{tag} {/tag}

IJCA Special Issue on International Conference on Communication, Computing and Information Technology

ICCCMIT - Number 2

Year of Publication: 2013

Authors:

S. Poonkuzhali

P. Sudhakar

K. Sarukesi

{bibtex}icccmit1021.bib{/bibtex}

Abstract

Web outlier mining is dedicated for finding web pages which differ significantly from the rest of the web document taken from the same category. Most of the existing algorithms for web content outlier mining is developed for structured documents, whereas WWW contains mostly unstructured and semi structured documents. Moreover, the false positive rate in the existing algorithms for mining web content outlier is more than 30%. Therefore, there is need to develop

a technique to mine web outliers from unstructured and semi structured document types with less false positive rate. This paper, concentrates on mining web content outliers which extracts the dissimilar web document taken from the group of documents of same domain. The proposed work implement a novel mathematical approach based on signed-with-weight technique for mining web content outliers which retrieves top n outlier web documents from both structured and unstructured web documents. The proven results show the performance measure of this approach in terms of precision and recall is more than 90%. Also, the false positive rate of this algorithm is less than 15%.

References

- Ali S. Hadi,A. H. M. Rahmatullah Imon(2009), Mark Werner, Detection of outliers Overview, Wiley Interdisciplinary Reviews: Computational Statistics, Volume 1, Issue 1, pp-57-70.
- Anguilli, F. , and Pizzuti, C. , Elomaa,T. (Eds.). Fast Outlier Detection in High Dimensional Spaces. PKDD, LNAI 2431, 2002, pp 15-27
- Bing Liu, Kevin Chen- Chuan Chang , Editorial: Special issue on Web Content Mining , SIGKDD Explorations, Volume 6, Issue 2.
- Breunig, M. M. , Kriegel, H-P. , Ng R. T. , and Sander, J. LOF: Identifying Outliers in Large Dataset. Proc. of ACM SIGMOD 2000, Dallas, TX 2000.
- Barnett, V. and Lewis, T. Outliers in Statistical Data. John Willey, 1994
- G Poonkuzhali, K Thiagarajan and K Sarukesi, Set theoretical Approach for mining web content through outliers detection International journal on research and industrial applications, Vol. 2, 2009, pp. 131-138
- G Poonkuzhali, K Thiagarajan, K Sarukesi and G V Uma, Signed approach for mining web content outliers. Proceedings of World Academy of Science, Engineering and Technology, Volume 56, 2009, pp -820-824.
- G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav and K. Sarukesi paper titled "Statistical Approach for Improving the Quality of Search Engine" " in the Book " RECENT RESEARCHES IN APPLIED COMPUTER AND APPLIED COMPUTATIONAL SCIENCE", included in ISI/SCI Web of Science and Web of Knowledge,Venice, Italy, 2011, pp-89-93.
- Malik Agyemang, Ken Barker and Rada S. Alhajj, Framework for Mining Web Content Outliersb. In: ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004, pp 590-594.
- Malik Agyemang, Ken Barker, Reda Alhajj, Web outlier mining: Discovering outliers from web datasets, Intelligent Data Analysis,Vol. 9, No (5)/2005, pp 473-486
- Malik Agyemang, Ken Barker and Rada S. Alhajj Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams' ACM Symposium on Applied Computing. , Santa Fe, New Mexico,2005, pp 482-487.
- Malik Agyemang Ken Barker and Rada S. Alhajj WCOND –Mine : Algorithm for detecting Web Content Outliers from Web Documents. IEEE Symposium on Computers and Communication. 2005.
- Malik Agyemang Ken Barker and Rada S. Alhajj, Hybrid Approach to Web Content Outlier Mining without Query Vector. Springer –Berlin, 2005,Vol. 3589.

- Malik Agyemang, Ken Barker, Reda Alhajj, A comprehensive survey of numeric and symbolic outlier mining techniques, Intelligent Data Analysis,Vol. 10, No (6)/2006, pp 521-538.

- Ramaswamy S, Rastogi R, Shim k, Efficient Algorithm for mining outliers from large data sets, proc. Of ACM SIGMOD 2000, pp 127 – 138.
- Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD, July 2000, Vol-2, pp 1-15.
- Xia Huosong, Fan Zhaoyan, Peng Liuyan, "Chinese Web Text Outlier Mining Based on Domain Knowledge," Intelligent Systems, WRI Global Congress on, vol. 2, pp. 73-77, 2010 Second WRI Global Congress on Intelligent Systems, 2010

Index Terms

Computer Science Information Technology

Keywords

Dissimilarity Weight Outlier Mining Term Frequency Weighted Approach Web Content Mining

Web Content Outliers