Term frequency-information content for focused crawling to predict relevant web pages.

ABSTRACT

With the rapid growth of the Web, finding desirable information on the Internet is a tedious and time consuming task. Focused crawlers are the golden keys to solve this issue through mining of the Web content. In this regard, a variety of methods have been devised and implemented. Many of these methods coming from information retrieval viewpoint are not biased towards more informative terms in multi-term topics (topics with more than one keyword). In this paper, by considering terms' information contents, we propose Term Frequency-Information Content (TF-IC) method which assigns appropriate weight to each term in a multi-term topic. Through the conducted experiments, we compare our method with other methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI). Experimental results show that our method outperforms those two methods by retrieving more relevant pages for multi-term topics.

Keyword: Focused crawling; Information content; Relevant page prediction; Web data mining.