IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VI (May-Jun. 2014), PP 17-22
www.iosrjournals.org

Improved Method for Pattern Discovery in Text Mining

Rakhi A. Dixit,[1] Prof. V.A.Chakrawar,[2]
PG Student, CSE Department GECA Aurangabad, India
Associate Professor CSE Department GECA, Aurangabad, India.

Abstract: *The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models. PBTM firstly generates pattern based topic representations to model user's information interests with multiple topics; then PBTM selects quality patterns for estimating the relevance of documents. The proposed approach incorporates the semantic topics from topic modeling and the specificity of the representative patterns. The proposed model has been evaluated by using RCV1 and TREC topics for the task of information filtering. Comparing with the state-of-the-art models, PBTM demonstrates excellent strength on document modeling and relevance ranking.*

Keywords: *Pattern mining, Text mining, Text classification*

I. Introduction

In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. In this paper, we focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining. Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models, rough set models, BM25 and support vector machine (SVM) based filtering models. The advantages of term based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term based methods suffer from the problems of ploys my and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

II. Related Work

The choice of a representation depended on what one regards as the meaningful units of text and the meaningful natural language rules for the combination of these units With respect to the representation of the content of documents, some research works have used phrases rather than individual words. In the combination of unigram and bigrams was chosen for document indexing in text categorization (TC) and evaluated on a variety of feature evaluation functions (FEF). A phrase-based text representation for Web document an augment was also proposed in data mining techniques have been used for text analysis by extracting co occurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency.

Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time, NLP was struggling for dealing with uncertainties in human languages. Recently, a new concept-based model, was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels. This model included three components. The first component analyzed the semantic structure of sentences; the second component constructed a conceptual onto logical graph (COG) to describe the sematic structures; and the last component extracted top concepts based on the first two components to build feature vectors using the standard vector space model. The advantage of the concept-based model is that it can effectively discriminate between non important terms and meaningful terms which describe a sentence meaning. Compared with the above methods, the concept-based model usually relies upon its employed NLP techniques.

III. Models

In the following, we introduce other two classes: the concept-based model and term-based methods.

3.3.1 Concept-Based Models

A new concept-based model was presented and which analyzed terms on both sentence and document levels. This model used a verb-argument structure which split a sentence into verbs and their arguments. For example, “John hits the ball,” where “hits” is a verb, and “John” or “the ball” are the arguments of “hits.” Arguments can be further assigned labels such as subjects or objects (or theme). Therefore, a term can be extended and to be either an argument or a verb, and a concept is a labeled term. For a document d, tf(c) is the number of occurrences of concept c in d; and ct f(c) is called the conceptual term frequency of concept c in a sentence s, which is the number of occurrences of concept c in the verb-argument structure of sentence s. Given a concept c, its tf and ctf can be normalized as tfweight(c) and ctfweight(c), and its weight can be evaluated as follows:

Weight(c)=tfreight(c)+ ctfweight(c):

To have a uniform representation, in this paper, we call a concept as a concept-pattern which is a set of terms. For example, verb “hits” is denoted as fhitsg and its argument “the ball” is denoted as “the, ball”

It is complicated to construct a COG. Also, up to now, we have not found any work for constructing COG for describing semantic structures for a set of documents rather than for an individual document for information filtering. In order to give a comprehensive evaluation for comparing the proposed model with the concept-based model, in this paper, we design n a concept-based model (CBM) for describing the features in a set of positive documents, which consists of two steps. The first step is to find all of the concepts in the positive documents of the training set, where verbs are extracted from Prop Bank data set. The second step is to use the deploying approach to evaluate the weights of terms based on their appearances in these discovery concepts. Unlike the proposed model, which uses 4,000 features at most, the concept-based model uses all features for each topic. Let CPi be the set of concepts in di 2 Dþ. To synthesize both tf and ctf of concepts in all positive documents, we use the following equation to evaluate term weights.We also designed another kind of the concept-based model, called CBM Pattern Matching, which evaluates a document d’s relevance by accumulating the weights of concepts that appear in d as follows:

$$weight(d) = \sum_{c \in d} weight(c).$$

3.3.2 Term-Based Methods

There are many classic term-based approaches. The Rocchio algorithm, which has been widely adopted in information retrieval, can build text representation of a training set using a Centroid =c as follows:

$$\vec{c} = \alpha \frac{1}{|D^+|} \sum_{\vec{d} \in D^+} \frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|D^-|} \sum_{\vec{d} \in D^-} \frac{\vec{d}}{\|\vec{d}\|},$$

where _ and _ are empirical parameters; D+and D- are the sets of positive and negative documents, respectively; d denotes a document.

Probabilistic methods (Prob) are well-known term-based approaches. The following is the best one:

$$W(t) = \log\left(\frac{r+0.5}{R-r+0.5} \Big/ \frac{n-r+0.5}{(N-n)-(R-r)+0.5}\right),$$

where N and R are the total number of documents and the number of positive documents in the training set, respectively; n is the number of documents which contain t; and r is the number of positive documents which contain t. In addition, TFIDF is also widely used. The term t can be weighted by W(t) = TF(d ,t) * IDF(t), where term frequency TF(d ,t) is the number of times that term t occurs in document d(d € D) (D is a set of documents in the data set); DF(t) is the document frequency which is the number of documents that contain term t; and IDF(t) is the inverse document frequency. Another well-known term-based model is the BM25approach, which is basically considered the state-of-the-art baseline. The weight of a term t can be estimated by using the following function.

$$W(t) = \frac{TF \cdot (k_1 + 1)}{k_1 \cdot ((1 - b) + b\frac{DL}{AVDL}) + TF} \cdot \log \frac{(r + 0.5)/(n - r + 0.5)}{(R - r + 0.5)/(N - n - R + r + 0.5)},$$

where TF is the term frequency; k1 and b are the parameters; DL and AV DL are the document length and average document length. The values of k1 and b are set as 1.2 and 0.75, respectively, according to the suggestion. The SVM model is also a well-known learning method introduced by Cortes and Vapnik [8]. Since the works of Joachims , researchers have successfully applied SVM to many related tasks and presented some convincing results. The decision function in SVM is defined as

$$h(x) = sign(W \cdot x + b) = \begin{cases} +1, & \text{if} \quad (W \cdot x + b) > 0, \\ -1, & \text{else,} \end{cases}$$

where x is the input space; b 2 R is a threshold and

$$W = \sum_{i=1}^{l} y_i \alpha_i x_i,$$

for the given training data

$$(x_i, y_i), \ldots, (x_l, y_l),$$

where xi € Rn and yi equals +1 (-1), if document xi is labeled positive (negative). α € R is the weight of the training example xi and satisfies the following on strains

$$\forall i : \alpha_i \geq 0, \quad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0.$$

Since all positive documents are treated equally before the process of document evaluation, the value of _i is set as 1.0 for all of the positive documents and thus the -value. In document evaluation, once the concept for a topic is obtained, the similarity between a test document and the concept is estimated using inner product. The relevance of a document d to a topic can be calculated by the function RðdÞ ¼ ~d _~c, where ~d is the term vector of d and ~c is the concept of the topic. For both term-based models and CBM, we use the following equation to assign weights for all incoming documents d based on their corresponding W functions

$$weight(d) = \sum_{t \in T} W(t)\tau(t, d).$$

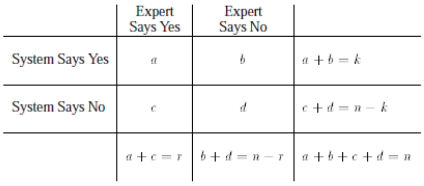
Hypotheses

The major objective of the experiments is to show how the proposed approach can help improving the effectiveness of pattern-based approaches. Hence, to give a comprehensive investigation for the proposed model, our experiments involve comparing the performance of different pattern-based models, concept-based models, and term based models.

In the experiments, the proposed model is evaluated in term of the following hypothesis: . Hypothesis H1. The proposed model, PTM (IPE), is designed to achieve the high performance for determining relevant information to answer what users want. The model would be better than other pattern based models, concept-based models, and state-of the art term-based models in the effectiveness. Hypothesis H2. The proposed deploying method has better performance for the interpretation of discovered patterns in text documents. This deploying approach is not only promising for pattern-based approaches, but also significant for the concept based model. In order to compare the proposed approach with others, the baseline models are grouped into three categories as mentioned the above. The first category contains all data mining-based (DM) methods, such as sequential pattern mining, sequential closed pattern mining, frequent item set mining, frequent closed item set mining, where min sup = 0.2. The second category includes the concept-based model that uses the deploying method and the CBM Pattern Matching model; and the last category includes n Gram, Rocchio, Probabilistic model, TFIDF, and two state-of-the art models, BM25 and SVM.

Measuring Effectiveness

The simplest and most common classification task is a binary one, where a system must decide whether or not an item belongs to a single class. Assume that a set of n documents has been classified by a binary text classification system and, separately, by an expert who judges the true classification. We can summarize the relationship between the system classifications and the expert judgments in a contingency table (Figure 1).



	Expert Says Yes	Expert Says No	
System Says Yes	a	b	$a+b=k$
System Says No	c	d	$c+d=n-k$
	$a+c=r$	$b+d=n-r$	$a+b+c+d=n$

Figure 1: A set of n classification decisions can be represented as a contingency table.

Each entry in the table specifies the number of documents with the specified outcome. For instance, is the number of times the system decided Yes, and Yes was in fact the correct answer. The effectiveness measures most widely used in IR are defined in terms of the contingency table: _ recall =a(+c) precision _ b fallout =(In words, recall is the proportion of class members that the system assigns to the class, and precision is the proportion of documents assigned to the class by the system that really are class members.

An ideal system would have recall and precision of 1. Fallout is the proportion of non class members that the system assigns to the class, and is an alternative to precision. An ideal system would have fallout of 0. Perfect recall can be achieved by a system that puts every document in the class, while perfect precision and fallout can be achieved by a system that puts no documents in the class, so using just one of these measures does not provide an interesting evaluation. One solution is to consider recall and precision, or recall and fallout, together and look at how the quantities trade off against one another under different parameter settings for the system. This is the usual approach in evaluating ranking systems, which do not have to commit to classification decisions.

Another approach is to define a single effectiveness measure

- error rate $= (b+c)/(a+b+c+d)$

that takes into account both errors of commission (b) and errors of omission (c). One such measure is error rate: _error rate A wide range of effectivenes measures have been defined. As early as 1973, Cooper commented that "too many ingenious and superficially plausible measures have been invented already".

Our goal in this paper will be not to propose new measures, or even to recommend one measure over another. Instead we will consider several families of single effectiveness measures, examining how the performance of systems on each can be estimated and optimized.

Information Filtering

Information Filtering (IF) is a system to remove irrelevant or unwanted information from an information or document stream based on document representations which represent users' interest.

Pattern-based Topic Model (PBTM)

 Benefits of using pattern mining

• Discover hidden associations among words to represent the documents and the collection • Patterns carry more semantic meaning than single words.

Combines pattern mining with statistical topic modeling to generate more discriminative and semantic rich topic representations

Latent Dirichlet Allocation

The statistical topic modeling technique has attracted great attention due to its robust and interpretable topic representations.

1. The most popular used topic modeling method is LDA (Latent Dirichlet Allocation), and its various extensions.
2. Each document is a mixture of topics
3. Each topic is represented by distributions of words

Corpus Level: A collection of documents is represented by a number of topics. Each topic is represented by group of words with probabilities.

Document Level: Every document in this collection is represented by a distribution of the topics, each topic is represented by group of words with probabilities.

Word Level: In each document, every word is assigned with a topic and a probability. This word-topic assignment indicates which words are important to which topics at document level.

PBTM for Information Filtering

Topic based User Interest Model:

Pattern Specificity: The specificity of a pattern X is defined as power function of the pattern length with the exponent less than 1, denoted as spe(X), spe(X) m =a|X|m , a and m are constant real numbers. In this paper, a = 1, m = 0.5.

Topic Significance:

Let d be a document, Zj be a topic in the user interest model, be matched patterns, k = 1,…, nj , to document d, and be the corresponding frequencies of the matched patterns within Zj, the topic significance Zj of to d is defined as:

$$sig(Z_j,d)=\sum_{k=1}^{n_j} spe(PA_{jk}^{d})\times f_{jk}=\sum_{k=1}^{n_j}| PA_{jk}^{d} |^{0.5}\times f_{jk}$$

Topic based Relevance Ranking:

The new idea of the proposed model is to use multiple topics to represent a collection, and represent each topic using semantic patterns. We choose two widely used patterns, frequent patterns and closed patterns to represent topics. The two models are PBTM_FCP and PBTM_FP.

Document ranking:

For an incoming document d, the relevance of d to the user interest model is estimated by topic significance and topic distribution:

$$rank(d)=\sum_{j=1}^{V} sig(Z_j,d)\times \vartheta_{D,j}$$

$$rank(d)=\sum_{j=1}^{V}\sum_{k=1}^{m_j}| X_{jk}^{d} |^{0.5}\times \vartheta_{D,j}$$

$$rank(d)=\sum_{j=1}^{V}\sum_{k=1}^{n_j}| c_{jk}^{d} |^{0.5}\times \vartheta_{D,j}$$

The experiment results clearly show that taking topics into consideration in generating user interest models can greatly improve the performance of information filtering.

- Pattern based topic models (PBTM) outperform word based topic model (LDA-word), which shows the benefit obtained by incorporating pattern mining into topic modeling, which is an important contribution of this paper.
- PBTM_FCP in most cases outperforms PBTM_FP, which indicates that using closed patterns to represent user interests is more accurate than using frequent patterns.
- The complexity of PBTM is determined by topic modeling or pattern mining, in most cases, by pattern mining. The complexity of efficient pattern mining methods such as FP-Tree has been proved acceptable in practice.

IV. Conclusion

In the future, we can select more discriminative and precise patterns for representing topics and document relevance These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective.

Acknowledgments

The authors would like to the editors and the anonymous reviewers for their insightful and helpful comments and suggestions, which resulted in substantial improvements to this work.

References

[1]. H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, “Text Classification Using String Kernels,” J. Machine Learning Research, vol. 2, pp. 419-444, 2002.

[2]. A. Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003.

[3]. C. Manning and H. Schu¨ tze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[4]. I. Moulinier, G. Raskinis, and J. Ganascia, “Text Categorization: A Symbolic Approach,” Proc. Fifth Ann. Symp. Document Analysis and Information Retrieval (SDAIR), pp. 87-99, 1996.

[5]. J.S. Park, M.S. Chen, and P.S. Yu, “An Effective Hash-Based Algorithm for Mining Association Rules,” Proc. ACM SIGMOD Int’l Conf. anagement of Data (SIGMOD ’95), pp. 175-186, 1995

[6]. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, “Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth,” Proc. 17th Int’l Conf. Data Eng. (ICDE ’01), pp. 215-224, 2001.

[7]. M.F. Porter, “An Algorithm for Suffix Stripping,” Program, vol. 14, no. 3, pp. 130-137, 1980.

[8]. S. Robertson and I. Soboroff, “The Trec 2002 Filtering Track Report,” TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER. FILTERING.ps.gz.

[9]. S.E. Robertson, S. Walker, and M. Hancock-Beaulieu, “Experimentation as a Way of Life: Okapi at Trec,” Information Processing and Management, vol. 36, no. 1, pp. 95-108, 2000.

[10]. J. Rocchio, Relevance Feedback in Information Retrieval. chapter 14,