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

- Bing Liu, Kevin Chen-Chuan Chang, Editorial: Special issue on Web Content Mining, SIGKDD Explorations, Volume 6, Issue 2.
- G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav and K. Sarukesi paper titled &quot;Statistical Approach for Improving the Quality of Search Engine&quot; in the Book &quot;RECENT RESEARCHES IN APPLIED COMPUTER AND APPLIED COMPUTATIONAL SCIENCE&quot;, 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 Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams(apos); ACM Symposium on Applied Computing, , Santa Fe, New Mexico, 2005, pp 482-487.
Signed-With-Weight Technique for Mining Web Content Outliers


Index Terms

Computer Science

Information Technology

Keywords

Dissimilarity Weight  Outlier Mining  Term Frequency  Weighted Approach  Web Content Mining

Web Content Outliers