, \dots, \text{ref}_n\}$ existing among the subjects in C .

4 Relevance Feature Acquisition

In general, the concept of relevance is subjective. People can easily determine the relevance of a topic (or document) in specificity or generality. However, it is hard to use these concepts for interpreting relevance features in text documents. This section first discusses how to use the concepts for understanding the different roles of the low-level feature terms for answering what users want. We also present the ideas for accurately weighting terms based on their specificity and distributions in the discovered higher level features. Finally, we describe algorithms for both the discovery of higher level features and the revision of weights of low-level terms.

4.1 Specificity of Low-Level Features

A term's specificity describes the extent of the term to which the topic focuses on what users want. It is very difficult to measure the specificity of terms because a

term's specificity depends on users' perspectives for their information needs [6]. The paper discusses how terms are grouped into three groups (SPE, general, and negative specific terms) based on their appearances in a training set. Given a term $t \in T$, its $coverage^+$ is the set of positive documents that contain t , and its $coverage^-$ is the set of negative documents that contain t . We assume that terms frequently used in both positive documents and negative documents are general terms. Therefore, we want to classify terms that are more frequently used in the positive documents into the positive specific category; and the terms that are more frequently used in the negative documents into the negative specific category. Based on the above analysis, we define the *specificity* of a given term t in the training set $D = D^+ \cup D^-$ as follows:

$$spe(t) = \frac{|coverage^+(t)| - |coverage^-(t)|}{n}$$

where $coverage^+(t) = \{d \in D^+ | t \in d\}$, $coverage^-(t) = \{d \in D^- | t \in d\}$, and $n = |D^+|$. $spe(t) > 0$ means that term t is used more frequently in positive documents than in negative documents. We present the following classification rules for determining the general terms G , the SPE terms T^+ , and the negative specific terms T^- :

$$G = \{t \in T | \theta_1 \leq spe(t) \leq \theta_2\},$$

$$T^+ = \{t \in T | spe(t) > \theta_2\}, \text{ and}$$

$$T^- = \{t \in T | spe(t) < \theta_1\}.$$

where θ_2 is an experimental coefficient, the maximum bound of the specificity for the general terms, and θ_1 is also an experimental coefficient, the minimum bound of the specificity for the general terms. We assume that $\theta_2 > 0$ and $\theta_2 \geq \theta_1$. It is easy to verify that $G \cap T^+ \cap T^- = \emptyset$. Therefore, $\{G, T^+, T^-\}$ is a partition of all terms. To describe relevance features for a given topic, we believe that specific terms are very useful for the topic in order to distinguish it from other topics.

4.2 Term Weighting

In Pattern Taxonomy Model (PTM), relevance features are discovered from a set of positive documents. To improve the efficiency of the PTM, an algorithm, *SPMining*(D^+, min_sup) [10], was proposed (also used in [1]) to find closed sequential patterns for all documents $\in D^+$, which used the well-known *Apriori* property in order to reduce the searching space. For all positive documents $d_i \in D^+$, the *SPMining* algorithm can discover all closed sequential patterns, SP_i , based on a given *min_sup*. (We omit this algorithm to save space.)

Let $SP_1, SP_2, \dots, SP_n$ be the sets of discovered closed sequential patterns for all documents $d_i \in D^+ (i = 1, \dots, n)$, where $n = |D^+|$. For a given term t , its support (or called *weight*) in discovered patterns can be described as follows:

$$support(t, D^+) = \sum_{i=1}^n \frac{|\{p | p \in SP_i, t \in p\}|}{\sum_{p \in SP_i} |p|} \quad (2)$$

where $|p|$ is the number of terms in p . After the supports of terms have been computed from the training set, the following rank will be assigned to every incoming document d to decide its relevance:

$$rank(d) = \sum_{t \in T} weight(t) \tau(t, d) \quad (3)$$

where $weight(t) = support(t, D^+)$; and $\tau(t, d) = 1$ if $t \in d$; otherwise $\tau(t, d) = 0$.

Because of many noises in the discovered patterns (an inherent disadvantage of data mining), the evaluated supports are not accurate enough. To improve the effectiveness of PTM, we use negative documents in the training set in order to remove the noises. If a document's rank (see Eq. (3)) is less than or equals to zero, this document is clearly negative to the system. If a negative document has a high rank, the document is called an offender [8] because it forces the system to make a mistake. The offenders are normally defined as the top- K negative documents in a ranked set of negative documents, D^- . The basic hypothesis is that the relevance features should be mainly discovered from the positive documents. Therefore, in our experiments, we set $K = \frac{n}{2}$, the half of the number of positive documents.

Once we select the top- K negative documents, the set of negative document D^- will be reduced to include only K offenders (negative documents). The next step is to classify terms into three categories, G , T^+ , and T^- , based on D^+ and the updated D^- . We can easily verify that the experimental coefficients θ_1 and θ_2 satisfy the following properties if $K = \frac{n}{2}$:

$$0 \leq \theta_2 \leq 1, \text{ and } -\frac{1}{2} \leq \theta_1 \leq \theta_2.$$

Here, we show the basic process of revising discovered features in a training set. This process can help readers to understand the proposed strategies for revising weights of low-level terms in different categories. Formally, let DP^+ be the union of all discovered closed sequential patterns in D^+ , DP^- be the union of all discovered closed sequential patterns in D^- and T be the set of terms that appear in DP^+ or DP^- , where a closed sequential pattern of D^+ (or D^-) is called a *positive pattern* (or *negative pattern*).

It is obviously that $\exists d \in D^+$ such that $t \in d$ for all $t \in T^+$ since $spe(t) > \theta_2 \geq 0$ for all $t \in T^+$. Therefore, for each $t \in T^+$, it can obtain an initial weight by the deploying method on D^+ (using the higher level features, see Eq. (2)).

For the term in $(T^- \cup G)$, there are two cases. If $\exists d \in D^+$ such that $t \in d$, t will get its initial weight by using the deploying method on D^+ ; otherwise it will get a negative weight by using the deploying methods on D^- .

The initial weights of terms finally are revised according to the following principles: increment the weights of the SPE terms, decline the weights of the negative specific terms, and do not update the weights of the general terms. The details are described as follows:

$$weight(t) = \begin{cases} w(t) + w(t) \times spe(t), & \text{if } t \in T^+ \\ w(t), & \text{if } t \in G \\ w(t) - |w(t) \times spe(t)|, & \text{if } t \in T^- \end{cases} \quad (4)$$

where w is the initial weight(or the support in Eq. (2)).

5 Concept Matching Method

The matching occurs between terms mined from previous sections and subjects appearing $s \in C$ in LCSH. Note that the mined terms are distinguished into three groups and their values can be calculated by the weight function in Eq. (4). In this method, we use only terms in the first scenario where $t \in T^+$ for matching the concepts among the LCSH. The SPE terms are sufficient and indispensable to represent user desired information according to our prior testing. The matching method is divided into several steps below:

1) To gather the importance of terms in T^+ , we first sort them based on weight values from Eq. (4). The action aims to identify the significant terms from text mining perspective and highlight them.

2) The second step consists of two phases. At phase one, we get the top 25% terms in T^+ based on their values. They are considered as the core portion of local information. Consequently, four most relevant subjects $s \in C$ are referred to each of the top 25% terms. Phase two, the relevance of subjects is computed by $rel(s) = |T^+ \cap s|/|s|$, where $|T^+ \cap s|$ denotes the number of overlapped terms between T^+ and subject s , $|s|$ stands for the total number of terms in the subject s . According to the rel values, four subjects can be confirmed. The idea of choosing 25% and 4 parameters is supported by empirical experiments. The matching performance can be impacted while modifying these parameters. A comparison of using other parameters will be shown in evaluation.

Algorithm 1. Concept Matching Algorithm

Input:

A set of SPE terms T^+ ; $weight(t)$ from Eq. (4); a set of LCSH subjects C .

Output:

A set of matched concepts SC .

- 1: Let $SC = \emptyset$;
 - 2: Sort T^+ using $weight(t)$ as descendant order;
 - 3: Let $K = |T^+|/4$, Let T_1^+ be the top- K terms in T^+ ;
 - 4: for each $t \in T_1^+$ {
 - 5: Let $s_1 = s_2 = s_3 = s_4 = t$;
 - 6: Let $c(t) = \{s \in C | t \in s\}$;
 - 7: Select the top-4 relevant concepts in $c(t)$ using $rel(s)$, and let s_1, s_2, s_3, s_4 be the subjects; //if $|c(t)| < 4$, t will be the default value of s_i because of step 5
 - 8: Let $SC = SC \cup \{s_1, s_2, s_3, s_4\}$ }
 - 9: for each $t \in (T^+ - T_1^+)$ {
 - 10: Let $s_1 = t$, $rel(s_1) = 0$;
 - 11: for each ($s \in C \& t \in s$)
 - 12: if $\frac{|s \cap T^+|}{|s|} > rel(s_1)$ then
 - 13: Let $s_1 = s$, $rel(s_1) = \frac{|s \cap T^+|}{|s|}$;
 - 14: Let $SC = SC \cup \{s_1\}$ }
-

3) This step is for the rest of terms (75%) in T^+ . To find out accurate subject for each of these terms, we select the most relevant one based on rel value instead of four.

4) A set of specified concepts and their references existing in the LCSH can be obtained to form a personalized ontology. By taking advantage of the ontology, a scope of user background knowledge can be defined, and search systems can offer tailoring results after understanding precise user preferences.

In most cases, a term $t \in T^+$ can successfully find a subject or a set of subjects $s = \{s | s \in C\}$ in the LCSH. It is a hard issue when a term cannot match concepts in the global knowledge base. For example, “dutroux” is not a valid word/term in vocabulary but appears frequently in the training documents. It should be important to describe user needs. However, no subject can be matched by the proposed method. To overcome this issue, we count the term itself as a subject directly as if $c(t) = \{s \in C | t \in s\} = \emptyset$.

Algorithm 1 illustrates an entire process for our concept matching method, which facilitates to repeat and optimize the work. Note that the output is a set of specific concepts SC , the process can be also understand as a transition of informative descriptor and conceptional descriptor. These acquired concepts and their semantic relations in LCSH are used to construct personalized ontologies.

6 Evaluation

The hypothesis in this paper is that the SPE terms extracted from RF contain user focusing needs. By adopting the RF for ontology matching, gathered concepts should be helpful to improve search effectiveness. Related experiments were conducted to support this hypothesis. This section states data collections, baseline models, information filtering measurements, and results.

6.1 Data Collections

A LCSH database was selected to build the world knowledge base. Its size is 719 mega bytes stored in Microsoft Office Access. Initially, 491,250 subject headings and their internal references between the headings were extracted.

RCV1 corpus consists of all and only English language stories produced by Reuter’s journalists between August 20, 1996, and August 19, 1997, a total of 806,791 documents that cover very large topics and information. TREC (2002) has developed and provided 50 assessor topics [11] for the filtering track, aiming at building a robust filtering system. The relevance judgements on RCV1 have also been made by the assessors. The assessor topics are more reliable than any artificially constructed topics [12]. For each topic, some documents in RCV1 are divided into a training set and a testing set. According to Buckley and others [13], first 50 topics are stable and enough for high quality experiments. This research uses RCV1 and the first 50 assessor topics to evaluate the proposed model. Documents in RCV1 are marked in XML. To avoid bias in experiments, all the documents have been preprocessed by stemming and stop-words removal.

6.2 Baseline Models and Measurements

For evaluation, we employed three baseline models: 1) PTM, is the up-to-date pattern mining based model. It discovers sequential closed patterns from positive

documents, deploys discovered patterns on their terms using Eq. (2) and ranks documents using Eq. (3). For the pattern-based models, we rank a document based on the total relative supports of discovered patterns that appear in the document. We also set $min_sup = 0.2$ (relative support) for PTM; 2) TF*IDF, is a term-based model for comparison with new retrieval models. For each topic, we sorted terms in the positive documents according to original TF*IDF values and chose first 150 terms; 3) ONTO, is a ontology-based model developed by Tao et al. in 2011 [6]. The idea of the model is to use a similarity to determine relevant concepts c for describing the themes of the local instances, $P(c|F) = P(c \cap F)/P(F)$, where F is a set of features discovered in the local instances. The comparison is worth to indicate our discovered concepts are more specific for user requisitions. The proposed model named POM is estimated by five state-of-the art IR measuring metrics, including top 20 precision based on the relevance judgement in RCV1 ($top@20$), the precision averages at 11 standard recall levels ($11 - points$), the *Mean Average Precision* (MAP), the F_1 -measure (F_1), and the *breakeven point* (b/p).

Table 1. Comparison Results for Different Parameter Settings. (Refer to Section 5, the table shows the reason to predefine parameters as top 25% terms with 4 subjects and 1 for the rest. A number of combination settings like 50% 33%, and 25% with different numbers of subject have been tested but omitted here for space.)

#subjects	%	$top@20$	MAP	F_1	b/p
4	50	0.43	0.3967	0.4103	0.3880
	33	0.42	0.3941	0.4084	0.3855
	25	0.46	0.4124	0.4195	0.4042
3	50	0.44	0.4029	0.4141	0.3968
	33	0.44	0.4008	0.4125	0.4001
	25	0.45	0.4053	0.4157	0.3971
2					

6.3 Experiment Design and Results

To prove that our matched concepts can truly contain user information needs, all terms in the concepts are deployed to revise a new weight as follows:

$$weight(t, \Theta) = \sum_{t \in s, s \in SC} rel(s)/|s|$$

For the POM, we use the revised weights computed by Eq. (3) to rank documents for each topic in the RCV1 testing set. For the baseline models, we apply features extracted by their methods, and then keep their original term weight pairs to implement our matching process. After that, the matched concepts are collected and deployed by the same formula upon. To indicate the influences affected, *the percentage change in performance* is used to compute the difference, which is formulated as:

$$\%chg = \frac{Result_{POM} - Result_{baseline}}{Result_{baseline}} \times 100$$

The larger $\%chg$ is, the more significant improvement it achieves.

As shown in Table 2, POM achieves excellent performance with 16.12% (max 25.13% and min 9.7%) in percentage change on average for all four measures. Another big different is the number of extracted terms by different techniques. There are 23.12 SPE terms in average, which is approximately 7 times less than the number of terms extracted by TF*IDF (147.32) and PTM (154.82). This demonstrates that SPE terms are quantitatively enough to summarize all user needs with a few number of words. The matched concepts should capture all concrete user knowledge. Fig. 2 shows the 11 – *points* comparison where the POM model distinctly outperforms than others.

Table 2. Comparison Results for All Models in First 50 Topics

	#terms	top@20	MAP	F ₁	b/p
POM	23.12	0.46	0.4124	0.4195	0.4042
TF*IDF	147.32	0.391	0.35	0.378	0.3536
PTM	154.82	0.382	0.3649	0.3915	0.3652
ONTO	75.96	0.335	0.3403	0.3787	0.3355
$\%cha$		+25.13%	+17.3%	+9.7%	+15.16%

In sum, to address the problem local instances and global knowledge mismatch, the proposed solution has been demonstrated successfully based on the evaluation. The substantial results indicate our method can effectively discover desired concepts to represent user specific needs by adopting RF. The dramatical improvements are significant after comparing with three classic baseline models.

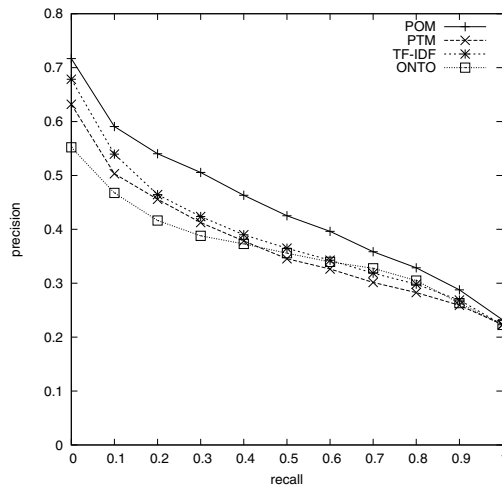


Fig. 2. 11 Points Result in First 50 Topics

7 Conclusion

This paper proposes a systematic method to build personalized ontologies by adopting relevance feature. The user orientated feature is used to represent user

information need and match concepts appearing in global knowledge base. It is an effective solution to solve the problem of local informative references and common knowledge mismatch. In evaluation, the standard topics and a huge testbed were employed for scientific experiments. The substantial results prove that the proposed model is reliable after comparing with baseline models.

In future, we plan to investigate the usage of learned ontologies by utilizing abundant semantic relations among concepts. Note that ontologies play an important role as a backbone to facilitate accessing information in knowledge management systems. The present work aims to define user wants in terms of extensive concepts in world knowledge. Besides, we are also interested in weight revising methods for performance enhancement. The investigation will extend the applicability for the majority of existing Web documents, and maximum the contribution of the present work.

References

1. Wu, S., Li, Y., Xu, Y.: Deploying approaches for pattern refinement in text mining. In: Proceedings of the 6th IEEE Conference on Data Mining (ICDM 2006), pp. 1157–1161 (2006)
2. Li, Y., Zhou, X., Bruza, P., Xu, Y., Lau, R.: A two-stage text mining model for information filtering. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008), pp. 1023–1032 (2008)
3. Li, Y., Algarni, A., Zhong, N.: Mining positive and negative patterns for relevance feature discovery. In: Proceedings of the 16th ACM SIGKDD Conference on Knowledge discovery and Data mining (KDD 2010), pp. 753–762 (2010)
4. King, J., Li, Y., Tao, X., Nayak, R.: Mining world knowledge for analysis of search engine content. *Web Intelligence and Agent Systems* 5(3), 233–253 (2007)
5. Sieg, A., Mobasher, B., Burke, R.: Web search personalization with ontological user profiles. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007), pp. 525–534 (2007)
6. Tao, X., Li, Y., Zhong, N.: A personalized ontology model for web information gathering. *IEEE Transactions on Knowledge and Data Engineering* 23(4), 496–511 (2011)
7. Zhong, N., Li, Y., Wu, S.: Effective pattern discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering* 24(1), 30–44 (2012)
8. Li, Y., Zhong, N.: Mining ontology for automatically acquiring web user information needs. *IEEE Transactions on Knowledge and Data Engineering* 18(4), 554–568 (2006)
9. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems* 1(3), 219–234 (2003)
10. Wu, S., Li, Y., Xu, Y., Pham, B., Chen, P.: Automatic pattern-taxonomy extraction for web mining. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), pp. 242–248 (2004)
11. Robertson, S., Soboroff, I.: The trec 2002 filtering track report. In: Proceedings of the 10th Text REtrieval Conference (TREC 2001), pp. 26–37 (2002)
12. Soboroff, I., Robertson, S.: Building a filtering test collection for trec 2002. In: Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), pp. 243–250 (2003)
13. Buckley, C., Voorhees, E.: Evaluating evaluation measure stability. In: Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pp. 33–40 (2000)