Improving multi-term topics focused crawling by introducing term frequency-information content (TF-IC) measure

ABSTRACT

By rapid growth of the Internet, finding desirable information would be a challenging and time consuming task. In order to tackle this issue, focused crawlers, as the ideal solution, through mining of the Web, help us to find web pages closely relevant to the desired information. For this purpose, a variety of methods are devised and implemented. Nonetheless, the majority of these methods do not favor more informative terms in a given multi-term topic. In this paper, we propose a new measure called Term Frequency-Information Content (TF-IC) to prioritize terms in a multi-term topic accordingly. Through conducted experiments, we compare our measure against both Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI) measures applied in focused crawlers. Experimental results indicate superiority of our measure over TF-IDF and LSI for collecting more relevant web pages of both general and specialized multi-term topics.

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