View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Universiti Putra Malaysia Institutional Repository

Classic term weighting technique for mining web content outliers

ABSTRACT

Outlier analysis has become a popular topic in the field of data mining but there have been less work on how to detect outliers in web content. Mining Web Content Outliers is used to detect irrelevant web content within a web portal. Term Frequency (TF) techniques from Information Retrieval (IR) have been used to detect the relevancy of a term in a web document. However, when document length varies, relative frequency is preferred. This study used maximum frequency normalization and applied Inverse Document Frequency (IDF) weighting technique which is a traditional term weighting method in IR to use the value of less frequent terms among documents which are considered as more discriminative than frequent terms. The dataset is from The 20 Newsgroups Dataset. TF.IDF is used in dissimilarity measure and the result achieves up to 91.10% of accuracy, which is about 17.77% higher than the previous technique.

Keyword: Information retrieval; Outliers; Term weighting; Web content