Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 9, September 2016, pg.112 – 116

THE COMPARISON OF TERM BASED METHODS USING TEXT MINING

S.Brindha[1], Dr. K.Prabha[2], Dr. S.Sukumaran[3]

[1](Ph.D Scholar, brindha.balajiee@gmail.com, Department of Computer Science Erode Arts and Science College, Erode, Tamilnadu, India)
[2](Assistant Professor of Computer Science, prabhaeac@gmail.com, Periyar University PG Extension Centre, Dharmapuri, Tamilnadu, India)
[3](Associate Professor, prof_sukumar@yahoo.co.in, Department of Computer Science, Erode Arts and Science College, Erode,Tamilnadu, India)

ABSTRACT: Increasing usability of internet and modern growth of information technology changing the fields of activity in modern days. Large number of people and organization would be required to interact more frequently with computer systems. The text documents are pre-processed; Term Frequency and Inverse Document Frequency (TF-IDF) are used to rank the document. In text mining the existing classification methods, the documents representing the traditional vector space model. A term graph model to represent not only the content of a document and also the relationship among keywords. Traditional text classification methods utilize term frequency (tf) and inverse document frequency (idf) as the main for information retrieval. High performance is applied in text classification using term weighting. TFIDF is one of the popular methods but it is not using the information retrieval. In this paper improved the supervised weighting in the TFIDF model and it also important to assume the low frequency terms. But using this TFIDF method, no need to concentrate the high frequency terms. So generate to consider the higher weights to the rare terms frequently. In this paper, compare a model for text document by combining the strengths of vector space model and frequently co-occurring terms together. Finally term graph model results are compared.

Keywords: TF-IDF, Concept Based Mining, Sentence Based Mining, Document Based Mining.

I. INTRODUCTION

In data mining text mining is one of the most important research areas in recent years. The rapid growth of evolution of technology the text documents usage can also increased. Such as, web pages, office documents and e-mails etc. Text mining can be automatically extracting information from different textual resources. It is multidisciplinary field, involving information retrieval, text analysis, and information extraction, clustering visualization, database technology, machine learning and data mining. The important aim of the text mining is to improve the textual database. All the paper publications and higher number of possible words and phrase types in the language, subtle and complex relationships between concepts in text. Information extraction is the task of automatically extracting structured information from unstructured and semi structured machine readable documents. Text mining is automatically extracting information from different textual resources. The goal of text mining is to discover previously unknown information. The challenges that arise due to unstructured text are large textual database. All publications are also in electronic form. Very high number of possible word and phrase types in the language, Complex and subtle relationships between concepts in text. A lot of research has been done to improve the quality of text representation and to develop high quality classifiers. Most of the machine learning methods as treats the text documents as bag of words. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. This activity concerns processing human language texts by means of natural language texts by means of natural language processing. The overall goal is to create a more easily machine readable text to process the sentences. Text classification is an important part of text mining. Current research
of text classification aims to improve the quality of text representation and develop high quality classifiers.

Basically Text Mining tools can be divided into two parts that is Text analysis tools, Web Searching Tools. The text analysis tools are divided into four the document can be extracted using feature extraction, categorization, summarization, clustering. The clustering can be identified Hierarchical and binary relational clustering. Web searching tools can be analysed using Text Search Engine, net Question solution, web Crawler. Text classification process includes collection of data documents, data pre-processing, indexing, term weighing methods, classification algorithms and measure. Machine learning techniques have been actively explored for text classification.

Text classification have many challenges for inductive learning methods since there can be millions of word features. The resulting classifiers have many advantages. The text can be easy to update and easy to construct and it depend only on information that is the people to provide and it can be customized to specific categories of individual interest. Using precision and recall can be used to analyze and to find the trade off values.

Text classification deals with the problem of automatically assigning single or multiple category (or class) labels to a new text document based after learning from a set of training documents with correct category labels. Most existing text classification methods (and text mining methods at large) adopt the approach of transforming the text mining problem into traditional machine learning problem. Usually, the conversion of a text document into a relational tuple is performed using the popular *vector-space model*. Intuitively, the document is parsed, cleaned and stemmed, in order to obtain a list of *terms* with corresponding frequencies. Then a corresponding vector can be constructed to represent the document. Therefore, a collection of documents can be represented by a *term-by-frequency* matrix, which can be subsequently interpreted as a relational table. However, the vector space model only allows preserving fundamental features of the document. Although a few alternative weighting schemes other than term frequency have been proposed, one common weakness is that they don't take into consideration the associations among terms. Recent studies have revealed that association among terms could provide rich semantics of the documents and serve as the basis of a number of text mining tasks [9]. However, the approach proposed in [9] discards the vector space model and uses frequently co-occurring terms only. The basic idea is to mine the associations among terms, and then capture all these information in a graph. They use text classification to illustrate the potential application of the new model. To that end, design two novel similarity functions. One is based on Google's page-rank style algorithm [3] to discover the weights of each term. The other is based on the distances of all pairs of terms.

In Term Based document is unity and used to identify content of text. In Term based method each term in document is associated with value known as weight. This is used to measure importance of term. That is, terms contribution in document. Word having semantic meaning is known as term and collection of such terms contributes meaning to documents. Mostly term based methods suffer from the problems of polysemy and synonmy. Polysemy means a word has multiple meanings. Same as Multiple words having the same meaning is known as synonmy. Information retrieval provided many term-based methods like supervised and also the term weighting methods to solve this challenge. The evolution of term weights is based on distribution of term in document. The term frequency TF (t, d) is number of times term't' occurs in document'd'. The document frequency DF (t) is number of documents in which term't' occurs at least once. The inverse document frequency IDF (t) can be calculated from document frequency.

The inverse document frequency of term is low if it occurs in many documents and highest if term occurs only in one document. The value of W means weight of term of document'd' calculated by product as Wi=TF(ti,d)* IDF(ti). Using this mathematical model term based method analyse document to extract features by TFIDF feature selection approach.

II. CONCEPT BASED TEXT MINING

The concept based analysis algorithm describes the concepts in the documents. This represents the semantic structures of the sentence and it is processed sequentially. Each concept in the present document is matched with other related concepts and also compared with previously processed documents concepts.

Most of text mining techniques are based on word and
/or phrase analysis of text. It is important to find term that contributes more semantic meaning to document this concept is known as concept based method. Only the importance of term within document is captured in statistical analysis of term based method. Only the importance of term within document is captured in statistical analysis of term based method. In concept based method the term which contributes to sentence semantic is analysed with respect to its importance at sentence and also document levels. The model is used to analyze term at sentence and document level by efficiently finding significant matching term rather than single term analysis. Conceptual term frequency to analyse each concept at sentence level is proposed. The 'ctf' is number of occurrences of concept c of sentence's' and 'tf' is term frequency to analyse each concept at document level. There are number of occurrences of concept 'C' in original document. The concept based term analyser algorithm describes how to calculate 'tf' and 'ctf' of matched concept in document.

The matched concepts list L algorithm starts with processing of new document has well defined sentence boundaries. The length of matched concepts and their verb argument stored concept based similarity calculation. For each sentence the concepts of the verb argument structures which represent the semantic structures of the sentence are processed sequentially. Each concept in the current document is processed sequentially. Each concept in the current document is matched with the other in the previously processed documents. To match the concept in previous documents is accomplished by keeping a concept list L that holds the entry for each of the previous documents that shares a concept with the current document. After the document is processed, L contains all the matching concepts between the current document and any previous document that shared at least one concept with the new document. After the document is processed, L contains all the matching concepts between the current document and any previous document that shares at least one concept with the new document. Finally, L is output as the list of document with the matching concepts and the needed information about them. The concept base analyser algorithm is capable of matching each concept in a new document (d) with all previously processed documents in O(m) time. Where m is the number of concepts in d. The sum between the two values of tfweight i1 and ctfweight i1 presents an accurate measure of the contribution of each concept to the meaning of the sentences and to the topics mentioned in a document. The term weight i2 is applied to the second document d2.

In Concept based mining model context information which is essential in determining an accurate similarity between documents. A concept-based analysis is similarity between documents. A concept based similarity measure, based on matching concepts at the sentence, rather than on individual terms words only is devised. There are three critical aspects. First one is the analyzed labelled terms are the concepts that capture the semantic structure of each sentence. Secondly the frequency of a concept is used to measure the contribution of the concept to the meaning of the sentence as well as to the main topics of the document. Finally the number of documents that contains the analyzed concepts is used to discriminate among documents in calculating similarity. These aspects are measured by the proposed concept based similarity measure which measures the importance of each concept at the sentence level by the ctf measure, document level by the tf measure, and corpus level by the df measure. The concept based mining model system is a text mining application that uses the concept based similarity measure to determining the similarity between the documents. A raw text document is input to the proposed system by the user. Each document has well defined sentence boundaries. Each sentence in the document is labelled automatically based on

the prop Bank notations. After running the semantic role labeller, each sentence in the document might have one or more labelled verb argument structures.

III. SENTENCE BASED TEXT MINING

The sentence based concept analysis is to analyze each concept at the sentence level. The concept based frequency measure called the conceptual term frequency (ctf) is proposed. The ctf is proposed concept c in sentence s and document d are as follows: calculating ctf of concept c in sentence s: the ctf is the number of occurrences of concept c in verb argument structures of sentence s. The concept c which frequently appears in different verb argument structures of the same sentence s will have a larger contribution to make to the meaning of S. Thus the ctf measure is a local measure on the sentence level. Calculating ctf on concept c in Document d: A concept can have many ctf values in different sentences in the same document d. The ctf value of concept C in document d is calculated by,

$$Ctf = \frac{\sum_{n=1}^{sn} ctfn}{sn}$$

Where sn is the total number of sentences that contain concept c in document d. The average of the ctf values of concept c in its sentences of document d measures the overall importance of its sentences in document d. A concept that has ctf values in most of the sentences in a document has a major contribution to the meaning of its sentences that leads to discover the topics of the document. The calculating the average of the ctf values measures the overall importance of each concept of the semantics of the document through its sentences. The number of generated verb argument structures is entirely dependent on the amount of information in the sentence. The sentence that has many labelled verb argument structures includes many verbs associated with their arguments. The labelled verb argument structures, the output of the role labelling task, are captured and analyzed by the concept based mining model on the sentence and labelled terms are considered concepts. One term can be an argument to more than one verb in the same sentence. It means this term can have more than one semantic role in the same sentence and hence it contributes more to the meaning of the sentence.

IV. DOCUMENT BASED TEXT MINING

Document based concept analysis is used to analyze the concept at the document level. The concept based term frequency tf, the number of the occurrences of a concept c in the original document is calculated. The tf is a local measure on the document level. Use feature-vector to represent documents, that is, take one document as a set of Term Sequences, including term t and term weight w. Then the document will be made up of the pairs of <t,w>. $t_1,t_2, t_3 \ldots t_n$ represents the features that is expressed the document content. And also treat them as an N-dimension coordinate. W1,W2,W3…Wn represent the value relevant to coordinate. So every document(d) is mapped to the target space as a feature-vector V(d)=(t1,w1,t2,w2,t3,w3….tn,wn). The most important of data pre-processing is to deal with the data resource and also build up the feature vectors. Also use the weight as the criterion of feature selection. The values of the vector elements w_i for a document d are calculated as a combination of statistics TF(t,d) and DF(t). The term frequency TF(t,d) is the number of times word t occurs in document d. The document frequency DF(t) refers to the number of document in which the word t occurs at least once. The inverse document frequency IDF(t) can be calculated from the document frequency.

Log(||)/DFtD*IDFt*|*D*|

|*D*| is the whole number of documents. The document frequency of inverse of a word is low if it occurs in many documents and is highest if the word occurs in only one. The values w_i of features t_i for document d is then calculated as the product (,) ( )(2) iii W=TF td.IDF t W_i is called the weight of the word t_i in document d. The heuristically says the word weighting a word t_i is an important indexing term for document d if it occurs in frequently in it. A word which occurs in many documents is rated less important indexing terms due to their low inverse document frequency. And also used to find the above IDF servers as an adjusting function to modulate the term frequency.

V. VECTOR SPACE MODEL

The Basic idea of vector space model is representing the document in computer understandable form. Bag of word model is one of forms to represent the document. In space Model, any text document is represented as vectors or dimensions. Each dimension of space is to represent as a single feature of the vector and the weight is calculated by various weighting schemes. The document can be represent as d = (t_1,w_1;t_2,w_2….t_n,w_n) which t_i is a terms w_i is the weight of the t_i in the document d. Reflect the importance of the term in a document use term weighting. There is a different term weighting methods proposed in the TC study. Four term weighting approaches which are proved to be prominent in TC, Term Frequency (TF). Each and every term is assumed to have a value proportional to the number of times it occurs in a document.

W(d,t)=TF(d,t)

Term Frequency

Inverse Document Frequency TF and IDF to weight the terms and the performance with reference accuracy that IDF and TF separately.

W(d,t)=Tf(d,t).IDF(t) N documents, if n document contains the term t, IDF

$$IDF(t)=\log(\frac{N}{n})$$

Term Frequency-CHI Square

It is a typical representation which combines TF factor with one feature selection metric i.e CHI-Square.

Term Frequency-Relevance Frequency (TF-RF)

The best term weighting approach for TC documents. Hence this method is considered mostly defined the document representation.

$$TF.RF(t)=TF*\log(2+\frac{a}{\max(1,c)})$$

Where a is the number of documents which contain the positive category term, c is the number of documents which contains negative category term.

VI. A TOPIC BASED DOCUMENT RELEVANCE MODEL

Pattern Enhanced Topic Model (PETM) is used. PETM determine document relevance based on topics distribution and maximum matched patterns. Latent Dirichlet Allocation (LDA) is one of the most admired probabilistic text modeling techniques. It is used to discover the hidden topics in collections of documents with the appearing words. PETM pattern mining is used to find the semantically meaningful and efficient patterns to represent topics and documents are implemented in two steps.

Firstly, construct a new transactional dataset from the LDA outcomes of the document collection.

Secondly, generate pattern based representation from the transactional dataset to represent user needs.

VII. EXPERIMENTAL RESULT

In order to compare the results of all possible combinations of term weighting methods with classifiers, and computed the precision and recall. F1 measure Precision is the proportion of examples labeled positive by the system that were truly positive and recall is the proportion of truly positive. Reuters Data set is one of the famous dataset. Reuters is a new version of the dataset collection. In this process used the 12,902 stories that had been classified in to 118 categories. The stories average about 200 words in length. The ModApte split in which 75% of the stories are used to build classifiers and the remaining 25% to test the accuracy of the resulting models in reproducing the manual category assignments. The stories are split temporally the training items all occur before the test items. As already mentioned, all classifiers output a graded measure of category membership, so different thresholds can be set to favour precision or recall depending on the application- for Reuters optimized the average of precision and recall. All model parameters and thresholds are set to optimize performance on a validation set and are not modified during testing. For Reuters, the training set contains 9603 stories and the test set 3299 stories. In order to decide which models to use they performed initial experiments on a subset of the training data, subdivided into 7147 training stories and 2456 validation stories for this purpose. They used to set the number of features (k), decision thresholds and document representations to use for the final runs. Estimated parameters for these chosen models using the full 9603 training stories and evaluated performance on the 3299 test items. They optimize performance by tuning parameters to achieve optimal performance in the test set.

VIII. CONCLUSION

Term weighting plays an important role to get high performance in text classification. The traditional tf-idf algorithm is a popular method for document representation and feature selection. Here, compared different term weighting scheme, makes use of a kind of information ratio to judge a term's contribution for category along with class information. This used to analyse the term weight based methods to analyse which is related to text classification method as well as text document methods. TF-Relevance Factor performs better for all categories.

REFERENCES

[1] An Efficient Concept-based Mining Model for enhancing Text Clustering, Shady Shehata, Fakhri Karray, Mohamed S.Kamel, IEEE Transactions on knowledge and Data Engineering, vol 22, no 10, October 2010.

[2] K.J. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 1998.

[3] B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.

[4] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report 941, Norwegian Computing Center, June 1999.

[5] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Comm. ACM, vol. 18, no. 11, pp. 112-117,1975.

[6] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[7] U.Y. Nahm and R.J. Mooney, "A Mutually Beneficial Integration of Data Mining and Information Extraction," Proc. 17th Nat'l Conf. Artificial Intelligence (AAAI '00), pp. 627-632, 2s000.

[8] L. Talavera and J. Bejar, "Generality-Based Conceptual Clustering with Probabilistic Concepts," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 196-206, Feb. 2001.

[9] H. Jin, M.-L. Wong, and K.S. Leung, "Scalable Model-Based Clustering for Large Databases Based on Data Summarization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1710-1719, Nov. 2005.

[10] T. Hofmann, "The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data," Proc. 16th Int'l Joint Conf. Artificial Intelligence (IJCAI '99), pp. 682-687, 1999.

[11] Eugene Agichtein , Silviu Cucerzan ,"Predicting Accuracy of Extracting Information from Unstructured Text Collections", *CIKM'05*, October 31-November 5, 2005.

[12] Eugene Agichtein, "Scaling Information Extraction to Large Document Collections", IEEE Computer Society Technical Committee on Data Engineering, 2005.

[13] Raymond J. Mooney and Razvan Bunescu, "Mining Knowledge from Text Using Information Extraction", SIGKDD Explorations, 2005.

[14] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, 2013.

[15] Shady Shehata,Fakhri Karray, Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions on Knowledge and Data Engineering, , 2010.

[16] Helena Ahonen and Oskari Heinonen "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections"Published in the Proceedings of ADL'98, April 22-24, 1998 in Santa Barbara, California,USA.

[17] X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In IJCAI'03, pages 587–594, 2003.

ABOUT AUTHORS

S.Brindha received B.Sc degree in Physics from Bharathiyar University. She done her Master Degree in Information Science and Management in Periyar University and she awarded M.Phil Computer Science from the Bharathiyar University. She has 3 years of teaching experience and 5 years of Technical Experience in Hash Prompt Softwares Pvt. Ltd. Currently She is doing her Ph.D computer Science in Erode Arts and Science College. Her Research area includes Data Mining and Text Mining.

K.Prabha received B.Sc Computer Science and M.Sc Computer Science Degree from Bharathiar University, Coimbatore and M.Phil in Periyar University, Salem. She received thePh.D degree in Computer Science at Bharathiar University. She has 7 years of teaching experience. She was working as Assistant Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu, India. She is working as a Associate professor in PG Extension Center, Periyar University, Dharmapuri, Tamilnadu, Erode. Currently she is Guiding 1 M.Phil Scholar and 1 Ph.D Scholar. She published around 15 research papers in national and international journals and conferences. Her research interests include Network Security and Data Mining.

Dr. S. Sukumaran graduated in 1985 with a degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 25 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu, India. He has guided for more than 50 M.Phil research Scholars in various fields and guided 5 Ph.D Scholars. Currently he is Guiding 5 M.Phil Scholars and 6 Ph.D Scholars. He is a member of Board studies of various Autonomous Colleges and Universities. He published around 40 research papers in national and international journals and conferences. His current research interests include Image processing and Data Mining.

