Abstract

Text mining has becoming an emerging research area now-a-days which helps in extracting the useful information from large amount of natural language text documents. The necessity of grouping the documents for different applications is gaining comprehensive review of the techniques used to improve the efficient time consumption, challenges, research issues are presented. The techniques presented in the review are k-means clustering, fuzzy c means clustering, support vector machine classifiers, naive Bayes classifier, Hidden Markov Model (HMM). Furthermore, discussion of the advantages and disadvantages of each technique is contributed to a better understanding and compared with the existing techniques based on the efficiency and computational time.

References

1. Gupta, Vishal, and Gurpreet S. Lehal. "A survey of text mining techniques and applications." Journal of emerging technologies in web intelligence 1, no. 1 (2009): 60-76
2. Feldman, Ronen, and Ido Dagan. "Knowledge Discovery in Textual Databases (KDT)." In KDD, vol. 95, pp. 112-117. 1995.
3. Sundari, D. Jasmine Guna, and D. Sundar. "A Study of Various Text Mining Techniques."
4. Ghosh, Sayantani, Sudipta Roy, and Samir K. Bandyopadhyay. "A tutorial review on Text Mining Algorithms." International Journal of Advanced Research in Computer and Communication Engineering 1, no. 4 (2012): 7.
5. Bisht, Sunita, and Amit Paul. "Document clustering: a review." International Journal of Computer Applications 73, no. 11 (2013).
6. Lian, Wang, Nikos Mamoulis, and Siu-Ming Yiu. "An efficient and scalable algorithm for clustering XML documents by structure." IEEE transactions on Knowledge and Data Engineering 16, no. 1 (2004): 82-96.
7. https://en.wikipedia.org/wiki/Document_Clustering
8. Shah, Neepa, and Sunita Mahajan. "Document clustering: a detailed review." International Journal of Applied Information Systems 4, no. 5 (2012): 30-38.
9. Chaurasia, Vikas, and Saurabh Pal. "A novel approach for breast cancer detection using data mining techniques." (2017).
10. Shen, Liyin, Hang Yan, Hongqin Fan, Ya Wu, and Yu Zhang. "An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design." Building and Environment 124 (2017): 388-401.
11. Turtle, Howard, and W. Bruce Croft. "Inference networks for document retrieval." In ACM SIGIR Forum, vol. 51, no. 2, pp. 124-147. ACM, 2017.
12. Desjardins, Guy, and Robert Godin. "Combining relevance feedback and genetic algorithms in an internet information filtering engine." In Content-Based Multimedia Information Access-Volume 2, pp. 1676-1685. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.
13. Forsati, Rana, Andisheh Keikha, and Mehrnoush Shamsfard. "An improved bee colony optimization algorithm with an application to document clustering." Neurocomputing 159 (2015): 9-26.
14. Lučić, Panta, and Dušan Teodorović. "Computing with bees: attacking complex transportation engineering problems." International Journal on Artificial Intelligence Tools 12, no. 03 (2003): 375-394.
15. Ludwig, Simone A. "MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability." International journal of machine learning and cybernetics 6, no. 6 (2015): 923-934.
16. Dumais, Susan, John Platt, David Heckerman, and Mehran Sahami. "Inductive learning algorithms and representations for text categorization." In Proceedings of the seventh international conference on Information and knowledge management, pp. 148-155. ACM, 1998.
17. Aggarwal, C. C., and C. Zhai. "Probabilistic Models for Text Mining: In Mining Text Data." (2012): 257-294.
18. Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." In KDD workshop on text mining, vol. 400, no. 1, pp. 525-526. 2000.
19. Yu, Hwanjo, and Sungchul Kim. "SVM tutorial—classification, regression and ranking." In Handbook of Natural computing, pp. 479-506. Springer, Berlin, Heidelberg, 2012.
20. Teh, Yee W., Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Sharing clusters among related groups: Hierarchical Dirichlet processes." In Advances in neural information processing systems, pp. 1385-1392. 2005.

Index Terms

Computer Science  Information Systems

Keywords

Clustering, text mining, k-means clustering