ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Improving Text Categorization Using Semantic Measures

Anusha Done
Persuing M.Tech,
Department of Computer Science and Engineering,
B.V.Raju Institute of Technology,
Narsapur, MedakDist.

K.Purnachand
Assistant Professor,
Department of Computer Science and Engineering,
B.V.Raju Institute of Technology,
Narsapur, MedakDist.

Abstract:

Classification of Text documents is a remarkably notified problem in library science, information science and also computer science. The task can be accomplished manually or algorithmically with different types of classifications like content based or request oriented. The predominant techniques in classifications are expectation maximization, nalve Bayes classifier, instantaneous trained neural networks, latent semantic indexing, support vector machines, artificial neural network, K-nearest neighbour algorithms, decision trees, concept mining and etc.. All these techniques are used to achieve the classification at maximum level. The existing methods are performing on frequency feature selection methods. The high accuracy can be achieved with the semantic indexing and measures only. The present paper is focusing on the Text Categorization Using Semantic Measures. The semantic measures like Term Rank Identifier with the class labels and Terms clustering algorithm are used to get increasing accuracy in the classification of Text categorization. The proposed paper is developed with a novel Semantically Enriched-Terms Clustering (SETC) Algorithm to achieve the better results.

Keywords:

Text Categorization – Clustering – Semantic measures – term rank identifier – Semantically enriched terms clustering - Semantically EnrichedTerms Clustering (SETC) Algorithm.

Introduction:

Internet browsing has become a part and parcel of the present day world for any requirement. Especially when a research scholars and readers who wants to read a specific concept or paper will mostly depend on the web browsing. Predominantly the search engines are playing vital role in providing billions of textual documents to the browsers. Of course, the web search engines are providing relative documents to the web browsers but not accurately what the browser needs and what is demanded in the search engines. Hence there is a great need to improve the quality and effective clustering on high dimensional data. The dimensional data has been improvised by many research scholars by developing new methods to analyze the document clustering. But all these approaches have concentrated on the factors like term frequency, document frequency, feature selection and support vector machines [SVM]. These approaches could not achieve the high accuracy in processing the dimensional data available in the net.

The present paper is developed to focus on improving the text categorization on text document clustering with the help of TRI and SETC to boost up the text categorization from the dimensional data. The proposed paper is enriched with the novel concepts like Term Rank Identifier and Semantically Enriched Terms Clustering (SETC) Algorithm to achieve high accuracy in processing the documents from high dimensional data. The important goal of the project is to measure the most frequent terms occurring on any text document clusters with the proposed measure Term Rank Identifier, wordnet tools. The targeted goal is to reduce the high dimensionality of data and to apply the clustering algorithms semantically and should be measured by TRI.

Related Work:

Y. H. LI [1998] has released a paper in Computer journal about the document clustering. The paper has illustrated three classifier combination approaches. These are simple voting, dynamic classifier selection and adaptive classifier combination. The performance results have revealed that the adaptive classifier combination method has achieved nearly 83 per cent of accuracy in clustering the documents. In another research work published in international journal stated that the advanced clustering can be possible with the help of keywords extraction.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The research work has focused on a lexical database of English words. The keyword extractions process has been done with the help of TF-IDF and WordNet. TF-IDF is an algorithm which works on K-Nearest Neighbor algorithms. The combination of these algorithms has given better accuracy for text classification. Menaka S et.al. [2013]. In the year 2013 another research work has introduced the text documents categorization method using frequent term sets. In this work a new concept through measuring AD-Sup has been introduced to extract discriminative features from frequent terms set for classification task.

The classification of documents results have been obtained with the help of Web-KB corpus demonstrate and declared that AD-Sup constraint is very effective in clustering the documents of same kind- Man Yuan [2013]. The accuracy has been increased by implementing a novel classification system with information retrieval technology and machine learning technology. The new technique used for text documents clustering is a combination of two different clustering algorithms Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents aimed to get increased accuracy- Bhumika [2013].

Text categorization Measures Text Mining:

While the text mining is incorporated the information retrievals are performed with the help of x2 statistics. X2 statistics are predominantly used to calculate the term frequencies and term category dependencies. The text clustering can be done by calculating the co-occurrences of the term and listed in contingency tables. It can be exemplified with an example of a corpus with a bunch of n documents with n number of categories. The stemming and removal of stop words the distinct terms can be extracted from the corpus.

The dependency test can be incorporated for X2 statistics with two strategies of null hypothesis and alternative hypothesis. The null hypothesis reveals that the two variables, term and category are independent of each other. The alternative hypothesis demonstrates that there is remarkable dependency between the two variables. The following formula can reveal the dependency calculation is

$$\chi^2 = \sum_{i=1}^{k} \left| \frac{(O_i - E_i)^2}{E_i} \right|$$

In the above calculation Oi– identified frequency in the ith cell of the table.
Ei – the expected frequency in the ith cell of the table
It is assumed that the degrees of freedom are (r-1) (c-1). In this r can be considered as number of rows and C can be calculated as number of columns.

Term Rank Identifier – TRI:

It is observed that x2 statistics did not fully explore the information facilitated in term category independence test. The research has focused on the positive term category dependencies. This mechanism has given rise to the text clustering based on the frequent words. Hence the research has diversified into term categoriy dependency measure denoted TRI. This has identified highly related terms based on the frequencies and each term is assigned with ranks and category by its semantics. Terms Clustering (SETC) Algorithm, it is integrated with TRI improves the clustering. The proposed project is developed to attain the Accuracy which leads to incremental text categorization Semantically Enriched. Terms Clustering (SETC) Algorithm, it is integrated with TRI improves the clustering accuracy which leads to incremental text categorization. The following table illustrates the text categorizing techniques achieved through the SETC algorithms. SETC algorithms are very effective and produce more accuracy in clustering the text documents.

Category / Term	C_1	C_2	C_3	Frequency	Rank
T_1	d_1	d_1, d_4	d_3	5	1
T_2	d_1, d_2	d_1, d_2		4	2
T_3	d_5	d_2		2	4
T_4	d_2, d_5	d_4		3	3

The above table is illustrating the classification with semantically related terms. The following table illustrates the results obtained from semantically related terms.

Category / Term	C_1	C_2	C_3
T_1	T_2, T_3	T_2, T_3	T_2, T_3
T_2	T_1	T_1	
T_3	T_1		T_2
T_4	T_2	T_2	
Total terms (union)	3	3	2

In this table the example is described as the database D contain 5 documents D= {d1, d2, d3, d4, d5}. These documents are categorized into three categories C1, C2 and C3.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

These categories are consists of documents C1={d1,d2,d5} C2={d1,d2,d4} and the last category C3 consists of d3 document only. In this process it is observed that four different terms t1, t2,t3 and t4 are found.

Text Clustering – Achieved:

Most of the text clustering algorithms are represented by the vector space model. This model is predominantly considering the documents as vectors. When it is considered as term frequency it will be identified as term space. The processing steps are rich with stop words, removal and stemming on the documents which have to be clustered. In this process IDF is implemented. IDF means Inverse Document Frequency. The implementation can give more clarity with the documents in the corpus. Different criterion will be used for different problems of clustering text documents. Cosine function is predominantly used to measure the similarity between two text documents as the correlation between the document vectors representing them. The following calculation can reveal the similarity between two documents di and dj.

Cos[di, dj] = di*dj / ||di|| ||dj||

In this above equation * represent the vector dot product and ||di|| represent length of vector di. The cosine value can be calculated as 1 when two documents are identical and 0 if there is nothing in common between them. These two documents are represented by the larger cosine value. At this juncture K-means algorithm is widely used for this type of problems of clustering a data set into K clusters. If the dataset consists of n documents spanned from d1,d2,d3…dn, then the clustering is the optimization process of grouping them into k clusters so that the global criterion function is either minimized or maximized.

$$\sum_{j=1}^{k}\sum_{i=1}^{n} f(d_i, Cen_j)$$

In the above equation Cenj is representing the centroid of a cluster cj, for j=1:….;k, and f(di,Cenj) is the clustering criterion function for a document di and a Centroid cenj. At the time of usage of cosine function every document is assigned to the cluster with the most similar centroid, and the global criterion function is maximized as a result.

Semantically Enriched Terms Clustering:

The proposed concept and clustering is performed with the novel functionality called semantically enriched terms clustering. This is related with the TRI metric derived from semantically enriched terms clustering. Word net3.0 is capable of calculating the synonyms and estimated relative frequencies of given terms. The basic goal of this algorithm is to achieve the generation of semantically highly related terms. The input can be given with set of different text documents and wordnet 3.0 for demantics. The output can be obtained as categorized class labels which generates taxonomies. he process can be defined in 5 processing steps.
In the first step the collection of documents D={d1,d2,d3,d4,d5} should be analyzed and calculate the unigrams, bigrams, trigrams and multigrams severally.

Unigram represent – frequently occurring 1 word
Bigram represents frequently occurring 2 words
Trigram represents frequently occurring 3 words

Multigram represents frequently occurring 4 or more words.In the step twoanalyze the documents based on the their relative frequencies in a single documents or in clustered documents. RANK can be calculated with the TERM Frequency [TF], Min_Support=2Step three initiates the identification of the semantic relationship between the terms by using a Lexical Analyzer with the help of Wordnet 3.0This can be declared as Sem_Rel[terms] is equal to Synonyms or estimated Relative Frequency. Step four should initiate the categorizing the semantically enriched terms into different categories by justifying the class labels.Step five is finalizing the construction of taxonomies which are generated by class labels. The paper is proposing Semantically Enriched Terms Clustering (SETC) Algorithm clusters all the documents into k no of clusters. The proposed method is different traditional K-Means and K-Medoids partition algorithms. These algorithms do clustering as a mean of the data objects and centroid values. This paper compare to these traditional algorithms our proposed SETC algorithm with TRI metric is out performing and improving the accuracy of text categorization by focusing the term semantics.

Experimental Results:

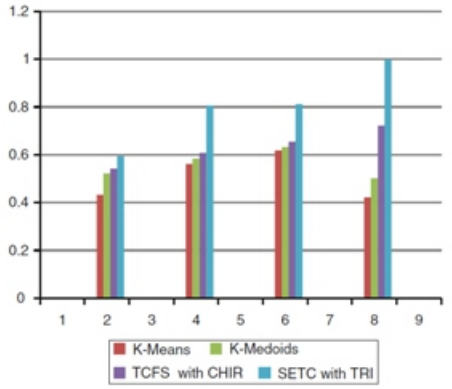
The performance comparison between x2 statistics and TRI

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Category	C_1		C_2		C_3	
Terms	χ^2 statistics	TRI	χ^2 statistics	TRI	χ^2 statistics	TRI
T_1	0.540	1.198	0.540	1.198	0.540	1.198
T_2	0.423	1.023	0.423	1.023	0	0
T_3	0.227	0.546	0	0	0.227	0.546
T_4	1.121	1.242	1.121	1.242	0	0

Performance comparison of SETC with other clustering methods

Data set	K-means	K-Medoids	TCFS with CHIR	SETC with TRI
20-News Groups	0.432	0.522	0.542	0.594
Reuters	0.562	0.584	0.608	0.806
PubMed	0.618	0.632	0.654	0.812
Wordsink	0.422	0.502	0.722	0.998



The above figure demonstrates the performance improvements of SETC with different clustering algorithms. The results are obtained on datasets like 20 –New Groups, Reuters, PubMed and Wordsink. The above results reveal that the performance of SETC is more obviously greater than the above mentioned clustering methods.

Conclusion:

The proposed paper is rich Term Rank Identifier to calculate highly related terms based on synonyms and expected relative frequencies. The comparison has been done on real data sets which available measures like x2 statistics and GSS Coefficients. In this project observation the performance of Term Rank Identifier is exceptionally good. In this project a novel Text Clustering algorithm named as Semantically Enriched Terms Clustering has been introduced to integrate with the TRI. The proposed clustering algorithm is compared with other clustering mechanism and demonstrated the results good. The outstanding performance has been achieved by the proposed text clustering algorithm used in this project.

Future Scope of Study:

In the future scope of the study the text categorization and clustering capabilities have to enhance. This can be achieved with the implementation of additional measures which are increasing the scope of the clusters independently.

The future scope of the study can demand the implementation of building ontologies automatically. This can be possible with the introducing NLP Lexical Analyzers.

References:

[1]Y. H. LI AND A. K. JAIN [1998] Classification of Text Documents published in The Computer Journal, Vol 41, No.8 1998

[2]Menaka S and Radha N [2013] Text Classification using Keyword Extraction Technique published in International Journal of Advanced Research in Computer Science and Software EngineeringVolume 3, Issue 12, December 2013.

[3]Man Yuan, Yuan Xin Ouyang1, And Zhang Xiong [2013] A Text Categorization Method using Extended Vector Space Model by Frequent Term Sets published by Journal of Information Science And Engineering 29, 99-114 (2013)

[4]Bhumika, Prof Sukhjit Singh Sehra, Prof AnandNayyar [2013] A REVIEW PAPER ON ALGORITHMS USED FOR TEXT CLASSIFICATION published in International Journal of Application or Innovation in Engineering & Management (IJAIEM)

[5]K. Purna Chand and G. Narsimha [2015]An Integrated Approach to Improve the Text Categorization Using Semantic Measures published in Springer India 2015

[6]Michael Steinbach George KarypisVipin Kumar [2013] A Comparison of Document Clustering Techniques published by springer India 2015.

[7]ChandanJadon Ajay Khunteta [2013] A New Approach of Document Clustering published in International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 4, April 2013

[8]Yogesh Jain, Amit Kumar Nandanwar [2013] A Theoretical Study of Text Document Clustering published by Yogesh Jain et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2246-2251

[9]RakshaK.Mundhe, AnkushMaind,R.B.Talmale [2014] Information Retrieval Using Document Clustering for Forensic Analysis published by International Journal of Recent Advances in Engineering & Technology (IJRAET) ISSN (Online): 2347 - 2812, Volume-2, Issue -5, 2014

[10]JayarajJayabharathy and SelvaduraiKanmani [2014] Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature published by Jayabharathy and Kanmani Decision Analytics 2014, 1:3 Springer