{tag} {/tag} International Journal of Computer Applications

Number 5 - Article 2

Year of Publication: 2010

Authors:

Rekha Baghel

DR.Renu Dhir

10.5120/826-1171

{bibtex}pxc3871171.bib{/bibtex}

Abstract

This paper presents a novel technique of document clustering based on frequent concepts. The proposed technique, FCDC (Frequent Concepts based document clustering), a clustering algorithm works with frequent concepts rather than frequent items used in traditional text mining techniques. Many well known clustering algorithms deal with documents as bag of words and ignore the important relationships between words like synonyms. the proposed FCDC algorithm utilizes the semantic relationship between words to create concepts. It exploits the WordNet ontology in turn to create low dimensional feature vector which allows us to develop a efficient clustering algorithm. It uses a hierarchical approach to cluster text documents having common concepts. FCDC found more accurate, scalable and effective when compared with existing clustering algorithms like Bisecting K-means , UPGMA and FIHC.

Reference

- J. Han and M. Kimber. 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.

- Jain, A.K, Murty, M.N., and Flynn P.J. 1999. Data clustering: a review. ACM Computing Surveys, pp. 31, 3, 264-323.
- M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. KDD Workshop on Text Mining'00.
- P. Berkhin. 2004. Survey of clustering data mining techniques [Online]. Available: http://www.accrue.com/products/rp_cluster_review.pdf.
- Xu Rui. 2005. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3):pp. 634-678.
- Miller G. 1995. Wordnet: A lexical database for English. CACM, 38(11), pp. 39–41.
- L. Zhuang, and H. Dai. 2004. A Maximal Frequent Itemset Approach for Document Clustering. Computer and Information Technology, CIT. The Fourth International Conference, pp. 970 – 977.
- R. C. Dubes and A. K. Jain. 1998. Algorithms for Clustering Data. Prentice Hall college Div, Englewood Cliffs, NJ, March.
- D. Koller and M. Sahami. 1997. Hierarchically classifying documents using very few words. In Proceedings of (ICML) 97, 14th International Conference on Machine Learning, pp. 170–178, Nashville, US.
- B.C.M.Fung, K.Wan, M.Ester. 2003. Hierarchical Document Clustering Using Frequent Itemsets", SDM'03.
- Green, S. J. 1999. Building hypertext links by computing semantic similarity. T KDE, 11(5), pp. 50–57.
- Sedding, J., & Kazakov, D. 2004. Wordnet-based text document clustering. 3rd Workshop on Robust Methods in Analysis of Natural Language Data, pp. 104–113.
- Y. LI, and S.M. Chung. 2005. Text Document Clustering Based on Frequent Word Sequences. In Proceedings of the. CIKM, 2005. Bremen, Germany, October 31- November 5.
- Zheng, Kang, Kim. 2009. Exploiting noun phrases and semantic relationships for text document clustering. Information Science 179 pp. 2249-2262.
- Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors,Proc 20th Int. Conf. Very Large Data Bases, VLDB, pp. 487–499.
- Agrawal, T. Imielinski, and A. N. Swami. 1993. Mining association rules between sets of items in large databases. In Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD93), pp. 207–216, Washington, D.C.
- Stanford Tagger available at http://nlp.stanford.edu/software/tagger.shtml
- K.M. Hammouda, M.S. Kamel. 2004. Document similarity using a phrase indexing graph model. Knowl. Inform. Syst. 6 (6) 710–727.
- StopWord List, http://www.lextek.com/manuals/onix/stopwords2.html
- Cognitive Science Laboratory at Princeton University Available at: http://www.cogsci. princeton .edu/.
- L. Kaufman and P. J. Rousseeuw. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons.
- G.Karypis. 2002. Cluto 2.0 clustering toolkit. http://wwwusers.cs.umn.edu/˜ karypis/cluto
- Classic. ftp://ftp.cs.cornell.edu/pub/smart/.

- E. H. Han, B. Boley, M. Gini, R. Gross, K. Hastings,G. Karypis, V. Kumar,B. Mobasher, and J. Moore. 1998. Webace: a web agent for document categorization and Exploration. In Proceedings of the second international conference on Autonomous agents, pp. 408–415. ACM Press.
- D.D.Lewis.Reuters. http://www.research.att.com/˜ lewis/.
- Crowe,M. 2000. Wordnet.net library. http://www.opensvn.csie.org/WordNetDotNet/
- M. F. Porter. 1980. An algorithm for suffix stripping. Program; automated library and information systems, 14(3), pp.130-137.
- CherryPicker Coreference resolution Tool. Available at http://www.hlt.utdallas.edu/~altaf/cherrypicker.html

Index Terms

Computer Science Data Mining

Key words

Document clustering Clustering algorithm
Frequent Concepts based Clustering

WordNet