MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

A SURVEY PAPER FOR TEXT MINING OF IMPORTANT TERM FROM RELEVANCE DOCUMENT USING PATTERN BASED MODEL

Vijaykumar Ganpatrao Ingawale, Prof. Sunil Damodar Rathod
Computer Department
Dr. D. Y. Patil School of Engineering, (Affiliated to Savitribai Phule Pune University)
Pune, India

Abstract: *Nowadays, the big challenge for community of Information Retrieval domain is to discover the relevance feature in text document which helps to decide whether document is relevant or irrelevant. Most existing text mining methods are based on term-based approaches which extract terms from a training set for describing relevant feature. In term-based approach multiple meaning of same word in different context leads to generate polysemy and synonymy issue. The term based approach also suffers from low level support problem. Even though pattern based text mining approach solve low level support problem but still this approach suffers from large number of noise pattern. In the propose work, a pattern discovery approach for text mining is explored. This approach discovers frequent sequential pattern and closed sequential patterns in text documents for identifying the most information contents of the documents and extract useful features for text mining. It also classifies extracted terms into three categories: positive terms, general terms, and negative terms. In this way, we update extracted features using multiple revising strategies. This technique discovers positive and negative patterns in text documents as higher level features in order to accurately weight low-level terms based on their specificity.*

Keywords: *Data Mining, Information Retrieval, text Classification, Feature Selection.*

1. INTRODUCTION

Web search engines return arrangements of website pages sorted by the page's relevance to the user query. The issue with web search relevance ranking is to establish relevance of a page to a query [12]. These days, business web-page search engines combine hundreds of features to approximate relevance [13].Information Retrieval (IR) Systems are the associates of Web and search engines. These systems are designed to retrieve documents from digital collections e.g. library abstracts, corporate reports, news and so forth. Generally, IR relevance ranking algorithms are designed to obtain high recall on medium sized document collections using detailed user queries. Moreover, textual documents in these collections had practically no structure or hyperlinks [12]. A web search engine uses many methods of the

standards and calculations of Information Retrieval Systems, however needed to adjust and stretch out them to fit their needs.

Data mining techniques help user to find useful information from a large amount of text documents on the Web. Many text mining methods have been developed in order to achieve the goal of retrieving useful information for users [12]. Most of them adopt the term-based approach whereas the others choose the pattern-based technique to construct a text representative for a set of documents.IR has provided many effective term-based methods to solve this challenge [17]. The advantages of term-based methods include efficient computational performance; as well as mature theories for term weighting.

In the recent work, various data mining techniques have been proposed for feature (e.g. term, pattern) discovery. These tasks include sequential pattern mining, frequent pattern mining and closed pattern mining. The synonymy and polysemy are the main issues associated with term-based methods [3], [9], and [11]. Polysemy implies same word has numerous meaning while synonymy implies a different word has the same meaning [3]. Also pattern-based methods face low frequency and miss understanding problems [3]. A highly relevant pattern is usually a specific pattern of low frequency. Many noisy patterns are discovered, if we reduce the minimum support. The measures used in pattern mining (support and confidence) turn out to be not suitable to discover useful patterns which lead to miss understanding. In text document, the difficult problem is how to use discovered patterns to exactly evaluate the weights of useful feature [3], [12].

2. LITERATURE REVIEW

These days web assets and its use is continuously increasing much over the time. User needs valuable data rapidly, while utilizing web. There are a large number of new documents in web and user want efficient results while searching the web. There are some issues in Web search [12], such as effective ranking and relevance, evaluation and information needs. The IR community faces the challenge of managing a huge amount of hyperlinked data, but members of this community can utilize modeling, document classification and categorization, user interfaces, and data visualization altering to accomplish their goals [12] [13]. Information Retrieval models are based on ranking algorithm, which is used in search engines to produce the ranked list of documents [6]. A ranking algorithm sorts a set of documents according to their relevance to a give query [8].

Feature selection is the process of selecting a subset of relevant features for use in model construction. In text documents feature can be term, pattern, sentence. However, the traditional feature selection methods are not effective for selecting text features for solving the relevance issue because relevance is a single class problem [13]. The well-organized way of feature selection for relevance is based on a feature weighting function. A feature weighting function indicates the amount of information represented by the feature occurrences in a document and reflects the relevance of the feature. The term-based IR models include the Rocchio algorithm [13], [19], Probabilistic models and Okapi BM25 [19] and language models, including model-based methods and relevance models [12], [13]. In a language model, the key fundamentals are the probabilities of word sequences which include both words and sentences. They are often approximated by n-gram models [13],

such as Unigram, Bigram or Trigram, for considering term dependencies. In the recent work important issue for feature selection in a text document is to identify format of the document. Text feature can be a single word or complex structure. It comprises various complex structures such as n-grams, pattern and term. A frame is a contiguous sequence of *n* items from a given sequence of text. The items can be phonemes, syllables, letters, words or base pairs according to the application. Pattern mining has been extensively studied in data mining communities for many years. A variety of efficient algorithms such as Apriori-like algorithms, PrefixSpan, FP-tree, SPADE, SLPMiner and GST [4], [5],[6],[7], [8] have been proposed. Patterns can be discovered by data mining techniques like frequent item set mining, sequential pattern mining and closed pattern mining [2]. To conquer the drawbacks of sequential patterns and closed patterns, taxonomy models have been developed in pattern discovery technique [18].

Feature classification is assigning different task according to predefine group of documents. There are numerous classification methods, for example, Naive Bayes, Rocchio, KNN and SVM have been produced in IR [14], [15], [16]. SVM is one of the main classification strategies used in machine learning domain [14]. The grouping issues incorporate the single and multi-marked issue. Term based model documents having semantic meaning and documents are analyzed on the basis of the term. The regular arrangement [13] to the numerous named issues is to break down it into a few classifiers, where a classifier allocates two predefined classifications. The two classifications are positive or negative classification. Term based methods suffer from the problems of polysemy and synonymy [10]. Polysemy implies a word has numerous meaning and synonymy implies different words having the same meaning. IR gave numerous term-based strategies to this test [2], [3]. The similar research was also published in [2], [11] for developing a new methodology of post-processing of pattern mining, pattern summarization, which grouped patterns into some clusters. Further patterns in the same clusters are into a master pattern that consists of a set of terms which are composed into a term-weight distribution. It is still a challenging issue for pattern-based methods to deal with low frequency patterns (noise).

In summary, the existing methods for finding relevance features are divided into three approaches. The first approach considers feature terms that appear in both positive samples and negative samples that are Rocchio-based models [19] and SVM [14]. The second approach is based on probabilistic based models [15] in which terms appear or do not appear in positive documents and negative documents which defines their importance. The third approach considers only positive patterns from the documents [11].

3. PROPOSED SYSTEM ARCHITECTURE

3.1 Term extraction

Most text analysis such as document classification include a step of text extraction to determine the words or terms that occur in each document. Text feature extraction depends on some definition of which characters are to be treated as word characters vs. non-word characters. Text feature extraction depends on some definition of which characters are to be treated as word characters vs. non-word characters.

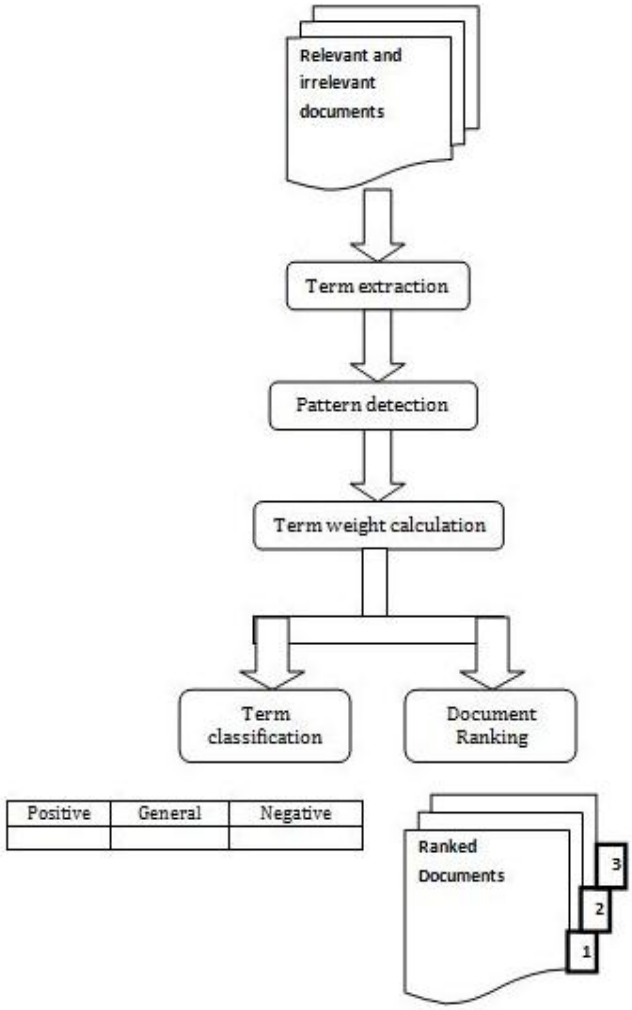

Fig-1. Proposed System Architecture

3.2 Pattern extraction

In pattern extraction documents are analyze on pattern basis. Patterns can be discovered by data mining techniques like frequent item set mining, sequential pattern mining and closed pattern mining [2]. The sequential pattern mining listed text documents in a certain order frequently in the same paragraph. A frequent closed pattern is a frequent sequential pattern such that it is not included in another sequential pattern having exactly the same support. The important thing to know is that closed sequential patterns are a compact representation of all sequential patterns .Patterns have possible for text mining since they have predictive power, and allow capturing semantic relationships existing among terms in sentences, paragraphs, or even the whole document [1],[2],[3]. The pattern based technique uses two processes, pattern deploying and pattern evolving [9].

3.3 Term weight calculation

The term weighting is one of the important works of information filtering and information retrieval. The term weighting function in text documents are used to find useful feature including pattern, terms and their weights occurs in text documents [10]. The weighting of term indicates how the term applies to topic by exploiting the statically variations in the

distribution of term with relevant documents. The frequency of a term in a document d can be used for document specific weighting. It is only a measure of a term with a document.

3.4 Document ranking

With the continuing growth of web information, it has become more and more important to provide improved mechanisms to find information rapidly. Usual Information Retrieval systems rank the documents based on maximizing relevance to the user query [12], [13]. A document ranking method is one where each document in the ranked according to query relevance and information

3.5 Term classification

Relevance Feature uses three specific features positive, negative and general features. Therefore, the key research question is how to find the best relevant classifier for positive documents and negative documents, for a given set of features. In this propose work, term classification method requires two empirical factors according to testing sets. In this section, propose work focus on approximation approach to find the relevance feature.

4. TAXONOMY CHART

The taxonomy chart given below shows the comparison of various existing system based on different approaches. The parameters are used here various functionalities of relevant text documents that are give clear idea about different approaches used in existing system.

Parameters / Systems	Feature Selection	Classification	Ranking Methods	Datasets
Comparison of Term frequency and document frequency	✓	X	X	✓
Fast Logistic Regression	✓	✓	X	✓
High-Precision Phrase-Based Document	X	✓	X	X
Query Dependent Ranking	X	✓	✓	✓
Mining Sequential Patterns	✓	✓	X	X
Topical Pattern Based Document	✓	✓	✓	✓

Fig-2. Taxonomy Chart

5. CONCLUSION

In this review of research work we have discussed different approaches for relevance feature discovery in text documents. Different data mining techniques have been proposed. Frequent item set mining, closed pattern mining, sequential pattern mining, and closed

pattern mining these all techniques are used in data mining techniques. The pattern deploying and pattern evolving techniques are used in proposed Method. In this survey work problems of low frequency and misinterpretation in pattern mining techniques are discussed. The paper also defines different approaches for relevance feature discovery.

ACKNOWLEGEMENT

We would like to thank MJRET for giving such wonderful platform for the PG students to publish their research work. Also would like to thanks to my guide & respected teachers for their constant support and motivation for us. We also extend our thanks to Dr. D.Y.PATIL SCHOOL OF ENGINEERING, CHARHOLI, PUNE for providing a strong platform to develop our skill and capabilities.

REFERENCES

[1] *Jaillet, S., Laurent, A., Teisseire, and M.: Sequential patterns for text categorization. Intelligent Data Analysis 10 (3), 199–214 (2006)*

[2] *Wu, S., Li, Y., Xu, Y., Pham, B., Chen, and P.: Automatic pattern-taxonomy extraction for web mining. In: 3th IEEE/WIC/ACM WI International Conf. In Web Intelligence, pp. 242–248 (2004)*

[3] *Zhong, N., Li, Y., Wu, S.: Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, doi: http://doi.ieeecomputersociety.org/10.1109/TKDE,2011*

[4] *D.B. Liu. Web data mining: exploring hyperlinks, contents, and usage data. Data-centric systems and applications. Springer, Berlin, 2007.*

[5] *A. Rakesh and R. Srikant. Mining sequential patterns. In proceedings of the 11th International Conference on Data Engineering, pages 3.14, 1995.*

[6] *X. Yan, J. Han, and R. Afshar. Clospan: Mining closed sequential patterns in large data sets. In Data Mining (SDM03), pages 166.177, 2003.*

[7] *J. Han and K. Chang. Data mining for web intelligence. IEEE Computer, 35 (11): 64:70, 2002.*

[8] *M. J. Zaki. Spade: an efficient algorithm for mining frequent sequences. In Machine Learning Journal, special issue on Unsupervised Learning, pages 31-60, 2001.*

[9] *S. -T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.*

[10] *G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management: An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.*

[11] *Y. Li, A. Agony, and N. Zhong. Mining positive and negative patterns for relevance feature discovery. In Proceedings of KDD'10 pages 753–762, 2010.*

[12] *C. C. Yang. Search engine information retrieval in practice. J.Am. Soc. Inf. SCI. Technol., 61:430–430, 2010.*

[13] *C. D. Manning, P. Raghavan, and H. Sects. Introduction to Information Retrieval. Cambridge University Press, 2009.*

[14] *D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res., 5:361–397, December 2004.*

[15] *X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In Proceedings of IJCAI'03, pages 587–592, 2003.*

[16] *X. -L. Li, B. Liu, and S. -K. Ng. Learning to classify documents with only a small positive training set. In Proceedings of ECML'07, pages 201–213, Berlin, Heidelberg, 2007.*

[17] *S. E. Robertson and I. Soboroff. The trek 2002 filtering track report. In Proceedings of TREC'02, 2002.*

[18] *Y. Li, X. Zhou, P. Bruce, Y. Xu, and R. Y. Lau. Two-stage Decision Model for Information Filtering. Decision Support Systems, 52 (3): 706-716, 2012.*

[19] *T. Joachims. A probabilistic analysis of the rich algorithm with tfidf for text categorization. In Proc. On ICML'97, pages 143–151, 1997.*