Mining an Open Forum with Similar Words Scoring

Rajkumar K
Department of Computer Science and Engineering, Indian Institute of Information Technology, Srirangam
Tiruchirapalli, Tamilnadu, India.

G. AnnaPoorani
Assistant Professor, Department of Information Technology, Anna University Trichy – BIT Campus
Tiruchirapalli, Tamilnadu, India.

Abstract – Many data mining techniques have been proposed so far to mine text documents. However, those are all not giving importance to similarity and relatedness between words. Since most existing text mining methods adopted term-based approaches, they all suffer from the problem of synonymy between words. Synonym is a word or phrase that means exactly or nearly the same as another word or phrase in the same language. This paper presents an innovative and effective text mining technique which includes the processes of detecting and scoring of similar and closely related words, to improve the efficiency of the text mining. Substantial experiments on an open forum demonstrate that the proposed solution achieves encouraging performance. Here similar words refers to the words which are similar by meaning and relatedness refers to the contextual relationship between words.

Index Terms – Text mining, Synonym based mining, phrase value, word similarity.

1. INTRODUCTION

Because of the quick development of digital information made accessible in late years, knowledge discovery and information mining have pulled in a lot of consideration with an imminent requirement for transforming such information into helpful data and knowledge. Numerous applications, for example, market examination and business administration, can benefit by the utilization of the data and learning removed from a substantial sum of information. Information revelation can be seen as the procedure of nontrivial extraction of data from huge databases, data that is certainly introduced in the information, already obscure and possibly valuable for clients.

Information mining is in this way a crucial stride during the time spent information revelation in databases. In the previous decade [5], a noteworthy number of information mining methods have been introduced keeping in mind the end goal to perform distinctive information undertakings. These systems incorporate association tenet mining, successive item set mining, consecutive example mining, greatest example mining, and shut example mining. The vast majority of them are proposed for the reason of creating proficient mining calculations to discover specific designs inside of a sensible and adequate time period. With an expansive number of examples produced by utilizing information mining methodologies, how to viably utilize and overhaul these examples is still an open examination issue. In this paper, we focus on the development of a knowledge detection model to effectively use the discovered similarity and apply it to the fieldof text mining.

Text mining is the discovery of interesting knowledge in text document. It is a challenging issue to find accurate knowledge (or features) in content reports to push clients to find what they need. At the outset, Information Retrieval (IR) if numerous term-based routines to settle this test, for example, Rocchio and probabilistic models [4], rough set models [23], BM25 and support vector machine (SVM) [34] based sifting models. The upsides of term-based techniques incorporate effective computational performance and in addition adult speculations for term weighting, which have developed in the course of the last couple of decades from the IR and machine learning groups. On the other hand, term-based systems experience the ill effects of the issues of polysemy and synonymy, where polysemy implies a word has different implications, and synonymy is different words having the same importance. The semantic importance of numerous found terms is indeterminate for noting what clients need.

Throughout the years, individuals have regularly held the theory that expression based methodologies could perform superior to anything the term-based ones, as expressions may convey more "semantics" like data. This theory has not fared too well in the history of IR [19], [20], [21]. In spite of the fact that expressions are less uncertain and more discriminative than individual terms, the conceivable purposes behind the demoralizing execution include: 1) expressions have sub-par factual properties to terms, 2) they have low recurrence of event, and 3) there are huge quantities of excess and uproarious expressions among them [21].

In the vicinity of these setbacks, sequential similarity used in data mining community have turned out to be a promising alternative to phrases [13], [15] because sequential similarity enjoy good statistical properties like terms. To overcome the disadvantages of phrase-based approaches, word similarity based text mining-based approaches (or similarity taxonomy models (STM) [15], have been proposed, which embraced the

idea of shut consecutive examples, and pruned no closed designs. These word similarity based text mining based approaches have demonstrated certain degree enhancements on the adequacy. Then again, the Catch is that individuals think similarity based methodologies could be a noteworthy option, be that as it may, therefore less huge changes are made for the adequacy contrasted and term-based techniques.

There are two crucial issues in regards to the effectiveness of example based methodologies: low recurrence and distortion. Given a predetermined point, an exceedingly visit example (ordinarily a short example with expansive backing) is typically a general example, or a particular example of low recurrence [18]. On the off chance that we diminish the base bolster, a great deal of uproarious examples would be found. Distortion implies the measures utilized as a part of example mining (e.g., "support" what's more, "certainty") end up being not suitable in utilizing found examples to answer what clients need. The troublesome issue consequently is the manner by which to utilize found examples to precisely assess the weights of helpful elements (information) in content reports.

In order to solve the above paradox, this paper presents an effective word similarity based text mining technique, which first calculates discovered specificities of similarity and then evaluates term weights according to the distribution of terms in the discovered similarity rather than the distribution in documents for solving the misinterpretation problem [10]. It also considers the influence of similarity from the negative training examples to find ambiguous (noisy) similarity and try to reduce their influence for the low-frequency problem. The process of updating ambiguous similarity can be referred as similarity evolution [13]. The proposed approach can improve the accuracy of evaluating term weights because discovered similarity are more specific than whole documents.

The rest of this paper is structured as follows: Section 2 discusses existing system. Section 3 proposes proposed system. Sections 4 provides the results. Finally, Section 5 gives conclusions.

2. EXISTING SYSTEMS

Many types of text representations have been proposed in the past. A well-known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In [21], the tf*idf weighting scheme is used for text representation in Rocchio classifiers. In addition to TFIDF, the global IDF and entropy weighting scheme is proposed in [9] and improves performance by an average of 30 percent. Various weighting schemes for the bag of words representation approach were given in [1], [14], [18]. The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid overfitting [14]. In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on. Details of these selection functions were stated in [19], [21].

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 [4]. With respect to the representation of the content of documents, some research works have used phrases rather than individual words. In [7], the combination of unigram and bigrams was chosen for document indexing in text categor- ization (TC) and evaluated on a variety of feature evaluation functions (FEF). A phrase based text representation for Web document management was also proposed in [14].

In [3], data mining techniques have been used for text analysis by extracting cooccurring terms as descriptive phrases from document collections. However, the effective- ness 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 for terms" as mentioned in [18].

Term-based ontology mining methods also provided some thoughts for text representations. For example, hier- archical clustering [8], [9] was used to determine synony- my and hyponymy relations between keywords. Also, the pattern evolution technique was introduced in [25] in order to improve the performance of term-based ontology mining.

Pattern mining has been extensively studied in data mining communities for many years. A variety of efficient algorithms such as Apriori-like algorithms [2], [3], [9], PrefixSpan [12], FP-tree [11], SPADE [16], SLPMiner [4], and GST [12] have been proposed. These research works have mainly focused on developing efficient mining algo- rithms for discovering patterns from a large data collection. However, searching for useful and interesting patterns and rules was still an open problem [2], [12], [20]. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent itemsets, cooccurring terms and multiple grams, for building up a representation with these new types of features. Nevertheless, the challenging issue is how to effectively deal with the large amount of discovered patterns.

For the challenging issue, closed sequential patterns have been used for text mining in [15], which proposed that the concept of closed patterns in text mining was useful and had the potential for improving the performance of text mining. Pattern taxonomy model was also developed in [15] and [15] to improve the effectiveness by effectively using closed patterns in text mining. In addition, a two-stage model that used both term-based methods and pattern- based methods was

introduced in [16] to significantly improve the performance of information filtering.

Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time [5], NLP was struggling for dealing with uncertainties in human languages. Recently, a new concept-based model [14], [16] 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 ontological 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 nonimportant 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.

3. PROPOSED SYSTEM

In this paper,we created a dictionary called as word similarity dictionary, which contain collection of similar words.

Example :

Stomach pain, stomach ache, abdomen pain

Let us consider the above example. Here the three words point to a same meaning of stomach ache. But people uses all three words in daily usage.

As already discussed when mining a public forum the main problem is this only, that is synonymy. The weightage is splitted while mining the public forum. To get perfect result we have to combine the result.

Our technique proposes an algorithm which combine the weightage of all the same meaning words to get a efficient result.

Documents used in Word Similarity Mining Algorithm :

1. A – Word similarity dictionary
2. B – Text to be mined
3. C – Result in word count
4. D – Word cloud of C

Word Similarity Mining Algorithm:

Main function:

```
Start
for each word 'w' in B :
    if( similar word already there in C)
        increase similar word of 'w'-'s count by 1
    else
        create new entry for 'w' in C with count 1
Draw word cloud for C
Stop
Similar Word Finder:
StringArray findSimilarWords(w)
    look for 'w' in A
    return line in which w located
```

Explanation:

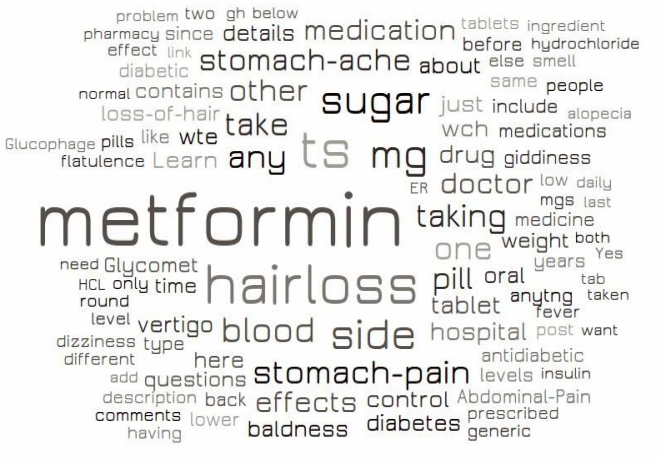
This algorithm read word by word in the text document to be mined. It checks whether the similar word is already there in the result ducument. If it is there means increase similar word of w's count by one.

Else create new entry for w in result with count by 1. The result is like word count. By the word count the word cloud is formed.

4. EXPERIMENTAL RESULTS

By Applying above algorithm in a medical open forum corpus the following results are obtained,

Without applying algorithm,



After applying the word similarity mining algorithm,

Result by count,
without applying algorithm,
stomach pain -421
stmach-ache-370
abdominal-pain - 178
baldness -217
hairloss - 737
hairfall - 89
loss-of-hair -181
hairlossing- 19
dizziness-99
giddiness-264
vertigo-320
after appling the word similarity mining algorithm,
stomach-ache- 969
hairloss- 1243
dizziness-683

From this result we can conclude that this algorithm is better than already existing algorithms.

5. CONCLUTION AND FUTURE WORKS

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent itemset 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. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance.

In this research work, an effective similar words mining technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The experimental results show that the proposed model outperforms on public open forums.

The future work is to implement the machine learning to find the similar words and same meaning words at run time, Instead of using word similarity dictionary.

REFERENCES

[1] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.
[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.
[3] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.
[4] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.
[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Kernel Methods for Document Filtering," TREC, trec.nist.gov/ pubs/trec11/papers/kermit.ps.gz, 2002.
[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word-Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.
[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07-2000, Instituto di Elaborazione dell'Informazione, 2000.
[8] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
[9] S.T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.
[10] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
[11] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.
[12] Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.
[13] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.
[14] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.
[15] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. European Conf. Machine Learning (ICML '98),, pp. 137-142, 1998.
[16] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 200-209, 1999.
[17] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
[18] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '92), pp. 37-50, 1992.
[19] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.
[20] D.D. Lewis, "Evaluating and Optimizing Automous Text Classification Systems," Proc. 18th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '95), pp. 246-254,1995.
[21] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI'03), pp. 587-594, 2003.