International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Improving Bisecting K-means by Applying Natural Language Processing

I.S.Choube[#1], Dr.S.D.Sharma [*2], Prof.Angad Singh [#3]

[#]Information Technology Department, RGPV University NIIST,1Sajjan Singh Nagar BOPAL-462021,INDIA
[1]choubeisha@gmail.com
[3]angada2007@gmail.com
[*]HOD Information Technology Department NIIST,1Sajjan Singh Nagar BOPAL-462021,INDIA
[2]smidhad2000@gmail.com

ABSTRACT
Text clustering techniques usually used to structure the text documents into topic related groups which can facilitate users to get a comprehensive understanding on corpus or results from information retrieval system. Most of existing text clustering algorithm which derived from traditional formatted data clustering heavily rely on term analysis methods and adopted Vector Space Model (VSM) as their document representation. But because of the essential characteristic underlying text such as high dimensionality features vector space, the problem of sparseness has a strong impact on the clustering algorithm. So feature reduction is an important preprocess step for improving the efficiency and accuracy of clustering algorithm by removing redundant and irrelevant terms from corpus. Even the clustering is considered as an unsupervised learning method, but in text, there is still some prior knowledge we can use from NLP analysis based approach. In this article, we propose a semantic analysis based feature reduction method which used in text clustering. Our method bases on a dedicated Part-of-Speech tags selection and Chunking reduce the feature space of documents more effectively compared with traditional feature reduction method tiff and stop words removal; meanwhile it preserves or sometimes even improves the accuracy of clustering algorithm. In our experiment, we tested our feature reduction method using bisecting k-means algorithm which was proved be efficient in text clustering. The results of our method is compared with k-means result applied on document clustering which shows that intra and inter cluster distance is reduced by our method than by using k-means. Our method can reduce the feature space significantly, and meanwhile have a better clustering accuracy compared with results in terms of the purity.
General Terms Clustering, Bisecting K-means algorithm, K-means Algorithm, Natural Language Processing
Keywords- Text clustering, feature selection, part-of-speech, chunking, K-means, Bisecting K-means.

I. INTRODUCTION

With the increasing prevalence of Web technologies, the amount of information which can be accessed by people has grown exponentially till now. How to find the useful information from the huge amount of data according to users intends at an effective and efficient way becomes more and more important. So, Web search engine has been an essential part in people's everyday life who suffering on the web. Based on the user's query, major commercial Web search engine usually return a huge list of related results which ranked by a sophisticated ranking algorithm [1-3] but usually the results are not all the user actually wants. A generally acknowledged issue in information retrieval systems, particularly in Web search engines, is that users queries are usually very short, sometimes even very ambiguous, so if the right results which user needed are not at the first several result pages, the searching will become a time consuming and annoying process, in which the user have to browse the result pages one by one.

Text clustering is suitable method to solve this kind of problem. As one of the most important text mining techniques, text clustering is developed to help users effectively navigate, summarize, and organize the results returned from search engine, and this lead to a significantly improvement on the precision and recall in information retrieval system [4]. Text clustering consists of four components, which are data representation model, similarity measure, clustering model and clustering algorithm. From all of these parts, the document representation is most important, because it determines the way that the other three parts choose. Most of the existing texts clustering methods are based on Vector Space Model [5], which represents documents as a feature vector of the words, a.k.a "bag of words", and statistical based word-weights, like *tfidf*, also accompany with it. Similarity between documents is measured by their distance or association coefficient like *Euclidean* distance or *cosine* measure, which mainly based on VSM. But due to the essential characteristic of text documents, the dimensionality of the feature vector is very huge, which imposes a big challenge to the performance of clustering algorithm. The clustering algorithm based on VSM could not work efficiently in high dimensional feature spaces due to the inherent sparseness of the data [6]. Not all features are useful for document clustering, and some of the features may be redundant or irrelevant. This situation gets worse especially in web documents for their incompact in content compared with formal text, and some of the features may even misguide the clustering results. In such cases, selecting a subset of original features often leads to a better performance. And also, feature selection not only reduces the high dimensionality of the feature space, but also provides a better data understanding, which can improve the accuracy of clustering results. The selected feature set should contain sufficient or more reliable information about the original data set.

The motivation behind the work in this article is that even text clustering is commonly treated as a unsupervised learning method, some kind of prior knowledge about nature language should helpful in text based feature selection process, which beyond the single word analysis. In this article, we proposed a novel feature selection method document clustering which based on semantic analysis, including a dedicated Part-of-Speech (PoS) tags selection and chunking.

II. RELATED WORK

The idea of text clustering derived from the traditional data clustering algorithm, so they share many same concepts. There are kinds of applications which can incorporate the text clustering technique to help users better organize their documents, such as clustering the results returned from search engine based on users' queries, like [7] clustering documents in a collection for automated construct the document taxonomies, like Yahoo directory1 and Open Directory Styles2; efficient information retrieval by focusing the query on relevant clusters rather than whole collections [8]. There are two general categories clustering algorithm used in text: one is agglomerative hierarchical algorithm, such as Hierarchical Agglomerative Clustering (HAC) [9], and the other is partitioning based methods, such as k-means algorithm [9-10]. Paper [11] compared these two kinds of clustering algorithm, and also proposed that Bisecting k-means is outperform both of these two categories algorithm in terms of accuracy and efficiency. Bisecting k-means is different from general k-means approach, and it splits a selected cluster abides by some criterion into two clusters until the number of clusters equal to the designated value.

Those clustering algorithm mentioned above are adopted from the traditional data clustering algorithm, which designed for clustering formatted data sets. So the special characteristics exist in text are not take care of well, such like the high dimensionally. To achieve a better result, a more informative feature unit – phrase has been considered in recent research. Paper [12] proposed a phrase-based document index model, named Document Index Graph, which allows the incremental construction of a phrase-based index for a document clustering. And the Suffix Tree Clustering (STC) algorithm [13-14] was proposed to be used in meta-searching engine to real-time cluster the document snippets returned from search engine. Compared with the traditional single-words based similarity computation, phrase-based document clustering approach achieved better accuracy.

Feature selection has been widely used in supervised learning, such as text categorization, and the class label information play a very important role to conduct the process of feature selection. For text clustering, there are just some unsupervised feature selection methods such as document frequency and term strength. Because there is no prior knowledge on the category structure can be used, so little research has been reported about the unsupervised feature selection in text clustering. Paper [15] proposed an Iterative Feature Selection (IF) method which utilizes the supervised feature selection to iteratively select features and perform text clustering. Paper [16] proposed a semi-supervised text clustering algorithm based on EM together with a feature selection technique based on Information Gain. Feature selections in both methods are semi-supervised. Latent Semantic Indexing [17] and Random Projection [18] can yield a considerable reduction in the dimension of the document representation, but their performance of clustering is not always remarkable [19].

Different with the statistics based feature selection method, there are kinds of approaches using the background knowledge underlying behind language to conduct feature selection, such as [20-22], which mainly depend on WordNet, and the results are encouraged. The relevant works similar with us were carried by Hotho et al. in [20-21], where they proposed background knowledge based feature standardization in text clustering using WordNet which can grasp the relationships between important terms that do not co-occur literally. But the works that have been reported in literature about using semantic feature selection to facilitate text clustering is limited. Part of speech (POS) tagging for English is often considered a solved problem. There are well established approaches such as Markov model trigram taggers [22], maximum entropy taggers [23], or Support Vector Machine based taggers (Gim´enez

and M`arquez, 2004), and accuracy reaches approximately 97%. However, most experiments in POS tagging for English have concentrated on data from the Penn Treebank [24]. If POS taggers trained on the Penn Treebank are used to tag data from other domains, accuracy deteriorates significantly.
Traditional approaches rely on preprocessing by an accurate POS tagger. Most work on shallow parsing is based on the English CoNLL'2000 shared task, which provided reference datasets for training and testing [25]. A number of approaches have been evaluated on these datasets, for general shallow parsing as well as for the simpler noun phrase chunking task: support vector machines (SVM) with polynomial kernel [26-27] and linear kernels [28], conditional random fields[29], maximum likelihood trigram models [30], probabilistic finite-state automata [31]. So far, SVM have achieved the best state of- the-art performances. The supervised English shallow parsing task and compare systems relying either on POS induction, on POS tagging, or on lexical features only as a baseline [32]. Michael Collins propose a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including part-of-speech tagging, chunking, named entity recognition, and semantic role labeling [33]. The encouraging results in tasks of classification make our approach appear promising for text clustering. Part-of-Speech also used in words meaning disambiguation and chunking parts of speech and short phrases and clustering the documents in topic related groups is similar to find the different meaning of words in documents in some sense.

III. PROPOSED METHOD

A Part Of Speech Selection

In our approach, we use Part-of-Speech selection. Using Part-of-speech, we can solve the problem of semantic ambiguity to some extent, so it is a very common tool in word sense disambiguation. The tags generated in our program are compatible with the Specification of Corpus Processing proposed by Peking University. This specification includes 35 Part-of-Speech categories with lots of related minor categories. For example the phrase in English need to be find with effort can be automatically labeled in our Part-of-Speech tagger and the tag related to phrase are divided into some Noun related minor categories. The tag set is listed in Table I.

TABLE 1. PART OF SPEECH TAG SET

SN	Tags	Explanation
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	to
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	WDT	Wh-determiner
33	WP	Wh-pronoun
34	WP$	Possessive wh-pronoun
35	WRB	Wh-adverb

B Chunking:
It is basically the identification of parts of speech and short phrases (like noun phrases). Part of speech tagging tells you whether words are nouns, verbs, adjectives, etc, but it doesn't give you any clue about the structure of the sentence or phrases in the sentence. Sometimes it's useful to have more information than just the parts of speech of words, but you don't need the full parse tree that you would get from parsing. Chunking fetch the action words that is usefull in searching and removed unimportant words.
C Clustering by Bisecting K-means:
Combination with PoS selection and chunking performs well in almost all datasets compared with each single one alone, and features are just half of them. So for unsupervised text clustering task, our unsupervised feature reduction based on PoS selection, chunking and combination of them two is very
efficient ,which can not only reduce the feature spaces accelerate the speed of clustering algorithm.

The algorirthm choosen is bisecting K-means which was proven to be best clustering method

IV. EXPERIMENTAL RESULTS

We use eight text documents dataset as input. At first we apply kmeans clustering on the dataset and form the required cluster of that dataset. Then we apply Bisecting k-means and form the cluster. We observe from the result that bisecting k-means with natural language processing works better than the k-means. The results are as follows:

For a given set of documents, we applied normal KMeans and Bisecting KMeans, and obtained some very good results, which are shown as follows:

Document Texts:
I am a good boy.
I am very bad boy.
He is going some where and he is a good boy.
I am a very good boy.
I am going to the market are you coming with me for shopping.
The fisherman went to the bank to catch fish.
The rule of the king has to be followed by the court.
This is a document for testing the algorithm. The document seems very good.

For 3 Clusters using KMeans

------------------------------[CLUSTER {1}]-----------------------------
I am a good boy

I am very bad boy

He is going some where and he is a good boy

I am a very good boy

------------------------------[CLUSTER {2}]-----------------------------
I am going to the market are you coming with me for shopping

The fisherman went to the bank to catch fish

------------------------------[CLUSTER {3}]-----------------------------
The rule of the king has to be followed by the court

This is a document for testing the algorithm. THe document seems very good
----

For 3 clustering using Bisecting k-means:
Action words fetch atfirst with natural language processing:-
good boy
bad boy
going good boy
good boy
going market coming shopping
fisherman went bank catch fish
rule of the king followed court
document testing algorithm THe document seems good

Clustering Results...
****
Cluster number 1
going market coming shopping
****
****
Cluster number 2
rule of the king followed court
document testing algorithm THe document seems good
****
****
Cluster number 3
good boy
bad boy
going good boy
good boy
fisherman went bank catch fish
****

Now after checking the clustering similarity the results of intra and inter cluster are:
1)Clusters form by k-means:
Enter Data in Cluster 1 (Separate documents by semi-colons(;)):
I am a good boy;I am very bad boy;He is going some where and he is a good boy;I am a very good boy
Enter Data in Cluster 2 (Separate documents by semi-colons(;)):I am going to the market are you coming with me for shopping;The fisherman went to the bank to catch fish
Cluster Analysis
Inter Cluster Similarity:59.0

Intra Cluster Similarity for Cluster 1:40.0
Intra Cluster Similarity for Cluster 2:6.0
2)Cluster form by Bisecting k-mean:
Enter Data in Cluster 1 (Separate documents by semi-colons(;)):going market coming shopping
Enter Data in Cluster 2 (Separate documents by semi-colons(;)):rule of the king followed court;document testing algorithm THe document seems good
Cluster Analysis
Inter Cluster Similarity:10.0
Intra Cluster Similarity for Cluster 1:0.0
Intra Cluster Similarity for Cluster 2:1.0

V. CONCLUSION

In This Article, We Proposed A Semantic Based Feature Reduction Method For Text Clustering Which Include A Dedicated Part-Of-Speech Tagging And Chunking. Experimental Results Will Show Its Efficiency In Reduction Of Features In Text Clustering Task Compared With Traditional Tfidf And Stopwords Removal Based Methods. Moreover The Feature Selection Method We Proposed Can Well Preserves Or Sometimes Even Improves The Accuracy Of Clustering Algorithm By Selecting The Most Meaningful Words And Proper Phrases.This Is Very Useful For Online Based Clustering Approach.

REFERENCES

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," : Technical report, Stanford Digital Library Technologies Project, 1998, 1998.

[2] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, pp. 604-- 632, 1999.

[3] M. Richardson, A. Prakash, and E. Brill, "Beyond PageRank: machine learning for static ranking," *Proceedings of the 15th international conference on World Wide Web*, pp. 707--715, 2006.

[4] R. C. Van, *Information Retrieval*: Butterworth-Heinemann Newton, MA, USA, 1979.

[5] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 613--620, 1975.

[6] C. C. Aggarwal and P. S. Yu, "Finding generalized projected clusters in high dimensional spaces," *ACM SIGMOD Record*, vol. 29, pp. 70--81, 2000.

[7] O. Zamir and O. Etzioni, "Grouper: A dynamic clustering interface to Web search results," *COMPUT. NETWORKS*, vol. 31, pp. 1361--1374, 1999.

[8] M. A. Hearst and J. O. Pedersen, "Reexamining the cluster hypothesis: scatter/gather on retrieval results," *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 76--84, 1996.

[9] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*: Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.

[10] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," *New York*, 1990.

[11] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD Workshop on Text Mining*, vol. 34, pp. 35, 2000.

[12] K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for Web document clustering," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, pp. 1279--1296, 2004.

[13] O. Zamir and O. Etzioni, "Web document clustering: a feasibility demonstration," *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 46--54, 1998.

[14] O. Zamir, O. Etzioni, O. Madani, and R. M. Karp, "Fast and intuitive clustering of web documents," *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 287--290, 1997.

[15] T. Liu, Liu, Shengping, Chen, Zheng, Ma, and Wei-Ying, "An Evaluation on Feature Selection for Text Clustering," , 2003.

[16] L. Rigutini and M. Maggini, "A Semi-Supervised Document Clustering Algorithm Based on EM," *Proceedings of the The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)-Volume 00*, pp. 200--206, 2005.

[17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391-- 407, 1990.

[18] R. I. Arriaga and S. Vempala, "An algorithmic theory of learning: Robust concepts and random projection," *Machine Learning*, vol. 63, pp. 161--182, 2006.

[19] M. Maggini, L. Rigutini, and M. Turchi, "Pseudo-Supervised Clustering for Text Documents," *Proceedings of the Web Intelligence, IEEE/WIC/ACM International*

Conference on (WI'04)-Volume 00, pp. 363--369, 2004.

[20] A. Hotho, S. Staab, and G. Stumme, "Wordnet improves text document clustering," *Proc. of the SIGIR 2003 Semantic Web Workshop*, 2003.

[21] J. Sedding and D. Kazakov, "Wordnet-based text document clustering," , vol. 113: Geneva, 2004.

[22] Thorsten Brants. 2000. TnT–a statistical part-of speech tagger. In *Proceedings of the ANLP-NAACL*, Seattle, WA.

[23] Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceeding of EMNLP*, Philadelphia, PA.

[24] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia.

[25] Tjong Kim Sang, E. F., & Buchholz, S. (2000). Introduction to the CoNLL-2000 shared task. *Proceedings of CoNLL 2000 - LLL 2000* (pp. 127-132).

[26] Kudo, T., & Matsumoto, Y. (2001). Chunking with support vector machines. *Proceedings of NAACL2001* (pp. 1-8).

[27] Goldberg, Y., & Elhadad, M. (2009). On the role of lexical features in sequence labeling. *Proceedingsof EMNLP 2009* (pp. 1142-1151).

[28] Lee, Y.-S., & Wu, Y.-C. (2007). A robust multilingual portable phrase chunking system. *Expert Systemswith Applications*, *33*(3), 590-599.

[29] Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of NAACL2003* (pp. 134-141).

[30] Shen, H., & Sarkar, A. (2005). Voting Between Multiple Data Representations for Text Chunking. In*Advances in Artificial Intelligence* (Vol. 3501, pp.389-400). Springer Berlin / Heidelberg.

[31] Araujo, L., & Serrano, J. I. (2008). Highly accurate error-driven method for noun phrase detection.*Pattern Recognition Letters*, *29*(4), 547-557.

[32] Marie Guégan ,Claude de Loupy Knowledge-Poor Approach to Shallow Parsing: Contribution of Unsupervised Part-of-Speech Induction Proceedings of Recent Advances in Natural Language Processing,pages 33–40,Hissar, Bulgaria, 12-14 September 2011.

[33] Michael Collins "Natural Language Processing (Almost) from Scratch" Journal of Machine Learning Research 12 (2011) 2493-2537