Abstract

Web Analytics is a way of turning raw data into actionable information. Large organisations own web based applications and connect to external databases which generate very large web logfiles. It then becomes crucial to estimate how information systems are accessed by staff, what their search preferences are, what documents are of greater demand. One challenge in obtaining this knowledge is that logfiles contain unstructured information where authentic search requests are not discriminated from crawler hits. Another challenge is that many proposed pattern mining techniques are usually tested on small benchmark datasets, so their performance on a large scale is hard to predict. This paper stresses the importance of data preprocessing and introduces an efficient method for mining patterns in large sized collections of web logs (of all types) based on classic set theory properties.

References

RecB: Set Theory based Technique for Large Scale Pattern Mining in Web Logs


R. Iv’ancsy, I. Vajk, Frequent Pattern Mining in Web Log Data, Acta Polytechnica Hungarica, 3 (1) 2006.


J. Pei, J. Han, B. Mortazavi-asl & H. Zhu, Mining Access Patterns Efficiently from Web Logs, in: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications (PADKK’00), Springer-Verlag London, UK, 2000.

L. Sun, X. Zhang, Efficient Frequent Pattern Mining on Web Log Data, University of Melbourne, Australia, 2004.


Index Terms

Computer Science  
Information Sciences

Keywords

Set theory, pattern mining, web mining, computational complexity, complexity reduction, big data analytics