UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

A Comparison Study between Entropy and Page Rank Techniques to Retrieve Relevant Web Pages

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Systems

Prepared By:

Mohammed Almontser Ali Babiker

Supervisor By:

Dr. Hassan Fadlallah

Jan 2017

Abstract

Within the rapid growth of the Web, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult. The reasons are that some web pages are not self-descriptive and that some links exist purely for navigational purposes. To confront this problematic in an accurate way Information Retrieval is a key factor for this issue. Information retrieval is a large field of research, it's specializes within obtaining information resources relevant to an information need from a collection of information resources. One of its types is web information retrieval which falls under web mining processes. It's about how to retrieve the relevant web pages for users who's seek for information. Recently many techniques had been developed on this field. But yet there is inaccurate results on web information retrieving process. One of the most common used techniques in search engines are Entropy and Page ranking techniques. The two techniques are differ from each other on many phases but the beautician of them is to give the relevant information for users queries. Making a comparison between these techniques on their accuracy measure is gainful process to elicit the superior and the functional technique and this is the aim of this thesis. On this thesis we implement Entropy and Page ranking algorithms using JAVA programming language. Also we prepare a dataset to act as an input for the implementation. The results of the two algorithms compared within a manual ranking results prepared in advance. Ultimately the experiments we conducted presented that Entropy technique is the best technique on retrieving the relevant information of web pages.

المستخلص

مع تسارع وزيادة عالم الويب اليوم أصبح توفير معلومات الصفحات بجودة ودقة للمستخدمين الذين يبحثون في الويب من الأمور الصعبة في ظل هذا التطور المستمر. والسبب في ذلك أن أغلبية هذه الصفحات ليست (self – descriptive)، وأيضا الكثير من الروابط التشعبية ليست معنية بتقديم المعلومات إنما مجرد روابط تساعد في عمليات البحث في الإنترنت. لمجابهة هذه المعضلة بصورة دقيقة وجد مفهوم إسترجاع البيانات ليكون الحل الأمثل في عملية إسترجاع المعلومات بصورة دقيقة ومثلى. مفهوم إسترجاع البيانات مجال بحثي واسع وهو يختص بإستنباط مصادر البيانات ذات الصلة بالمعلومات المطلوبة من أصل مجموعة مصادر مختلفة. إحدى أنواع هذا المجال هو إسترجاع البيانات من صفحات الإنترنت، الذي يتدرج تحت عمليات تنقيب الويب. يهتم مجال إسترجاع البيانات من الويب بكيفية إسترجاع المعلومات ذات الصلة بطلبات المستخدمين عند البحث في الويب. مؤخرا تم تطوير الكثير من التقنيات في هذا المجال ولكن ما ذال هنالك بعض القصور في عمليات إسترجاع البيانات. من أكثر التقنيات إستخداما في محركات البحث هما Page Ranking و ال Entropy. تختلف هذه التقنيات عن بعضهما البعض في أوجه شتى ولكن المجمل انهم يختصوا بعملية إسترجاع بيانات الصفحات بناء على إستعلام المستخدمين وفقا لمعايير معينة. المقارنة بين هاتين التقنيتين وعلى وجه الخصوص المقارنة بين معايير الدقة والمثالية بينهما عملية مربحة للجانب المعرفي، حيث أنها تتيح معرفة وإستخلاص أي من التقنيتين لها الأفضلية والأداء الفعال في عملية إسترجاع البيانات من الويب وهذا هو الهدف من هذا البحث. في هذا البحث طبقنا خوارزمية هاتين التقنيتين بإستخدام لغة البرمجه جافا، وأيضا قمنا بإعداد مستودع بيانات ليكون المدخل للتطبقين. المخرج أو النتائج من التطبيقين تمت مقارنته مع ترتيب الصفحات اليدوي الذي تم تجهيزه مسبقا. في نهاية المطاف هذه التجربة التي قمنا بتطبقيها أظهرت أن تقنية ال Entropy هي التقنية الأمثل والأفضل أداء وأكثر دقة في عملية إسترجاع أو إستخلاص البيانات من الويب.

1.1 Introduction

In the highly competitive world and with the broad use of the Web in e-commerce, e-learning, and e-news, finding users' needs and providing useful information are the primary goals of website owners. Therefore, analyzing users' patterns of behavior becomes increasingly important. Web mining is used to discover the content of the Web, the users' behavior in the past, and the webpages that the users want to view in the future [1].

With the rapid growth of the Web, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult. The reasons are that some web pages are not self-descriptive and that some links exist purely for navigational purposes. Therefore, finding appropriate pages through a search engine that relies on web contents or makes use of hyperlink information is very difficult.

To address the problems mentioned above, several algorithms have been proposed. Among them are PageRank and Entropy algorithms. PageRank is a commonly used algorithm in Web Structure Mining. It measures the importance of the pages by analyzing the links. PageRank has been developed by Google and is named after Larry Page, Google's co-founder and president. PageRank ranks pages based on the web structure [2].

The concept of entropy based on information theory was introduced by Shannon and other computer scientists also known as Shannon entropy, it is actually the measure of uncertainty, inconsistency, unstructured data etc. of the random variable. From this, information content of web page can be obtained. Information Content of document tends to give efficient result. A web page having low information content, then that web page have the lowest unstructured data, in other word the web page has lowest uncertainty that means web page leads towards the relevancy. If a web page is more relevant and accurate then that web page having low Information Content [3].

The comparison of these algorithms is feeding back the web researches to improve and develop more accurate algorithms that can help in web pages extraction and providing the relevant data within highest quality to users who's searching for information's.

1.2 Problem Statement

The tremendous growth of data sources available on the web makes information retrieval a tedious and difficult for users because some web pages are not self-descriptive and that some links exist purely for navigational purposes. Many techniques had been developed to solve this problem. Comparison between those techniques or algorithms is a key for evaluate and decide which algorithm is suitable for information retrieval process.

1.3 Research Objective

Thesis objective are:

1. Compare and measure the accuracy between (PageRank, and Entropy) web mining techniques.
2. Determine which algorithm have the most accurate results on information retrieval process.

1.4 Research Methodology

The research methodology composed of the following steps:

1. Study the needed algorithms.
2. Implement the algorithms.
3. Test the Algorithms.
4. Analyze the results of the algorithms.
5. Compare the accuracy of the algorithms.

1.5 Thesis Structure

The thesis contains five chapters: the second chapter is a literature review chapter, which discuss the Information Retrieval, web mining background and related work. The third chapter describes the research methodology. The fourth chapter illustrates the results and discussions of the comparison. The fifth chapter includes conclusion and recommendations.