Abstract

Importance of document clustering is now widely acknowledged by researchers for better management, smart navigation, efficient filtering, and concise summarization of large collection of documents like World Wide Web (WWW). The next challenge lies in semantically performing clustering based on the semantic contents of the document. The problem of document clustering has two main components: (1) to represent the document in such a form that inherently captures semantics of the text. This may also help to reduce dimensionality of the document, and (2) to define a similarity measure based on the semantic representation such that it assigns higher numerical values to document pairs which have higher semantic relationship. Feature space of the documents can be very challenging for document clustering. A document may contain multiple topics, it may contain a large set of class-independent general-words, and a handful class-specific core-words. With these features in mind, traditional
agglomerative clustering algorithms, which are based on either Document Vector model (DVM) or Suffix Tree model (STC), are less efficient in producing results with high cluster quality. This paper introduces a new approach for document clustering based on the Topic Map representation of the documents. The document is being transformed into a compact form. A similarity measure is proposed based upon the inferred information through topic maps data and structures. The suggested method is implemented using agglomerative hierarchal clustering and tested on standard Information retrieval (IR) datasets. The comparative experiment reveals that the proposed approach is effective in improving the cluster quality.

Reference

- Jain, A. K., Murty, M. N., and Flynn, P. J., "Data Clustering: a review," ACM Computing Survey, pp. 264-323, 1999.
- Campi, A. and Ronchi, S., "The Role of Clustering in Search Computing ," in 20th International Workshop on Databases and Expert Systems Application , Linz, Austria, 2009, pp. 432-436.
- Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W., "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," in Fifteenth Annual International ACM SIGIR Conference, June 1992, pp. 318-329.
- Hearst, M. A. and Pedersen, J. O., "Reexamining the cluster hypothesis: scatter/gather on retrieval results," in 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, Switzerland , 1996, pp. 74-84.
- Hammouda, K.M. and Kamel, M.S. , "Efficient Phrase-Based Document Indexing for Web Document Clustering," IEEE Transaction on Knowledge and Data Engineering, vol. 16, no. 10, pp. 1279-1296, 2004.
- Hung, C. and Xiaotie, D., "Efficient Phrase-Based Document Similarity for Clustering," IEEE Transaction on Knowledge and Data Engineering, vol. 20, no. September, pp. 1217-1229, 2008.
- Fung, B.C.M., Wang, K., and Ester, M., "Hierarchical document clustering using frequent Itemsets," Proceedings of SIAM International Conference on Data Mining, 2003.
- Soon, M. C. , John, D. H., and Yanjun, L., "Text document clustering based on frequent word meaning sequences," Data & Knowledge Engineering, vol. 64, pp. 381-404, 2008.
- Pepper, S., “Topic Maps,” Encyclopedia of Library and Information Sciences, Third Edition 2010
- Pepper, S.; Moore, G., Eds. XML Topic Maps (XTM) 1.0; TopicMaps.Org 2001, http://www.topicmaps.org/xtm/1.0/
- Maicher, L.; Garshol, L.M.; Eds. Scaling Topic Maps; In Third International Conference on Topic Maps Research and Applications, TMRA 2007, Leipzig, Germany, October 2007; Springer-Verlag: Berlin, Heidelberg
- Steianbach, M., Karypis, G. , and Kumar, V., “A comparison of document clustering techniques,” in KDD-Workshop on Text Mining , 2000.
- Rafi, M.,Maujood, M. ,Fazal, M. M., Ali, ,M.,“Acomparison of two suffix tree-based document clustering algorithms", The Second IEEE ICIE,14-16 June Karachi, Pakistan, 2010.
- Document on Wandora Implementation and Usage. Can be found at http://www.wandora.org/wandora/wiki/index.php?title=Documentation
Document Clustering based on Topic Maps

Index Terms

Computer Science  Information Retrieval

Key words

Text Document  Document Clustering  Algorithm

Performance measure