International Journal of Information Processing, 9(2), 47-56, 2015
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

# Hybrid Clustering Model for Text Documents with Semantic Based Document Representation

B Drakshayani[a], E V Prasad [b]

[a]Government Polytechnic, Nalgonda, Andhra Pradesh, India

[b]Lakkireddy Balireddy College of Engineering, Vijayawada, Andhra Pradesh, India

Recently, large documents are organized into clusters, as clustering in text document has a major role in perceptive navigation and browsing. The existing methods performances well, if the documents do not contain any idioms/metaphors and semantics. In this paper, a model is proposed by considering idioms/metaphors based on semantics. It follows replacement of idiom/metaphors with their original meaning, tagging and assignment of semantic weights to the document words. The similarity measure is obtained between the documents and then the documents are clustered using a hybrid clustering algorithm. The performance of the proposed model is evaluated on different data sets. The results are analyzed using standard parameters such as: precision, recall, F-measure, purity and entropy. These parameters are compared with the Vector Space Model (VSM) and Phase Base Method (PBM). Finally, the results confess that the proposed model outperforms the existing methods.

**Keywords :** Chameleon Algorithm, Document Clustering, Hybrid Clustering, Idiom, Leaders Algorithm, Metaphor, Natural Language Processing, Semantic weight, Tagging.

## 1. INTRODUCTION

Due to the enormous usage of Internet, the information resources are also increasing and available in terms of text. The text data is un-structured, hence very difficult to process them. Researchers has presented knowledge discovery in text system, which uses the simplest information extraction to get interesting information and knowledge from unstructured text collection. So, there are lot of text mining techniques are designed to understand the information from the Internet. All the existing approaches convert the text documents into simplistic intermediate forms consists of term vectors and keywords. The usage of individual terms in simplistic representations, they lose their semantic relations and original meaning. Now a day's, in all the existing search engines, documents are clustered with the help of similarity score. However, search engines do not organize documents automatically; they just retrieve related documents to a certain query. The clustering techniques improve the performance when a query is submitted to the system and provide support for scalability, complex data types and shapes.

Document clustering is a fundamental task in text mining that is concerned with organizing documents into clusters according to their content and to visualize the collection, providing an overview of the range of documents and of their relationships, so that they can be browsed more easily. Different aspects of similarity between documents can be defined. Document clustering can be investigated and for use in a number of different areas of text mining and information retrieval [1-6]. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems and as an efficient way of finding the nearest neighbors of a document. More recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to a users query. Text document clustering techniques plays a crucial role in text

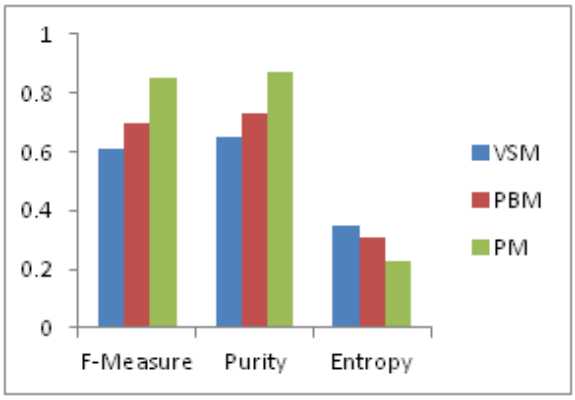


Figure 4. Evaluation of Clustering Algorithm on Reuters Transcribed Dataset

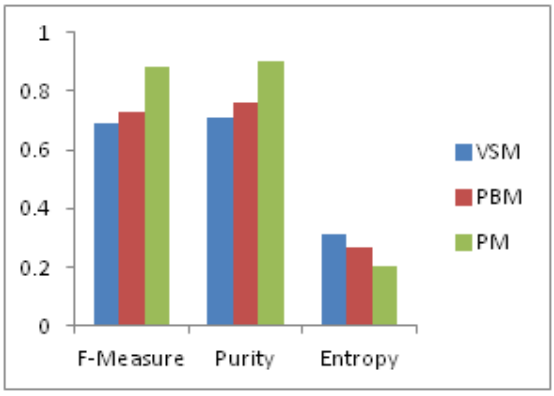


Figure 5. Evaluation of Clustering Algorithm on Reuters-21578 Dataset

as follows:

$$Purity(C, L) = \sum_{c \in C} \frac{|C|}{|D|} \max_{l \in L} P(c, l) \quad (8)$$

The entropy of each cluster c is measured as:

$$E(c) = -\sum_{l \in L} P(c, l). \log(c, l) \quad (9)$$

Purity measures the purity of the resulting clusters when evaluated against a pre-categorization, where entropy evaluates the homogeneous of a cluster. The objective of the proposed model is to maximize the purity and minimize the entropy of clusters to achieve high quality clustering. These values dictate the clusters quality.

### 4.3. Results and Discussion

The clusters are formed using existing leader algorithm on Reuters Test, Reuters-21578 and20 news Groups. The results obtained in this model are compared with the Vector space model and Phrase based clustering model. The quality measures of clustering results are shown in Figures 4 to 6. The methods have been executed on the three data sets. The cluster quality measures Precision (P), Recall (R), Purity (Pu), Entropy (En) and F-measure (Fm), for the three data sets values are reported in Table 1. For the three datasets vector space model and phrase based clustering model has found to exhibit poorly and the proposed clustering model has obtained best indices. A more interesting observation is that purity and entropy values indicate better clusters for 20 news groups datasets than Reuters Test, Reuters-21578 datasets. This quality improvement is done because of the idiom/metaphor processing, POS tagging, semantics and hybrid clustering. The proposed Model outer forms the VSM and PBM in terms of $F$-Measure, Purity and Entropy. The results reveal that idiom/metaphors, semantics and hybrid clustering plays a crucial factor in precisely judging the relation between documents.

## 5. CONCLUSIONS

In this paper, an idiom/metaphor semantic based mining model with hybrid clustering

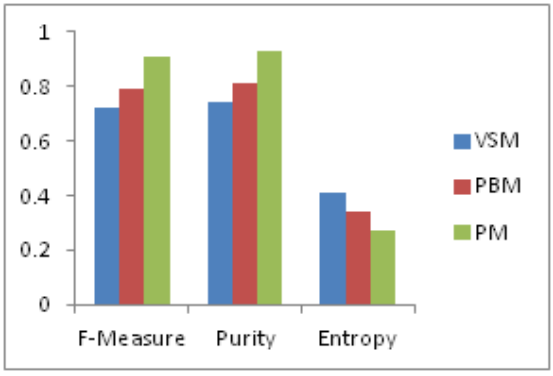


Figure 6. Evaluation of Clustering Algorithm on 20 News Groups Dataset

Table 1
Multi-column table

| Database/Model | Vector Space Model | | | Phrase Based Model | | | Proposed Model | | |
|---|---|---|---|---|---|---|---|---|---|
| | $P_u$ | $F_m$ | $E_n$ | $P_u$ | $F_m$ | $E_n$ | $P_u$ | $F_m$ | $E_n$ |
| Reuters Test | 0.65 | 0.61 | 0.35 | 0.73 | 0.70 | 0.31 | 0.87 | 0.85 | 0.23 |
| Reuters-21578 | 0.71 | 0.69 | 0.31 | 0.76 | 0.73 | 0.27 | 0.90 | 0.88 | 0.20 |
| 20News Groups | 0.74 | 0.72 | 0.41 | 0.81 | 0.79 | 0.34 | 0.93 | 0.91 | 0.27 |

for enhancing text document clustering is proposed. The main finding of this work is that documents are clustered based on their meaning using idiom/metaphor processing, semantic weight and similarity measured followed by hybrid clustering algorithm. We have considered the sentences with compositional semantics and NLP techniques to deal with idioms/metaphors. The performance of the proposed model is evaluated on different data sets. The results are analyzed using standard parameters such as: precision, recall, F-measure, purity and entropy. These parameters are compared with the Vector Space Model(VSM) and Phase Base Method (PBM). The results confess that the proposed model gives meaningful clusters than the existing methods.

## REFERENCES


1. K P Supreethi and E V Prasad. Web Document Clustering using Case Grammar Structure, *International Conference on Computational Intelligence and Multimedia Applications*, 2:98–102, Dec 2007.
2. M Rafi, M Manjood, M M Fazal and S M Ali. A Comparison of Two Suffix Tree based Document Clustering Algorithms, *In Information and Emerging Technologies*, Page 1–5, 2010.
3. Y S Lin, J Y Jiang and S J Lee. A Similarity Measure for Text Classification and Clustering, *In IEEE Transactions on Knowledge and Data Engineering*, 26(7):1575–1590, 2014.
4. D Bollegala, Y Matsuo and M Ishijuka. A Web Search Engine Based Approach to Measure Semantic Similarity between Words, *In IEEE Transactions on Knowledge and Data Engineering*, 23(27):977–990, 2011.
5. S Shehata, F Karray and M S Kamel. An Efficient Concept-Based Mining Model for Enhancing Text Clustering, *In IEEE Transactions on Data and Knowledge Engineering*, 22(10):1360–1371, 2010.
6. Z Elberrichi and M Simonet A Amine. Evaluation of Text Clustering Methods using WordNet, *International Arab Journal of Information Technology*, 7(4), Oct 2010.
7. K Shutova. Automatic Metaphor Interpretation as a Paraphrasing Task Human Language Technologies, in *Annual Conference of the North American Chapter of the ACL, California*, pages 1029–1037, 2010.
8. David Holmes. Idioms and Expressions, a Method for Learning and Remembering Idioms and Expressions.
9. U S Tiwari and T Siddiqui. Natural Language Processing and Information Retrieval, *In Oxford University Press*, August 2008.
10. R Guo and F Ren. Towards the Relationship between Semantic Web and NLP, *International Conference Natural Knowledge Processing and Knowledge Engineering*, 2009.
11. H T Zheng, B Y Kang and H G Kim. Exploiting Noun Phrases and Semantic Relationships for Text Document Clustering, *Elsevier journal of Information Science*, pages 2249–2262, February 2009.
12. A Hotho, S Staab and G Stumme. Wordnet Improves Text Document Clustering, *SIGIR Semantic Web Workshop*, pages 541–544, 2003.
13. G A Miller. WordNet: A Lexical Database for English, *ACM Communications* , 38(11):39–41, 1995.
14. Xu Rui and Donald Wunsch. Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks,* 16(3):645–678, May 2005.
15. A K Jain. Data Clustering: 50 Years Beyond K-Means, in *International Conference in Pattern recognition*, pages 651–666, 2010.
16. D A Keim and A Hinneburg. Clustering Techniques for the Large datasets- from the Past to the Future, *Tutorial Notes for ACM SIGKDD*, pages 141–181, Aug. 1999.

17. I Cheng-Ru and M S Chen. Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging, *IEEE transactions on Knowledge and Data Engineering*, 17(2):145–159, February 2005.
18. S Dumais and S T L Deerwester. Indexing by Latent Semantic Analysis, *Journal of the Society for Information Science*, pages 391–407, 1990.
19. H Chim and X Deng. Efficient Phrase based Document Similarity for Clustering, *IEEE Transactions on Knowledge and Data Engineering*, 20(9):1217–1229, September 2008.
20. W K God and M S Kamel. PH-SSBM: Phrase Semantic Similarity Based Model for Document Clustering, *IEEE Second International Symposium on Knowledge Acquisition and Modeling*, 2:197–200, Dec 2009.
21. K M Hammouda and M S Kamel. Efficient Phrase based Document Indexing for Web Document Clustering, *IEEE Transaction on Knowledge and Data Engineering*, 16(10):1279–1296, 2004.
22. A Hotho, S Staab and G Stumme. Wordnet Improve Text Document Clustering, *in SIGIR 2003 Semantic Web Workshop*, pages 541–544, 2003.
23. A Hotho, S Staab and G Stumme. Text Clustering based on Background Knowledge, *Technical Report No. 425*.
24. UCIKDD ARCHIEVE, kdd.ics.uci.edu
25. B Drakshayani and E V Prasad. Comparison of Single level, Multilevel and Hybrid Clustering Methods for Text Documents, *IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions*, IIT Kanpur, India, pages 214–218, July 2013.
26. B Drakshayani and E V Prasad. Metaphor based Document Representation Model for Text Document Clustering, *IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions*, IIT Kanpur, India, pages 74–78, July 2013.
27. B Drakshayani and E V Prasad. Semantic Based Model Text Document Clustering with Idioms, *International Journal of Data Engineering*, 4(1):1–13, 2013.
28. The Stanford Parser, nlp.stanford.edu/software/lex-parser.shtml.
29. W Gad and M Kamel. New Semantic Similarity based Model for Text Clustering using Extended Gloss Overlaps, *in International Conference on Machine Learning and Data Mining (MLDM)*, pages 663–677, July 2009.
30. B Danushka, M Yutaka and I mitsuru. Measuring Semantic Similarity between Words using Web Search Engines, *16th WWW*, pages 757–766, 2007.
31. M Srinivas and C K Mohan. Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods, *International Joint Conference on Neural Networks*, pages 1–7, July 2010.

**E V Prasad** has thirty six years of academic experience in technical education and is currently the Director, Lakireddy Bali Reddy College of Engineering, Mylavaram, Krishna District, Andhra Pradesh. Earlier to the current assignment he was the Rector, Jawaharlal Nehru Technological University (JNTU), Kakinada, Registrar and Director of Institute of Science and Technology, JNTU Kakinada and distinguished himself as Principal and Vice-Principal of University College of Engineering (Autonomous), JNTU, Kakinada.

**B Drakshayani** has 11 years of teaching experience in various engineering colleges and is currently working as Lecturer, Govt. Polytechnic, Nalgonda. Earlier she worked as Head of Deprtment in Engineering College. She did her BTech and MTech through JNTU, Kakinada, Andhrapradesh and currently submitted her Ph.D thesis to JNTU, Kakinada.