IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 243-245

A Novel Document Representation Model for Clustering

Supreethi.K.P[1] **& E.V. Prasad**[2]

[1]Dept. of CSE, JNTUHCE, Hyderabad
[2]Dept. of CSE, UCE, JNTU, Kakinada.
Email: [1]supreethi.pujari@gmail.com, [2]drevprasad@yahoo.co.in

ABSTRACT

Text document plays an important role in providing better document retrieval, document browsing and text mining. Traditionally, clustering techniques do not consider the semantics relationships between words, such as synonymy and hypernymy. Existing clustering techniques are based on the syntactic structure of the document. To exploit semantic relationships, WordNet has been used to improve clustering results. However, WordNet-based clustering methods mostly rely on single-term analysis of text; they do not perform any phrase-based analysis. To address these issues, we derive the semantic structure of the document. Case grammar structures from the field of natural language processing, are used as semantic structure. These structures are used as document representation model and used for clustering. Semantic similarity measure is used to compare the documents' similarity. The experimental results show the effectiveness of semantic relationships for clustering. Quality of the cluster has been improved. Moreover, semantic structure improves the WordNet-based clustering method.

Keywords: Document Clustering, Case Grammar, Compositional Semantics, Semantic Similarity Measure, WordNet, Hypernym, Troponym, Entailment, Coordinate Terms

1. INTRODUCTION

In an effort to keep up with the tremendous growth of the information, many research projects were targeted on how to organize such information in a way that will make it easier for the end users to find the information they want efficiently and accurately [1].

Text mining shares many concepts with traditional data mining methods. Data mining includes many techniques that can unveil inherent structure in the underlying data. One of these techniques is clustering. When applied to textual data, clustering methods try to identify inherent grouping of the text documents so that a set of clusters is produced in which clusters exhibit high intra cluster similarity and low inter cluster similarity. Generally speaking, text document clustering methods attempt to segregate the documents into groups where each group represents some topic that is different than those topics represented by the other groups [2].

Text document clustering techniques were initially developed to improve precision and recall of information retrieval systems by effectively partitioning texts into previously unseen categories [6]. Documents contain paragraphs, paragraphs contain sentences and sentences contain words. Most of the documents clustering methods that are in use today are based on the Vector Space Model that represents documents as a feature vector of words that appear in all the document set. A set of measures are used for pair wise document similarity by these clustering methods .These clustering methods are well suited for the web sites and search engines, which are keyword based.

Keyword-based search engines such as Google and Bing are the main tools for use of today's web. It is clear that the huge success of the web is due to search engines. These search engines take the user's requirement, converts it into a list of words and displays the links to the relevant documents. A document is said to be relevant, if the words in the users' query match with the document words.

Document content means not only the syntax, but also the relationship among the words called semantics [3]. This relationship depicts the meaning of the sentences. The existing document clustering techniques have concentrated more or less the syntax of the sentence, rather than the semantics. Key word based document clustering techniques are not suitable for the Semantic web site searching process [8]. These issues have motivated for the current work discussed in this paper.

2. PROPOSED METHOD

A novel document representation model has been proposed. This model considers the semantic structure of the document for clustering. Document semantics is derived from the syntax, and Semantic structure viz. case grammar structure is derived from the syntactic structure. Case grammar structure is a type of semantic grammar used in natural language processing.

Document is parsed to fetch Parts Of Speech (POS) tagging for the sentences in the document. Stanford parser [10] is used for generating the syntactic structure i.e., POS tagged text. This POS tagged text is processed further, stop words are removed and stemming is performed. Standard stop words contain sometimes, some words which are essential to retain the meaning of the sentences. For eg., the word 'by' a standard stop word many a times necessary to depict the voice of the verb. Hence the authors have created their own stop word list.

Stemming is the process of dropping the suffixes and prefixes of a word to get the root word. Standard stemmers many a times perform stemming by losing the meaning of the word. To retain the meaning the existing rules are useless, hence authors have framed stemming rules exclusively for the verb phrases. Development of case grammar structure follows stemming. Case grammar structure is a triplet, giving weight to the verb of the sentence and considering the subject and object of the sentence i.e verb (subject, object). Authors have considered only the sentences with compositional semantics and disambiguity. Compositional semantics does not consider metaphors or idioms. For eg., the sentence '*The old man kicked the bucket*' does not have compositional semantics, the words collectively do not give the meaning of *death*.

The synsets for the words in the case grammar structures are gathered using the WordNet lexical database [9]. Most synsets are connected to other synsets via a number of semantic relations. These relations for the verb include hypernyms, troponyms, entailments, and coordinate terms. A verb Y is a *hypernym* of the verb X if the activity X is a kind of Y. A verb Y is a *troponym* of the Verb X if the activity Y is doing X in some manner. A verb Y is *entailed* by X if by doing X you must be doing Y. *Coordinate terms* are those verbs sharing a common hypernym. All these synsets for the verbs are fetched and are used for semantic similarity measure [4] calculation. Highly similar documents are grouped into clusters. Similarity among the documents is measured using the equation

$$\text{Sim}(D_1, D_2) = [\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n}\text{Sim}(W_{1,i}, W_{2,j})]/\text{mn}$$

Authors have used Leader algorithm for clustering. Leader algorithm is an incremental partition based algorithm. A document is identified as a leader for every cluster. New document will be compared with the leader, if semantic similarity is more than the threshold then the document will be thrown into that leader's cluster, otherwise the document will be compared with other leader. If the document does not follow any leader it emerges as a leader and increments the cluster count.

3. EXPERIMENT RESULTS

The effectiveness of semantic structure based document representation model has been proved by conducting experiments using WordNet, the lexical data base and compared with vector space model [7], phrase based model [2]. The experiments were performed on J2SE 5.0, Windows XP, Pentium 4, 3.0 GHz CPU with 2 GB RAM.

3.1. Experimental Set Up

To ensure the experimental results are independent of one special test collection, we used three collections to test our proposed method. They are 20-newsgroup, Reuters-transcribed-set and Reuters-21578 test collection. These are available from the UCI KDD archive [11]. Leader algorithm has been executed considering three different document representation models.

3.2. Evaluation Measures

To prove the superiority of the semantic based structure, we have considered the information retrieval measures for evaluation. Our model is suited for semantic web search process, so we have taken precision and recall measures. The measures are defined as

Precision = (Relevant and Retrieved) / Retrieved

Recall = (Relevant and Retrieved) / Relevant

Both measures have the interval [0, 1] as a range. Both the measures need to be maximized to satisfy the users of the semantic web site.

3.3. Results and Discussion

The results are shown in figs.1 and 2. For 20-newsgroup, the precision of these methods ranges from 0.613 to 0.985; the recall of these methods ranges from 0.726 to 0.831. For Reuters-transcribed-set, the precision of these methods ranges from 0.345 to 0.587; the recall of these methods ranges from 0.436 to 0.731. For Reuters-21578, the precision of these methods ranges from 0.515 to 0.698; the recall of these methods ranges from 0.832 to 0.906.

The table 1 lists the values for the evaluation measures of our experiments. Our proposed model outperforms the vector space model and phrase based representation model.

Table 1
Results of the Experiments on Vector Space Model, Phrase Based and Semantic Based Models

	Precision			*Recall*		
	VSM	*PBM*	*SBM*	*VSM*	*PBM*	*SBM*
20-newsgroup	0.613	0.785	0.985	0.726	0.798	0.831
Reuters-transcribed-set	0.345	0.472	0.587	0.436	0.625	0.731
Reuters-21578	0.515	0.591	0.698	0.832	0.864	0.902

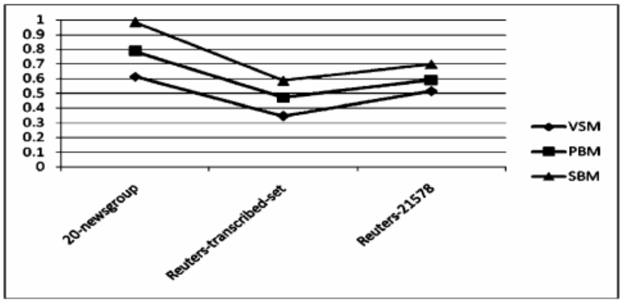

Fig.1: Precision Measure

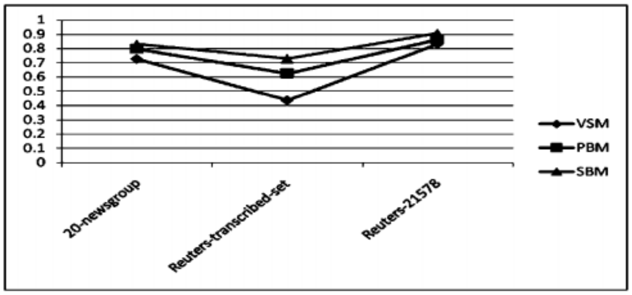

Fig. 2: Recall Measure

4. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a novel document representation model for clustering. Documents are clustered based on their meaning. We have considered only the sentences with compositional semantics and without ambiguity. We will conduct further research to improve our work, by incorporating more NLP techniques to deal with idioms, metaphors and ambiguity. The work has considered only .txt and .doc files, it can also be extended to html, xml files i.e., web document clustering can be done.

REFERENCES

[1] A.K.Jain, "Data Clustering: 50 Years Beyond K-Means", *Proc. Int'l Conf Pattern Recognition, Pattern Recognition Letters*, **31**, Issue 8, pp.651-666, June 2010.

[2] K.M.Hammouda and Mohamed S.Kamel, "Efficient Phrase based Document Indexing for Web Document Clustering", *IEEE Transactions on Knowledge and Data Engineering*, **16**, No.10, pp1279-1296, 2004.

[3] Tanveer Siddiqui, U.S.Tiwary, *Natural Language Processing and Information Retrieval*, Oxford University Press, 2008.

[4] B.Danushka, M.Yutaka and I.Mitsuru, "Measuring Semantic Similarity between Words using Web Search Engines," *Proc. 16th WWW*, pp 757-766, 2007.

[5] O.Zamir and O.Etzioni, " Web Document Clustering: A Feasibility Demonstration", *21st Annual Int'l ACM SIGIR Conference*, Melbourne, Australia, pp.46-54, 1998.

[6] M.Steinbach, G.Karypis and V.Kumar, " A Comparison of Document Clustering Techniques," *Proc. KDD-2000 Workshop Text Mining*, Aug. 2000.

[7] Ruiqiang Guo, Fuji Ren, "Towards the Relationship between Semantic Web and NLP," *Proc. Int'l Conf. Natural Knowledge Processing and Knowledge Engineering*, 2009.

[8] J.Sedding, D.Kazakov, "WordNet-based Text Document Clustering," *COLING 2004 Third Workshop on Robust Methods in Analysis of Natural Language Data*, Geneva, Switzerland, 29th Aug.2004.

[9] The Stanford Parser, nlp.stanford.edu/software/lex-parser.shtml.

[10] UCIKDD ARCHIVE, kdd.ics.uci.edu