Position score weighting technique for mining web content outliers.

ABSTRACT

The existing mining web content outlier methods used stemming algorithm to preprocess the web documents and leave the domain dictionary in their root words. The stemming algorithm was usually used to reduce derived words to their stem, base or root form. The stemming algorithm sometimes does not leave a real word after removing the stem and it caused a problem to match words in the full word profile with the domain dictionary. Therefore this study uses stemmed domain dictionary and applies it with Term Frequency with Position Score (TF.PS) weighting technique which is derived from TF.IDF weighting technique from Information Retrieval (IR) in dissimilarity measure phase to see the efficiency of these technique for determining the outliers in the web content. The dataset is from The 20 Newsgroups Dataset. The result for stemmed domain dictionary with TF.PS weighting technique achieves up to 98.19% of accuracy and 90% of F1-Measure which is higher than previous techniques.

Keyword: Information retrieval; Outliers; Web content; Weighting technique.