A Hybrid Approach for Web Change Detection

Sakher Khalil Alqaaidi, Middle East University for Graduate Studies, Amman, Jordan

ABSTRACT

Search engines save copies of crawled web pages to provide instant search results. Saved pages may become old and un-updated as original pages change providing new information and new links, and most of websites don’t submit these new changes to search engines so search engines don’t depend mainly on websites techniques of submitting changes. Keeping pages fresh and updated in search engine is important for giving real page ranks and for providing real time information. Techniques were invented to improve the page update process by search engine. In this paper the author combines two of good known techniques and implements the new one via experiments that improve better results in different experiment cases.

Keywords: Crawled Web Pages, Page Freshness, Page Update Process, Search Engine, Web Crawler

1. INTRODUCTION

Search engines are the easiest and fastest tool to get information on the web, they consist of many programs that index pages, crawl over the web, and serve requests. One of these programs is web crawler (Ezz, 2007; Castillo, 2004) which is required to take copies of web pages to provide instant search results by search engine, as web contains now billions of pages (Antonio & Signorini, 2005) with dynamic content due to the technology of dynamic pages generated by web server software and programming languages that uses databases to store content like PHP and MySQL (Gilmore, 2006), adding pages to websites does not require generating HTML files (Nielsen 2005) and can be done easily via editors in same website interface, so these pages could be generated or changed at any time or in defined time periods like a daily newspaper website, so web crawler should effectively follow these updates and add new content to improve search engine performance in providing real time information and faster reaching to the required result through ordering search results which depends on page rank (Andrew, 2004) factor calculated by algorithms that depend mainly on URL links found in the crawled pages or the ones that points to a required page, and these links also change as information changes in different periods of time. Many techniques were found that organize the process of taking copies of web pages by web crawler through determining the best time to perform a visit to the page to get the new updates immediately after they occur by expecting the change time which is important for web crawler in different situations because doing many visits to the page in short periods of time as illustrated in Figure 1 trying to detect the updates once it...
occur will affect the performance of the crawler as it is required to do the same for billions of pages in the same way and will require using many resources to perform this, and in case a visit has been performed without getting new updates in the page this also will be considered wasting of resources, visiting pages so many times in short periods of time will affect also the server that hosts the page when it suffer from heavy visits performed by search engines, also making visits in long periods of time will cause missing possible updates in pages as shown in Figure 1. This study describes two of best policies used in detecting page freshness and contributes in proposing new policy by combining the previously described ones and simulate it in different cases like suffering from low resources or handling huge number of web pages to crawl.

The rest of the paper is divided as follows. Section 2 present related work. Section 3 describes sampling policies. Section 4 covers the proposed policy. Section 5 presents the experiments. Section 6 present conclusions and future work.

2. RELATED WORK

Ezz (2007) came up with a fact that it is impossible to keep visiting all pages in the web in a short period of time trying to detect fresh updates due to the huge amount of pages there and it requires 8.5 days to visit all the pages if we assumed that a web page has an average size of 12 kilobytes, a simple equation has been proposed to calculate the speed of any web crawler that indexes all the web as follow:

$$\frac{10^8}{8.5 \text{days}} \times \frac{12 \text{bytes}}{1 \text{page}} = 1.3 \text{Gbit/sec}$$
Related Content

Web Retrieval of XML Documents: Practice and Challenges
www.irma-international.org/chapter/web-retrieval-xml-documents/31415/

How Culture May Influence Ontology Co-Design: A Qualitative Study
www.irma-international.org/article/culture-may-influence-ontology-design/55380/

Patterns for Improving the Pragmatic Quality of Web Information Systems
www.irma-international.org/chapter/patterns-improving-pragmatic-quality-web/21965/

A Novel Cache Replacement Policy for Web Proxy Caching System Using Web Usage Mining
www.irma-international.org/article/a-novel-cache-replacement-policy-for-web-proxy-caching-system-using-web-usage-mining/159155/

City Brands and their Communication through Web Sites: Identification of Problems and Proposals for Improvement
www.irma-international.org/chapter/city-brands-their-communication-through/37683/