Abstract

Due to daily quick growth of the information, there are considerable needs to extract and discover valuable knowledge from data sources such as World Wide Web. The common methods in text mining are mainly based on statistical analysis of term either phrase or word. These methods consider documents as bags of words and they will not give any importance to meanings of document content. In addition, statistical analysis of term frequency extracts the significance of term within a document only. Whenever any 2 terms might have same frequency in their documents, but only 1 term pays more to meaning of its sentences than other term. The concept-based model that analyses terms on corpus, document and sentence levels instead of ancient analysis of document is introduced. The planned model consists of, concept-based analysis, clustering by using k-means, concept-based similarity measure.

Term that contributes to sentence meaning is assigned with 2 dissimilar weights by concept-based statistical analyzer. These 2 weights are united into new weight. Concept-based similarity is used for computing similarity among documents. The concept based similarity
method takes full benefit of using concept analysis measures on the corpus, document, and sentence levels in computing the similarity among documents. By using k-means algorithm experiments are done on concept based model on different datasets in text clustering. The experiments are done by comparing the concept-based weight obtained by concept-based model and statistical weight. The results in text clustering show the significant progress of clustering feature using: concept-based term frequency (tf), conceptual term frequency (ctf), concept-based statistical analyzer, and concept-based combined model. In text clustering the results are evaluated using f-measure and entropy.

References

1. [Berry Michael W., (2004), “Automatic Discovery of Similar Words”, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, 24-43
2. Navathe, Shamkant B., and ElmasriRamez, (2000), “Data Warehousing and Data Mining”, in “Fundamentals of Database Systems”, Pearson Education pvtInc, singapore, 841-872.
3. HaralamposKaranikas and BabisTheodoulidis Manchester, (2001), “Knowledge Discovery in Text and Text Mining Software”, Centre for Research in Information Management, UK
4. https://en.wikipedia.org/wiki/Concept_mining
5. P. Kingsbury and M. Palmer, “Propbank: The Next Level of Treebank,” Proc. Workshop Treebanks and Lexical Theories, 2003.
6. G. Salton and C. Buckley. Term Weighting Approaches in AutomaticText Retrieval, 1960, Information Processing and Management, 24, Vol5, 513-52
7. G. Salton and C. Buckley. Term Weighting Approaches in AutomaticText Retrieval, 1960, Information Processing and Management, 24, Vol 5, 513-523
8. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th VLDB Conference, 1994
9. Agrawal R, Imielinski T, Swami A, “Mining association rules between sets of items in large databases”. Proc of the 1993ACM SIGMODInternational Conference on Management of data
10. Bing Liu, Yiming Ma, “Discovering unexpected information from your competitors’ Web Sites in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, USA.
11. An Efficient Concept-Based Mining Model for Enhancing Text Clustering, Shady Shehata, Member, IEEE, FakhriKarray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE 2010
12. Concept mining from natural language texts, Rockai V. Dept. of Cyber. & Artificial Intelligent, Tech. Univ. of Kosice, Kosice, Slovakia Mach. M IEEE 2012
13. Concept Mining using Association Rules and Combinatorial Topology Sutojo, A, San Jose State University, San Jose IEEE 2007
14. Webpage Clustering and Concept Mining, an Approach to Intelligent Information Retrieval. Fang Li, Martin Mehlitz, Li Feng, Huanye Sheng, DEPT of CSE, Shanghai Jiaotong University, Shanghai ,China IEEE 2006
15. A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs: Prentice Hall, 1988.
16. L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, ser. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley &
Sons Inc., 1990.

17. K. J. Cios, W. Pedrycz, and R. W. Swiniarski, Data mining methods for knowledge discovery," IEEE Transactions on Neural Networks, vol. 9, no. 6, pp. 1533{1534, 1998.

18. B. V. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, 1991.

19. D. R. Hill, A vector clustering technique, “in FID-IFIP, Samuelson(ed), N-H 1968, 1967.

20. A survey paper on Concept Mining in Text documents K.n.s.s.v.prasad, ,Dr.S.K.Saritha, ,Dixa saxena. International Journal of Computer Applications (0975 – 8887)Volume 166 – No.11, May 2017

Index Terms

Computer Science
Data Mining

Keywords

Concept Mining