Capturing Semantics of Web Pages using Weighted TAG-Tree for Information Retrieval

R. Vishnu Priya, Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamilnadu, India
A. Vadivel, Department of Computer Applications, National Institute of Technology, Tamilnadu, Tiruchirappalli, India

ABSTRACT
Web pages are highly dynamic and it's difficult to retrieve the relevant web pages in top 10 search results. This is based on some ranking mechanism incorporated retrieval system. The Retrieval system is designed for ranking the relevant web pages for user query. Usually, the retrieval system considers many techniques for ranking such as link based, connectivity based and keyword based techniques. The authors’ rank the web pages using the keywords and its associated TAGs. Based on the importance of each TAGs, weights are assigned and the semantics of the page is captured. In addition, the semantic information is represented in compact tree form, which supports both incremental and interactive mining with refined retrieval. From the experimental result, the authors have observed that the performance of the proposed approach is encouraging compared to the recently proposed approach.

Keywords: Compact Tree, Incremental and Interactive Mining, Information Retrieval, Web Semantics, Weighted TAGs

1. INTRODUCTION
In current scenario, the web is considered as a major information source in everyday and in every body’s life. Nearly, one million web pages are added every day and several hundred gigabytes are changed every month. Due to this fact of booming web data and web users, it is found to be tedious to find relevant or interesting information in top 10 retrieved results. In this situation, the web drew attention of many researchers for extracting knowledge from the web, which could also be the base process that helps Web Searching (WS), Information Retrieval (IR) and Web Mining (WM).

IR deals with the searching of relevant web pages from heterogeneous data, such as text, semi structure database, unstructured database and multimedia. The amount of web pages retrieved is on the higher side compared to the number of relevant web pages. Hence, retrieving relevant web pages is becoming an
essential issue. In order to retrieve the relevant pages, the information retrieval systems calculate a numeric score for each web page based on how well it is relevant to the user queries. The web pages are ranked based on the scores and displayed to the users. This process of web page ranking mechanism is performed in most of the well-known search engine systems.

Majority of the users use Google, MSN and Yahoo search engines for retrieving the relevant information. Currently, one of the popular search engine is Google and it indexes more than 3 billion web pages in the world as well as this number increases with the rate of 7.3 million pages per day (Forsati et al., 2009). Google use a well-known algorithm for ranking pages called page rank.

Page rank algorithm (Page et al., 1998) use link-base concept, where query independent fixed score is assigned to each element of hyperlinked set of web pages to measure relative importance of each web page within the result set. The algorithm uses the web graph, where nodes are World Wide Web pages and edges are hyperlinks. Both rank and hyperlinks are considered for ranking, where rank value indicates the importance of a page and hyperlinks are counted as vote of support. The rank of each page is defined as the weighted sum of ranks of all pages having link to the page. In addition, the value of damping factor ($d$) is added for removing the effect of sink pages. Usually, a user randomly surfs the web by clicking the links on the current page. This process of surfing a page is continued and again jumped to a random page if the user reaches a page with no output links. Therefore, the damping factor is calculated, while a user is in web pages with probability of $d$ will be selected as one output link randomly or will jump to other web pages with the probability of $1-d$. In this way, a rank for a page is calculated.

A page has a high rank if it has more back links or page having links to this page have higher ranks (Bidoke & Yazdani, 2008). If there is no links to a web page, then the page has no rank. Once the logic of ranking mechanism of Google is known, some organization has not developed their business instead they have shown interest in increasing the page rank of their pages in web. This is done with the purpose of displaying their pages in top-10 results, as the users usually browse only the first or second pages of the search result. The well-known tactics to increase page rank are publishing articles on article directories, submitting your website to web directories, exchanging links with other websites, commenting on other people’s blogs, posting on question and answer sites like Yahoo! Answers, using Twitter, Facebook and other networking sites, using bookmarking sites, participating in forums, providing an RSS feed on your website, using link building services and tools and buying links (rank.html). In addition, a score for a page is increased based on the number of time a page is visited/clicked. We can understand that the total count of clicks can be increased with a simple source code. Due to this fact all commercial web pages, social networking web pages like Facebook, Google+, orkut and so on are displayed as the top result in current search engines, which will not provide relevant information for user. It has also been found that the page rank concept is vulnerable to manipulate (PageRank).

The web search engines have applied another technique for ranking called keyword search techniques. In this technique, the web search engines fetch the web pages from web using Crawler and the collected web pages are parsed to extract keywords from the texts of web page. The extracted keywords are indexed to facilitate fast and accurate information retrieval. Query is given in the form of keywords and the web search engines examine its index. Based on its page ranking criteria, the best-matching web pages related only to the keywords query are displayed. Initially, this technique has used only frequency of occurrence of keyword for ranking. However, this scheme has considerably failed to provide preferred web pages to the user. In order to retrieve user preferred web pages, the semantics of web pages are captured from the syntax of HTML and the engines rank the web pages based on semantics rather than using only keywords. It is observed that this
Guidelines for Preparing Organizations in Developing Countries for Standards-Based B2B
[www.igi-global.com/chapter/guidelines-preparing-organizations-developing-countries/54770?camid=4v1a](www.igi-global.com/chapter/guidelines-preparing-organizations-developing-countries/54770?camid=4v1a)