A Topic based Document Relevance Ranking Model

Yang Gao, Yue Xu, Yuefeng Li
Faculty of Science and Engineering
Queensland University of Technology, QLD, Australia
y10.gao@student.qut.edu.au, yue.xu@qut.edu.au, y2.li@qut.edu.au

ABSTRACT

Topic modelling has been widely used in the fields of information retrieval, text mining, machine learning, etc. In this paper, we propose a novel model, Pattern Enhanced Topic Model (PETM), which makes improvements to topic modelling by semantically representing topics with discriminative patterns, and also makes innovative contributions to information filtering by utilising the proposed PETM to determine document relevance based on topics distribution and maximum matched patterns proposed in this paper. Extensive experiments are conducted to evaluate the effectiveness of PETM by using the TREC data collection Reuters Corpus Volume 1. The results show that the proposed model significantly outperforms both state-of-the-art term-based models and pattern-based models.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Topic models, pattern mining, relevance ranking

1. INTRODUCTION

Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent users' interest. Traditional IF models were developed based on a term-based approach (i.e., BM25) or pattern mining techniques (i.e., closed pattern), which have achieved good performance. But all these data mining and text mining techniques hold the assumption that user's interest is only related to a single topic. However, the reality is not necessarily the case.

Topic modelling, such as LDA [1], has become one of the most popular probabilistic text modelling techniques and quickly been accepted by machine leaning and text mining communities. It is reasonable to expect that applying LDA to IF could make a breakthrough for current IF models due to two advantages of LDA: first, the topic based representation generated by using LDA conquers the problem of semantic confusion compared with the traditional term based document representation. Second, LDA can describe documents at a general level with multiple topics instead of a single topic in traditional IF.

Considering the benefits from data mining, a pattern mining based LDA method, called two-stage LDA model, has been proposed in [2] which alleviates the problem of semantic ambiguous topics in LDA by providing a promising way to meaningfully represent topics by patterns rather than single words. But the two-stage model can't represent documents with the discovered patterns. In this paper, we propose a new document relevance ranking model, called Pattern Enhanced Topic Model (PETM), that determines the document relevance based on the topic distribution and most discriminative patterns.

2. PATTERN ENHANCED TOPIC MODELLING

Latent Dirichlet Allocation (LDA) [1] is a typical statistical topic modelling technique and the most common topic modelling tool currently in use. It can discover the hidden topics in collections of documents with the appearing words. The resulting representations of LDA are at two levels, words distribution over topics at collection level and topics distribution at document level. In our proposed PETM, pattern mining is userd to discover semantically meaningful and efficient patterns to represent topics and documents, two steps are proposed: firstly, construct a new transactional dataset from the LDA outcomes of the document collection D; secondly, generate pattern based representations from the transactional dataset to represent user needs of D.

(1) **Construct Transactional Dataset**

Let R_{d_i,Z_j} represent the word-topic assignment to topic Z_j in document d_i. R_{d_i,Z_j} is a sequence of words assigned to topic Z_j. Construct a set of words from each word-topic assignment R_{d_i,Z_j}. Let I_{ij} be a set of words which occur in R_{d_i,Z_j}, $I_{ij} = \{w | w \in R_{d_i,Z_j}\}$, i.e., I_{ij} contains the words which are in document d_i and assigned to topic Z_j by LDA. I_{ij}, called a *topical document transaction*, is a set of words without any duplicates. From all the word-topic assignments R_{d_i,Z_j} to Z_j, we can construct a transactional dataset Γ_j. Let $D = \{d_1, \cdots, d_M\}$ be the original document collection, the transactional dataset Γ_j for topic Z_j is defined as $\Gamma_j = \{I_{1j}, I_{2j}, \cdots, I_{Mj}\}$. For the topics in D, we can construct V transactional datasets.

(2) **Generate Pattern Enhanced Representation**

The basic idea of the proposed pattern based method is to use patterns generated from each transactional dataset Γ_j to represent Z_j. In the two-stage topic model [2], frequent pat-

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577289.

terns are generated in this step. For a given minimal support threshold σ, an itemset X in Γ_j is frequent if $supp(X) >= \sigma$, where $supp(X)$ is the support of X which is the number of transactions in Γ_j that contain X.

3. AN IF MODEL BASED ON PETM

Representations generated by PETM can contain huge number of frequent patterns and can be difficult to accurately represent topics. A closed pattern covers all information that its subsets describe and can greatly reduce the number of frequent patterns. A generator is one subset of the closed pattern which has the same support as the closed pattern.

Definition 1.*Equivalence Class*: for a transactional dataset Γ, let X be a closed itemset and $G(X)$ consists of all generators of X, the equivalence class of X in Γ, denoted as $EC(X)$, is defined as $EC(X) = G(X) \cup \{X\}$.

Definition 2. *Maximum Matched Pattern*: Let $EC_{j1}, \cdots, EC_{jn_j}$ be the pattern equivalence classes of Z_j, a pattern in d is considered a maximum matched pattern to equivalence class EC_{jk}, denoted as MC^d_{jk}, if (1) $MC^d_{jk} \subseteq d$ and $MC^d_{jk} \in EC_{jk}$; (2) $\nexists\, X$ such that $X \in EC_{jk}, X \subseteq d$ and $MC^d_{jk} \subset X$; (3) $|MC^d_{jk}| > 1$.

Our novel IF model, which is based on PETM, consists of two parts, training part to generate user interests from a collection of training documents (i.e., document modelling) and filtering part to determine the relevance of incoming documents based on the user information interests generated in training part (i.e., document ranking).

(1)**Topic based User Interest Models**

User interest model $U = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$, can be generated by PETM, in which $\mathbf{X}_{Z_i} = \{X_{i1}, X_{i2}, \cdots, X_{im_i}\}$ is a set of frequent patterns generated for topic Z_i. Moreover, assume that there are n_i closed patterns in $\mathbf{X}_{Z_i}$, $c_{i1}, \cdots, c_{in_i}$, $\mathbf{X}_{Z_i}$ can be partitioned into n_i equivalence classes, $EC(c_{i1}), \cdots, EC(c_{in_i})$, their corresponding statistical significance are $f_{i1}, \cdots, f_{in_i}$, respectively. For simplicity, the equivalence classes are denoted as $EC_{i1}, \cdots, EC_{in_i}$ for $\mathbf{X}_{Z_i}$, or simply for topic Z_i.

Let $\mathbb{E}(Z_i)$ denote the set of equivalence classes of $\mathbf{X}_{Z_i}$ for topic Z_i, i.e., $\mathbb{E}(Z_i) = \{EC_{i1}, \cdots, EC_{in_i}\}$. In this paper, the equivalence classes $\mathbb{E}(Z_i)$ are used to represent user interests, i.e., $\mathbb{U} = \{\mathbb{E}(Z_1), \cdots, \mathbb{E}(Z_V)\}$.

(2)**Topic based Relevance Ranking**

For an incoming document d, we propose to estimate the relevance of d ($rank(d)$) to the user's interest based on topics distribution $\vartheta_{D,j}$ which represents the proportion of topic j in the collection D and topic significance $f_{j,k}$ which represents the user's interest to the topic. According to the user interest model, all the patterns in one equivalence class have the same frequency which indicates the statistical significance. The difference among them is the size. A longer pattern is more significant and specific than a shorter pattern in the same equivalence class. Based on this consideration, in this paper, maximum matched pattern is the most significant feature among all frequent patterns, which can estimate the relevance of d to the user interest. The document relevance is estimated using the following equation:

$$rank(d) = \sum_{j=1}^{V}\sum_{k=1}^{n_j} |MC^d_{jk}|^m \times f_{jk} \times \vartheta_{D,j} \qquad (1)$$

Table 1: Comparison of all models over all assessing collections of RCV1

Methods	*top20*	*b/p*	*MAP*	F_1
PETM	**0.529**	**0.446**	**0.463**	**0.450**
PTM	0.406	0.353	0.364	0.390
FCP	0.428	0.346	0.361	0.385
BM25	0.434	0.339	0.401	0.410
improvement%	21.9	28.9	15.5	9.8

where m is the scale of pattern specificity, we set $m = 0.5$. The higher the $rank(d)$ is, the more likely the document is relevant to the user's interest.

4. EVALUATION

The main hypothesis proposed in this paper is that user information needs involve multiple topics, document modelling by taking multiple topics into consideration can generate more accurate user information needs.

In RCV1, the first 50 collections are used in the experiments. For each collection, documents in RCV1 are divided into a training set and a testing set. The effectiveness is assessed by five different measures: average precision of the top K ($K = 20$) documents, $F_\beta(\beta = 1)$ measure, Mean Average Precision (MAP), break-even point (b/p) and $F_1 = \frac{2pr}{p+r}$. The larger the measure score is, the better the system performs. The experiments tested cross the 50 collections of independent datasets, which satisfy the generalized cross-validation for statistical estimation model.

The proposed PETM IF model is compared with three state-of-the-art models, which are frequent closed pattern model (FCP), sequential closed pattern model (conducted by PTM model [3]) and term based model, BM25.

The results given in Table 1 indicate the proposed PETM model significantly outperforms all the other baseline models and the improvements are consistent on all four measures.

5. CONCLUSION

We conclude that the PETM is an exciting achievement in discovering high-quality features in text documents mainly because it represents the text documents not only using the topics distribution at general level but also using most representative patterns, which are particularly maximum matched patterns, at detailed specific level, both of which contribute to the accurate document relevance ranking.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] Y. Gao, Y. Xu, Y. Li, and B. Liu. A two-stage approach for generating topic models. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PADKDD'13*, pages 221–232. Springer, 2013.

[3] N. Zhong, Y. Li, and S.-T. Wu. Effective pattern discovery for text mining. *Knowledge and Data Engineering, IEEE Transactions on*, 24(1):30–44, 2012.