Web Algorithms for Information Retrieval: A Performance Comparative Study

Bouchra Frikh, LTTI Laboratory, Ecole Supérieure de Technologie, Fes, Morocco
Brahim Ouhbi, LM2I Laboratory, Ecole Nationale Supérieure d’Arts et Métiers, Moulay Ismail University, Meknes, Morocco

ABSTRACT

The World Wide Web has emerged to become the biggest and most popular way of communication and information dissemination. Every day, the Web is expending and people generally rely on search engine to explore the web. Because of its rapid and chaotic growth, the resulting network of information lacks of organization and structure. It is a challenge for service provider to provide proper, relevant and quality information to the internet users by using the web page contents and hyperlinks between web pages. This paper deals with analysis and comparison of web pages ranking algorithms based on various parameters to find out their advantages and limitations for ranking web pages and to give the further scope of research in web pages ranking algorithms. Six important algorithms: the Page Rank, Query Dependent-PageRank, HITS, SALSA, Simultaneous Terms Query Dependent-PageRank (SQD-PageRank) and Onto-SQD-PageRank are presented and their performances are discussed.

Keywords: HITS, Onto-SQD-PageRank, PageRank, Performance Comparative Study, SQD-PageRank, SALSA, SQD-PageRank

INTRODUCTION

The World Wide Web is a huge, widely distributed, global information service for every kind of information such as news, advertisements, consumer information, financial management, education, government, e-commerce, health services, and many other information services. The amount of information on the Web is growing rapidly, as well as the number of Web sites and Web pages per Web site. It contains a rich and dynamic collection of hyperlink information, data of all types: structured tables, texts, multimedia data (e.g., images and movies), semi-structured HTML code information, etc. Development of such successful sites is usually an iterative
process during which the developers get some feedback from users of the current version of the site: Which pages are visited most frequently? When? What is the typical order in which pages are visited? Which sequences led to an action (such as placing an order, request for an offer or information)? Etc.

Because of its rapid and chaotic growth the information on the Web is heterogeneous, noisy, and redundant since multiple Web pages may present the same or similar information using completely different formats or syntaxes. Consequently, the resulting network of information lacks of organization and structure. So, it has become more difficult to find relevant and useful information for Web users. In that context, predicting the needs of a Web user as she visits Web sites has gained importance. The requirement for predicting user needs in order to guide the user in a Web site and improve the usability of the Web site can be addressed by recommending pages to the user that are related to the interest of the user at that time. The modeling of behavior of web users is often labeled as Web Usage Mining, which, together with Web Content Mining and Web Structure Mining forms a dynamic field of research called Web Mining. It consists in use of data mining techniques to automatically discover and extract information from the World Wide Web (WWW).

The Web can be seen as a directed labeled graph whose nodes are the documents or pages that can be browsed using a normal Web browser and the edges are the hyperlinks between them. These links serve as an information organization tool and also as indications of trust/authority in the linked pages and sites. Deep Web is mainly composed of databases that can only be accessed through parameterized queries using query forms.

Due to the heterogeneity and lack of structure of Web data, there are various challenges associated with the ranking of web pages such that some web pages are made only for navigation purpose and some pages of the web do not possess the quality of self descriptiveness. Automated discovery of targeted or unexpected knowledge/information is a challenging task. It calls for novel methods that draw from a wide range of fields spanning data mining, machine learning, natural language processing, statistics, databases, and information retrieval. The objectives of this paper is to analyze the currently important algorithms for ranking web pages to find out their relative strengths, limitations and to provide a future direction for the research in the field of efficient algorithm for ranking web pages. For ranking web pages, There are a number of algorithms proposed based on content and (or) link analysis in the literature (Duhan et al., 2009; Langville & Meyer 2005; Scime 2005; Liu 2008; Chevalier et al., 2003). Six important algorithms Page Rank, Query Dependent-PageRank, HITS, SALSA, Simultaneous Terms Query Dependent-PageRank(SQD-PageRank) and Onto-SQD-PageRank are discussed below.

This paper is organized as follows: section 2 deals with the related works. Section 3 presents important existing
Related Content

Transforming eMaintenance into iMaintenance with Mobile Communications Technologies and Handheld Devices
[www.irma-international.org/chapter/transforming-emaintenance-into-imaintenance-mobile/60219/](www.irma-international.org/chapter/transforming-emaintenance-into-imaintenance-mobile/60219/)

Security Framework for Mobile Agents-Based Applications
[www.irma-international.org/chapter/security-framework-mobile-agents-based/62198/](www.irma-international.org/chapter/security-framework-mobile-agents-based/62198/)

A Context Transfer Model for Secure Handover in WiMAX/LTE Integrated Networks
[www.irma-international.org/article/a-context-transfer-model-for-secure-handover-in-wimaxlte-integrated-networks/130481/](www.irma-international.org/article/a-context-transfer-model-for-secure-handover-in-wimaxlte-integrated-networks/130481/)

New Transaction Management Model
[www.irma-international.org/chapter/new-transaction-management-model/17158/](www.irma-international.org/chapter/new-transaction-management-model/17158/)

Web Page Adaptation and Presentation for Mobile Phones
[www.irma-international.org/chapter/web-page-adaptation-presentation-mobile/41635/](www.irma-international.org/chapter/web-page-adaptation-presentation-mobile/41635/)