An Integrated Approach to Improve the Text Categorization Using Semantic Measures

K. Purna Chand and G. Narsimha

Abstract Categorization of text documents plays a vital role in information retrieval systems. Clustering the text documents which supports for effective classification and extracting semantic knowledge is a tedious task. Most of the existing methods perform the clustering based on factors like term frequency, document frequency and feature selection methods. But still accuracy of clustering is not up to mark. In this paper we proposed an integrated approach with a metric named as Term Rank Identifier (TRI). TRI measures the frequent terms and indexes them based on their frequency. For those ranked terms TRI will finds the semantics and corresponding class labels. In this paper, we proposed a Semantically Enriched Terms Clustering (SETC) Algorithm, it is integrated with TRI improves the clustering accuracy which leads to incremental text categorization. Our experimental analysis on different data sets proved that the proposed SETC performing better.

Keywords Text categorization · Clustering · Semantic knowledge · Term rank identifier · Semantically enriched terms clustering

1 Introduction

Today the world became web dependent. With the booming of the Internet, the World Wide Web contains a billion of textual documents. To extract the knowledge from high dimensional domains like text or web, our search engines are not enough smart to provide the accurate results. This factor leads the WWW to urgent need for effective clustering on high dimensional data.

K. Purna Chand (✉) · G. Narsimha
Department of CSE, JNTU College of Engineering, Kakinada, Andhra Pradesh, India
e-mail: purnachand.k@gmail.com

G. Narsimha
e-mail: narsimha06@gmail.com

L.C. Jain et al. (eds.), *Computational Intelligence in Data Mining - Volume 2*,
Smart Innovation, Systems and Technologies 32, DOI 10.1007/978-81-322-2208-8_5

Many traditional approaches are proposed and developed to analyze the high dimensional data. Text Clustering is one of the best mechanisms to identify the similarity between the documents. But most of the clustering approaches are depends upon the factors like term frequency, document frequency, feature selection and support vector machines (SVM). But there is still uncertainty while processing highly dimensional data.

This research is mainly focuses on improving the text categorization on text document clusters. The proposed TRI and SETC will boost up the text categorization by providing semantically enriched document clusters. The primary goal is to measure the most frequent terms occurring on any text document clusters with our proposed metric Term Rank Identifier (TRI). For those frequent terms the semantic relations are calculated with Wordnet Tools. The basic idea behind the frequent item selection is to reduce the high dimensionality of data. The secondary goal is to apply our proposed text clustering algorithm Semantically Enriched Terms Clustering (SETC) to cluster the documents which are measured by TRI.

2 Related Work

There exist two categories of major text clustering algorithms: Hierarchical and Partition methods. Agglomerative hierarchical clustering (AHC) algorithms initially treat each document as a cluster, uses different kinds of distance functions to compute the similarity between all pairs of clusters, and then merge the closest pair [1]. On other side Partition algorithms considers the whole database is a unique cluster. Based on a heuristic function, it selects a cluster to split. The split step is repeated until the desired number of clusters is obtained. These two categories are compared in [2].

The FTC algorithm introduced in used the shared frequent word sets between documents to measure their closeness in text clustering [3]. The FIHC algorithm proposed in [4] went further in this direction. It measures the cohesiveness of a cluster directly by using frequent word sets, such that the documents in the same cluster are expected to share more frequent word sets than those in different clusters. FIHC uses frequent word sets to construct clusters and organize them into a topic hierarchy. Since frequent word sequences can represent the document well, clustering text documents based on frequent word sequences is meaningful. The idea of using word sequences for text clustering was proposed in [5]; However, STC does not reduce the high dimension of the text documents; hence its complexity is quite high for large text databases.

The sequential aspect of word occurrences in documents should not be ignored to improve the information retrieval performance [6]. They proposed to use the maximal frequent word sequence, which is a frequent word sequence not contained in any longer frequent word sequence. So, in view of all the text clustering algorithms discussed above we proposed TRI and SETC.

Table 1 General notation of 2 × 2 contingency table

Category	Category 1	Category 2	Total
Term			
Term 1	a	b	a + b
Term 2	c	d	c + d
Total	a + c	b + d	a + b + c + d = n

2.1 Traditional Text Categorization Measures

2.1.1 χ^2 Statistics

In text mining for the information retrievals, we frequently use χ^2 Statistics in order to measure the term frequencies and term-category dependencies. It can be done by measuring the co-occurrences of the terms and listed in contingency tables (Table 1). Suppose that a corpus contains n labeled documents, and they fall into m categories. After the stop words removal and the stemming, distinct terms are extracted from the corpus.

For the χ^2 term-category dependency test, we consider two strategies one is the null hypothesis and the alternative hypothesis. The null hypothesis states that the two variables, term and category, are independent of each other. On the other hand, the alternative hypothesis states that there is some dependency between the two variables.

General formula to calculate the dependency is

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

where

$\mathbf{O_i}$—the observed frequency in the ith cell of the table.
$\mathbf{E_i}$—the expected frequency in the ith cell of the table

The degrees of freedom are (r − 1) (c − 1). Here r = # of rows and c = # of columns.

2.2 Term Rank Identifier (TRI)

In our exploration, we found that $\boldsymbol{\chi^2}$ does not fully explore all the information provided in term-category independence test. We point out where the problem is due to identifying only positive term category dependencies based upon the frequent words. In view of this, we proposed a new term-category dependency measure, denoted TRI, which identifies highly related terms based upon their frequencies and each term is assigned with ranks and is categorized by its semantics.

Table 2 Term-ranking based on their frequencies

Category	C_1	C_2	C_3	Frequency	Rank
Term					
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

Table 3 Calculating semantically related terms

Category	C_1	C_2	C_3
Term			
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

Example 1 For suppose a database D consists of 5 documents D = {d_1, d_2, d_3, d_4, d_5} are categorized as three categories c_1 = {d_1, d_2, d_5}, c_2 = {d_1, d_2, d_4} and C_3 = {d_3} and we observed four different terms t_1, t_2, t_3 and t_4.

The above illustrated example is represented in Table 2. If we observe closely that the term T_1 almost all occurred in all documents except in d_2, d_5. And coming to the term T_2 even its rank is 2 but it is occurred only in d_1, d_2 documents. Likewise by analyzing all the occurrences of different terms we concluded that term-category frequency is not much better in all cases. So our proposed metric Term Rank Identifier (TRI) measures the semantic relatedness (Table 3) of each term in every document.

So from Table 3 we can say that the terms T_1, T_2 and T_3 are semantically related to each and every category. Compare to c_3; c_1 and c_2 categories consists of highly related terms. So we can determine that documents of c_1 = {d_1, d_2, d_5}, c_2 = {d_1, d_2, d_4} and consisting of similar information and these documents are clustered by our proposed Semantically Enriched Terms Clustering (SETC) Algorithm.

3 Proposed Text Clustering Algorithm

3.1 Overview of Text Clustering

In many traditional text clustering algorithms, text documents are represented by using the vector space model [7]. In this model, each document d is considered as a vector in the term-space and is represented by term-frequency (TF) vector: Normally, there are several preprocessing steps, including the stop words removal and the stemming, on the documents. A widely used refinement to this model is to weight each term based on its inverse document frequency (IDF) [8] in the corpus.

For the problem of clustering text documents, there are different criterion functions available. The most commonly used is the cosine function [8]. The cosine function measures the similarity between two documents as the correlation between the document vectors representing them.

For two documents d_i and d_j, the similarity can be calculated as

$$\cos\left(d_i,\, d_j\right) = d_i * d_j / \|d_i\|\ \|d_j\| \tag{2}$$

where * represents the vector dot product, and $\|d_i\|$ denotes length of vector 'd_i'. The cosine value is 1 when two documents are identical and 0 if there is nothing in common between them. The larger cosine value indicates that these two documents share more terms and are more similar. The K-means algorithm is very popular for solving the problem of clustering a data set into k clusters. If the dataset contains n documents, d_1; d_2;…; d_n, then the clustering is the optimization process of grouping them into k clusters so that the global criterion function is either minimized or maximized.

$$\sum_{j=1}^{k}\sum_{i=1}^{n} f(d_i, Cen_j) \tag{3}$$

where Cen_j represents the centroid of a cluster cj, for j = 1;…; k, and $f(d_i,Cen_j)$ is the clustering criterion function for a document d_i, and a Centroid Cen_j. When the cosine function is used, each document is assigned to the cluster with the most similar centroid, and the global criterion function is maximized as a result.

3.2 Semantically Enriched Terms Clustering (SETC)

In the previous section we described that our proposed metric TRI identifies the semantically highly related terms. The semantic relativeness is calculated with the help of Wordnet 3.0. (Lexical Semantic Analyzer). It is used to calculate the synonyms and estimated relative frequencies of given terms.

Algorithm: The objective of the algorithm is to generate semantically highly related terms

Input: Set of different text documents and Wordnet 3.0. for Semantics.
Output: Categorized Class labels which generates taxonomies.

Step 1: Given a collection of text documents D = {d_1, d_2, d_3, d_4, d_5}. Finds the unigrams, bigrams, trigrams and multigrams for every document.

Unigram—Frequently Occurring 1 Word

Bigram—Frequently Occurring 2 Words
Trigram—Frequently Occurring 3 Words
Multigrams—Frequently Occurring 4 or more Words.

Step 2: Assign ranks to the each term based upon their relative frequencies in a single document or in clustered documents.

$$\mathbf{Rank} = \mathbf{Term\ Frequency}\ (\mathbf{TF}),\ \mathbf{Min\_Support} = \mathbf{2}$$

Step 3: Identify the semantic relationship between the terms by using a Lexical Semantic Analyzer **Wordnet 3.0**

$$\mathbf{Sem\_Rel}(\mathbf{Terms}) = \mathbf{Synonyms\ or\ Estimated\ Relative\ Frequency}$$

Step 4: Categorizing the semantically enriched terms into different categories by assigning the class labels.

Step 5: Construct taxonomies which are generated by class labels.

Primarily, we considered a single document d_1 and measured the term-category dependency and identified frequent terms and these terms are assigned with ranks based upon their frequencies in that particular document d_1. Next the semantic related ness between each terms can be measured with our metric TRI and terms are categorized according to synonymy and expected related frequencies with the help of Wordnet 3.0. Lexical Semantic Analyzer. Like that each document $d_2 \ldots d_n$ can be categorized with the help of our proposed metric TRI.

Later, our proposed Semantically Enriched Terms Clustering (SETC) Algorithm clusters all the documents into k no of clusters. Our proposed method is quite differentiated from traditional K-Means and K-Medoids partition algorithms. These algorithms do clustering as a mean of the data objects and centroid values. But 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.

4 Experimental Results

In this section, we compared our proposed metric with the existing measures like $\chi 2$ Statistics (Table 4) and observed that our metric TRI is identifying the semantically highly related terms effectively.

The performance of our integrated approach is compared with traditional and most familiar clustering algorithms like K-Means, K-Medoids and TCFS are applied

Table 4 Performance comparisons between χ2 statistics and TRI

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

Table 5 Performance comparisons 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

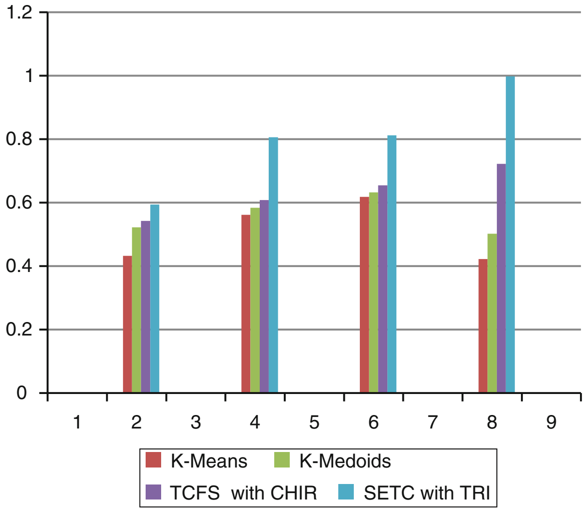

Fig. 1 Performance improvements of SETC with different clustering algorithms

on datasets like 20-News Groups, Reuters, PubMed and Wordsink, we observed that SETC (Table 5) with TRI is producing good results. The statistics are shown here.

Figure 1 represents the performance improvements of our proposed algorithm by comparing with traditional and well-known clustering algorithms.

5 Conclusion

In this paper, we introduced a new metric named as Term Rank Identifier (TRI) which calculates the highly related terms based upon their synonyms and expected relative frequencies. The comparison is made on real data sets with available measures like $\chi 2$ Statistics and GSS Coefficients; we observed that, it is performing well. And we proposed a Text Clustering algorithm named as Semantically Enriched Terms Clustering (SETC), which is integrated with TRI. Our proposed SETC algorithm is compared with other clustering and feature selection algorithms like K-Means, K-Medoids, TCFS with CHIR.The experimental results shows that our SETC is outperforming in terms of clustering accuracy on different data sets.

In Future, we enhance the text categorization and clustering capabilities by proposing additional measures which are independent of scope of the cluster. And we are planning to build ontologies automatically by introducing NLP Lexical Analyzers.

References

1. Liu, X., Song, Y., Liu, S., Wang, H.: Automatic taxonomy construction from keywords. In: Proceedings of KDD’12, pp. 12–16, August, Beijing, China (2012)
2. Li, Y., Luo, C., Chung, S.M.: Text clustering with feature selection by using statistical data. IEEE Trans. Knowl. Data Eng. **20**(5), 641–651 (2008)
3. Doucet, A., Ahonen-Myka, H.: Non-contiguous word sequences for information retrieval. In: Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004). Workshop on Multiword Expressions and Integrating Processing, pp. 88–95 (2004)
4. Fung, B.C.M., Wang, K., Ester, M.: Hierarchical document clustering using frequent itemsets. In: Proceedings of SIAM International Conference on Data Mining, pp. 59–70 (2003)
5. Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 436–442 (2002)
6. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD-2000 Workshop on Text Mining, pp. 1–20 (2000)
7. Ahonen-Myka, H.: Finding all maximal frequent sequences in text. In: Proceedings of ICML-99 Workshop on Machine Learning in Text Data Analysis, pp. 11–17 (1999)
8. A Clustering Toolkit, Release 2.1.1. http://www.cs.umn.edu/karypis/cluto/
9. Beydoun, G., Garcia-Sanchez, F., Vincent-Torres, C.M., Lopez-Lorca, A.A., Martinez-Bejar, R.: Providing metrics and automatic enhancement for hierarchical taxonomies. Inf. Process. Manage. **49**(1), 67–82 (2013)
10. Pont, U., Hayegenfar, F.S., Ghiassi, N., Taheri, M., Sustr, C., Mahdavi, A.: A semantically enriched optimization environment for performance-guided building design and refurbishment. In: Proceedings of the 2nd Central European Symposium on Building Physics, pp. S. 19–26, 9–11 Sept 2013, Vienna, Austria. (2013). ISBN 978-3-85437-321-6

11. Ahonen-Myka, H.: Discovery of frequent word sequences in text. In: Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining, pp. 16–19 (2002)
12. The Lemur Toolkit for Language Modeling and Information Retrieval. http://www-2.cs.cmu.edu/lemur/
13. Data Mining: Concepts and Techniques—Jiawei Han, Micheline Kamber Harcourt India, 3rd edn. Elsevier, Amsterdam (2007)