Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on Improving Text Mining Using Discovery of Relevant Features by Natural Language Processing

Roshani S. Khule
CSIT, HVPM, COET, Amravati & SGBAU, Maharashtra, India

Ranjit R. Keole
HVPM, COET, Amravati & SGBAU, Maharashtra, India

Abstract— Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance. It is a big challenge to guarantee the quality of discovered relevance features in text documents for describing user preferences because of large scale terms and data patterns. Most existing popular text mining and classification methods have adopted term-based approaches. However, they have all suffered from the problems of polysemy and synonymy. Over the years, there has been often held the hypothesis that pattern-based methods should perform better than term-based ones in describing user preferences; yet, how to effectively use large scale patterns remains a hard problem in text mining.

Keywords: Text mining, text feature extraction, text classification.

I. INTRODUCTION

Text Mining [1] is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. Text mining is a variation on a field called data mining [2],that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information, yet, unstructured texts remain the largest readily available source of knowledge.

The objective of relevance feature discovery (RFD) is to find the useful features available in text documents, including both relevant and irrelevant ones, for describing text mining results. This is a particularly challenging task in modern information analysis, from both an empirical and theoretical perspective [3], [6]. This problem is also of central interest in many Web personalized applications, and has received attention from researchers in Data Mining, Machine Learning, Information Retrieval and Web Intelligence communities [2]. There are two challenging issues in using pattern mining techniques for finding relevance features in both relevant and irrelevant documents [3]. The first is the low-support problem. Given a topic, long patterns are usually more specific for the topic, but they usually appear in documents with low support or frequency. If the minimum support is decreased, a lot of noisy patterns can be discovered. The second issue is the misinterpretation problem, which means the measures (e.g., "support" and "confidence") used in pattern mining turn out to be not suitable in using patterns for solving problems. For example, a highly frequent pattern (normally a short pattern) may be a general pattern since it can be frequently used in both relevant and irrelevant documents.

Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used

depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system. Hence, the difficult problem is how to use discovered patterns to accurately weight useful features. There are several existing methods for solving the two challenging issues in text mining. Pattern taxonomy mining (PTM) models have been proposed [5], [6], [7], in which, mining closed sequential patterns in text paragraphs and deploying them over a term space to weight useful features.

Concept-based model (CBM) [5], [1] has also been proposed to discover concepts by using natural language processing (NLP) techniques. It proposed verb-argument structures to find concepts in sentences. These pattern (or concepts) based approaches have shown an important improvement in the effectiveness [7]. However, fewer significant improvements are made compared with the best term-based method because how to effectively integrate patterns in both relevant and irrelevant documents is still an open problem. Over the years, people have developed many mature term-based techniques for ranking documents, information filtering and text classification [7], [3], [4]. Recently, several hybrid approaches were proposed for text classification. To learn term features within only relevant documents and unlabelled documents, paper [2] used two term-based models. In the first stage, it utilized a Rocchio classifier to extract a set of reliable irrelevant documents from the unlabeled set. In the second stage, it built a SVM classifier to classify text documents. A two-stage model was also proposed in [3], [5], which proved that the integration of the rough analysis (a term-based model) and pattern taxonomy mining is the best way to design a two-stage model for information filtering systems. For many years, we have observed that many terms with larger weights are more general because they are likely to be frequently used in both relevant and irrelevant documents [2]. For example, word "LIB" may be more frequently used than word "JDK"; but "JDK" is more specific than "LIB" for describing "Java Programming Languages"; and "LIB" is more general than "JDK" because "LIB" is also C++. Therefore, we recommend the consideration of both terms' distributions and specificities for relevance feature discovery. Given a topic, a term's specificity describes the extent to which the term focuses on the topic that users want [3]. However, it is very difficult to measure the specificity of terms because a term's specificity depends on users' perspectives of their information needs [5]. We proposed the first definition of the specificity in [3], [1], which calculated the specificity score of a term based on its appearance in discovered positive and negative patterns. However, this definition required an iterative algorithm (three loops) in order to weight terms accurately. In order to make a breakthrough in relation to the two challenging issues, we proposed the first version of the RFD model In accordance with the distributions of terms in a training set, it provided a new definition for the specificity function and used two empirical parameters to group terms into three categories: "positive specific terms", "general terms", and "negative specific terms". Based on these definitions, the RFD model can accurately evaluate term weights according to both their specificity and their distributions in the higher level features, where the higher level features include both positive and negative patterns. The term classification method proposed in [3] requires manually setting two empirical parameters according to testing sets. In this paper, we continue to develop the RFD model, and experimentally prove that the proposed specificity function is reasonable and the term classification can be effectively approximated by a feature clustering method. We also design a comprehensive approach for evaluating the proposed models. In addition, we conducted some new experiments by using six new sliding windows to adaptively update the training sets and also applying the RFD model for binary text classification to test the robustness of the proposed model. This paper proposes an innovative technique for finding and classifying low-level terms based on both their appearances in the higher-level features (patterns) and their specificity in a training set. It also introduces a method to select irrelevant documents (so-called offenders) that are closed to the extracted features in the relevant documents in order to effectively revise term weights. Compared with other methods, the advantages of the proposed model include:

Effective use of both relevant and irrelevant feedback to find useful features; and Integration of both term and pattern features together rather than using them in two separated stages. To justify these claims for the proposed approach, we conducted substantial experiments on standard data collections, namely, the Reuters Corpus Volume 1 (RCV1), TREC filtering assessor topics, the Library of Congress Subject Headings (LCSH) ontology and Reuters-21578. We also used five measures and the t-test to evaluate these experiments. The results show that the proposed specificity function is adequate, the clustering method is effective and the proposed model is robust. The results also show that the proposed model significantly outperforms both the state-ofthe-art term-based methods underpinned by Okapi BM25,

II. SURVEY RELATED DETAILS

Text mining is the application of algorithm as well as methods from the machine learning and statistics to text with goal of finding useful pattern, Whereas data mining belongs in the corporate world because that's where most databases are, text mining promises to move machine learning technology out of the companies and into the home" as an increasingly necessary Internet adjunct (Witten & Frank, 2000) – i.e., as "web data mining" (Hearst, 1997). Laender, Ribeiro-Neto, da Silva, and Teixeira (2001) provide a current review of web data extraction tools. Text mining is also referred to as text data mining, roughly equivalent to text analytics, it refers to process of deriving high quality information form text. and high quality of information is derived through devising of patterns. Text analysis involves information retrieval, lexical analysis, word frequency distributions, pattern recognition, information extraction, and data mining techniques including link and association analysis, visualization to turn text into data for analysis via..natural language processing and analytical methods. On otherhand we called -Text mining is a variation on field called data mining, that tries to find interesting patterns from large datasets. Recently, one of the important issues for multimedia data is the identification of the optimal feature set without any

redundancy [6]; however, the challenging issue for text feature selection in text documents is the identification of which format or where the relevant features are in a text document because of the large amount of noisy information in the document[2]. Text features can be simple structures (words),LI ET AL.: RELEVANCE FEATURE DISCOVERY FOR TEXT MINING 1657complex linguistic structures or statistical structures. We mainly discuss three complex structures below for selecting relevant features: n-grams, concepts and patterns. n-grams (or phrases) are more discriminative and carry more "semantics" than words. They were useful for building good ranking functions [2], [4], [5]. In [49], a phrase based text representation for Web document management was also proposed that used rule-based Natural Language

Processing and Context Free Grammar techniques. Language models were proposed to calculate weights for n grams, which are often approximated by Unigram, Bigramor Trigram models for considering word dependencies [8], [3], [8]. A concept-based model [5] was also presented to find concepts in text documents by using NLP techniques, which analyzed terms' associations based on the semantic structure of sentences. This model included three components. The first one analyzed the semantic structure of sentences; the second one then constructed a conceptual ontological graph (COG) to represent the semantic structures; and the last one found top concepts according to the first two components to generate feature vectors by using the standard vector space model.

III. FUNDAMENTALS OF RFD MODEL

For a given topic, the goal of relevance feature discovery in text documents is to find a set of useful features, including patterns, terms and their weights, in a training set D, which consists of a set of relevant documents, Dþ, and a set of irrelevant documents, D_. In this paper, we assume that all text documents, d, are split into paragraphs, PSðdÞ. In this section, we introduce the basic definitions about patterns and the deploying method. These definitions can also be found in. In this section, we introduce the RFD model for relevance feature discovery, which describes the relevant features in relation to three groups: positive specific terms, general terms and negative specific terms based on their appearances in a training set. We first discuss the concept of "specificity" in terms of the relative "specificity" in training datasets and the absolute "specificity" in domain ontology .We also present a way to understand whether the proposed relative" specificity" is reasonable in term of the absolute "specificity". Finally, we introduce the term weighting method in the RFD model.

3.1 Specificity Function

In the RDF model, a term's specificity (referred to as relative specificity in this paper) is defined [32] according to its appearance in a given training set. Let T2 be a set of terms which are extracted from D_ and T ¼ T1 [T2. Given a term t 2 T, its coverageþ is the set of relevant documents that containt, and its coverage_ is the set of irrelevant documents that contain t. We assume that the terms frequently used in both relevant documents and irrelevant documents are general terms. Therefore, we want to classify the terms that are more frequently used in the relevant documents into the positive specific category; the terms that are more frequently used in the irrelevant documents are classified into the negative specific category. Based on the above analysis, we defined the specificity of a given term t in the training set D ¼ Dþ [D_ as follows:speðtÞ ¼ j coverage þðtÞj _ j coverage_ðtÞj n; (4)where coverageþðtÞ ¼ fd 2 Dþjt 2 dg, coverage_ðtÞ ¼ fd 2D_jt 2 dg, and n ¼ jDþj. speðtÞ > 0 means that term t isused more frequently in relevant documents than in irrelevant documents. Based on the spe function, we have the following classification rules for determining general terms G, positive specific terms Tþ and negative specific terms T_: G ¼ ft 2 Tju1 _ speðtÞ _ u2g, Tþ ¼ ft2 TjspeðtÞ > u2g, and T_ ¼ ft2 TjspeðtÞ < u1g, where u2 is an experimental coefficient ,the maximum boundary of the specificity for the general terms, and u1 is also an experimental coefficient, the minimum boundary of the specificity for the general terms.We assume that u2 > 0 and u2 _ u1. It is easy to verify that G \ Tþ \ T_ ¼ ;. Therefore, fG; Tþ; T_g is a partition
of all terms. A term's relative specificity describes the extent to which the term focuses on the topic that users want. It is very difficult to measure the relative specificity of terms because a term's specificity depends on users' perspectives of their information needs [55]. For example, "knowledge discovery" will be a general term in the data mining community; however, it may be a specific term when we talk about information technology. In this paper, we propose a way to understand whether the proposed relative "specificity" is reasonable in term ofthe absolute"specificity" in domain ontology, where "absolute" means the specificity is independent to any training dataset.Normally, people consider terms to be more general if they are frequently used in a very large domain ontology; otherwise, they are more specific. Therefore, we define the absolute specificity of a term in the ontology as follows: speontoðtÞ ¼ 1jcoverageðtÞj, where coverageðtÞ denotes the set of concepts of subjects that use term t for describing their meaning.To clearly illustrate the spe values between 0 and 1, we normalize the above equation as follows: speontoðtÞ ¼ log10NjcoverageðtÞj_ _=log10NM_ _; (5)where N is the total number of subjects and M is the maximum of coverageðtÞ for all We call a relative spe function reasonable if the average absolute specificity of its positive specific terms (Tþ) is greater than the average absolute specificity of its general terms (G).

IV. PROPOSED WORK

In the current approach, researchers are using N-Gram based technique for detection of relevant features in text mining. This approach is well known for accuracy if the related words in the vicinity of the current word are proper action words, and are relevant to the meaning of the sentence. In the research done, the researchers have used various techniques of N Gram like Bi-gram and tri-gram to incorporate the detection of relevance features in the given text, but these techniques if used alone give in-accurate results due to the lack of pre-processing done on the text to find the input keywords Our approach uses an improved and efficient Natural Language Processor based on the Word Net API, which performs pre-processing to give better and improved results for relevant feature detection and its application to text mining. This approach will first apply NLP on the input text to get only the action words, these action words will be found out on a per sentence basis, and then application of these words to the N-Gram Approach will help us to get the proper mining features and improve the accuracy of the mining in the system.

Proposed approach works in the following manner,

STEP 1: Collection of input text documents

- Some Text Datasets

-Popular Text Data Sets in Matlab Format

STEP 2: Application of Natural Language Processing on documents
-Natural Language Processing (Almost) from Scratch
-The Stanford Core NLP Natural Language Processing Toolkit
-Recursive Deep Learning for Natural Language Processing

STEP 3: Development of text mining for relevance feature discovery
-Relevance Feature Discovery for Text Analysis
-using data mining methods knowledge discovery for text ...

STEP 4: Integration of Text Mining with NLP for improved mining
-This is combination of 2 and 3

STEP 5: Result analysis and comparison
-We will be finding results in this module

V. AIMS AND OBJECTIVES

Improving text mining using discovery relevant features by natural language processing with the goal of achieving following objectives:

- Improved accuracy of text mining
- Getting better recommendations of the text mining

VI. CONCLUSION

The first RFD model uses two empirical parameters to set the boundary between the categories. It achieves the expected performance, but it requires the manually testing of a large number of different values of parameters. The new model uses a feature clustering technique to automatically group terms into the three categories. Compared with the first model, the new model is much more efficient and achieved the satisfactory performance as well. This paper also includes a set of experiments on RCV1 (TREC topics), Reuters-21578 and LCSH ontology. These experiments illustrate that the proposed model achieves the best performance for comparing with term-based baseline models and pattern-based baseline models. The results also show that the term classification can be effectively approximated by the proposed feature clustering method, the proposed spe function is reasonable and the proposed models are robust. The main objective of the review paper was to throw some light on the previous proposed work. It provides a promising methodology for developing effective text mining models for relevance feature discovery.

REFERENCES

[1] Yuefng Li, Abdulmohsen Algarni,Mubarak Albathan, Yan shen, and moch Arif Bijaksana"Relevance feature discovery for text mining" IEEE transaction on knowledge and data engineering,vol.27,no.6,june2015

[2] Ning Zhong, Member of the IEEE, Yuefeng Li, Member, IEEE, and Sheng Tang Wu, Member, IEEE " Effective Pattern Discovery for Text Mining"IEEE, transaction on knowledge and data engineering,vol.24,no. 1, January 2012

[3] D.M.Kulkarni and S.K.Shirgave"using data mining methods knowledge discovery for text mining" QUT E-Discovery lab,

[4] V.Sharmila,I.Vasudevan, Dr.g.Tholkappia Arasu "Pattern based classification for text mining using fuzzy similarity algorithm" Journal of theoretical and applied information technology,vol 63,May 2014.

[5] Mallareddy kiran, R.Ravikanth ME "Classification of documents using effective pattern taxonomy"International journal of computer applications,vol 86, no. 6, January 2014.

[6] Miss. Dipti charjan and prof. Mukesh Pund "Pattern discovery for text mining using pattern taxonomy"International journal of engineering trends and technology,vol 4,issue 10,October 2013

[7] Yuefeng Li,Xiaohui Tao, Abdulmohsen Algarni,Sheng Tang Wu " Mining specific and general features in both positive and negative relevance feedback" Australian research council, July 2010

[8] Vishal Gupta and Gurpreet S. Lehal*"* A survey of text mining techniques and application" Journal of emerging technologies in web intelligence, vol 1, no. 1, August 2009

