International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 337 - 339

Implementation of Clever Crawler

Anuja Lawankar
Department of Computer Science Engineering,Yeshwantrao Chavan College of Engineering,Nagpur, Maharashtra, India.
lawankar.anuja08@gmail.com

Nikhil Mangrulkar
Department of Computer Science Engineering,Yeshwantrao Chavan College of Engineering,Nagpur, Maharashtra, India.
lawankar.anuja08@gmail.com

***Abstract*—** Now a days due to duplicate documents in the World Wide Web while crawling, indexing and relevancy. Search engine gives huge number of redundancy data. Because of that, storing data is waste of rankings quality, resources, which is not convenient for users. To optimized this limitation a normalization rules is use to transform all duplicate URLs into the same canonical form and further optimized the result using Jaccard similarity function which optimized the similar text present in the content of the URL's these function define some threshold value to reduce the duplication.

Keywords- *Web crawler, Non Duplicated URLs, Normalized URLs, Jaccard Similarity Function.*

*****

I. INTRODUCTION

To fetch topic related information search engine use the web crawler. A crawler is a program that visits Web sites and reads their web pages and other information in order to create entries for a search engine index. Crawler starts collecting the set of URLs, called seed URLs. Many legitimate sites, in particular search engines, use crawling as a means of providing up-to-date data. World Wide Web have graphical structure, i.e. the links given in a page can be used to open other web pages. Actually Internet is a directed graph, webpage as node and hyperlink as edge, so the search operation could be abstracted as a process of traversing directed graph. Syntactically different URLs that have similar content is a common phenomenon on the web. For instance, in order to facilitate the user's navigation, many websites define links or redirections as alternative paths to reach a document. In addition, webmasters usually mirror content to balance load and ensure fault tolerance. Generally duplication of contents are due to generation of dynamic web pages that are invoked by the web crawler. On web there are large-scale deduplication of documents. Web pages which have the same content but it redirect to different URLs, are known to cause a host of problems. There are in fact two types of near duplicates for this application. Figure 1(a) and 1(b) shows a pair of same-core web pages that only differs in the framing, advertisements, and navigational banners added each by the web site.

Fig 1(a) Near-duplicate Web pages

Both articles exhibit almost identical core contents, reporting Bridgestone marketing company pre-election phase, because both were delivered by Associated Press same-core pair to be near duplicates, since the core articles are their focus. In particular for domains like stock markets, news sites often use very uniform layouts and the actual contents-of-interest only constitute a fraction of the page. Crawler that harvests such a website daily will keep stumbling into those areas. However, rather than re-collecting identical pages, the crawler is served with different framing ads upon each access, which introduces near duplicates into the archived. Near duplicates would not be discarded but could be collected into a same set which is then added into batch [3]. To find out such a duplicate results is an extremely important task for search engines since crawling this redundant content leads to many limitation such as waste of resources. Resources are wasted in fetching duplicate web pages, indexing and relevancy of results are listed out for a query [1].

Fig 1(b) Identical core content with different framing and banners

To overcome these problems, several authors have proposed methods for detecting and removing such duplicated content from search engines. The web crawler firstly fetches the URLs from the application server. Large number of URLs are served by the web crawler to optimize this URLs, first focus on forming the clusters of the similar content are formed by use cluster normalization rules that transform duplicate URLs into a unified canonical form and optimizes the URLs. Further optimization of each cluster, comparing document content by using the Jaccard similarity coefficient which is commonly used to measure the overlap between two sets. The similar words are extracted from web page and only if the similar words are cross the threshold value which represent the similarity. In this way relevant links are reduced.

II. RELATED WORK

The proposed techniques like normalization and jaccard similarity function which given the more relevant results for small dataset (upto 100kb). But this techniques were not gives the optimized result for big dataset (more than 1MB). To optimize the URLs on big dataset the normalization and jaccard similarity function is used with the WARC dataset and Google results.Collected the dataset of WARC file. Working on collecting the raw data, which will help for improve the scalability of program code.The Wamp server is used to show the whole data extracted like URLs, content and all the metadata related to that webpage. The dataset present in WARC file is contain some URLs and summary of the particular webpage. To collect the maximum data related to the search text Google is used and all the URLs and contents of that URLs are fetched. Further, the content of that URLs are extracted from the web are store in the database. Normalized and optimized URLs are served on web in search engine. Normalization Techniques is apply on the duplicated URLs by giving a set of URLs U (i.e., a training set) partitioned into groups of similarpages (referred to as dup-cluster) from one or moredomains. The strategy of the URL-based de-dupingmethods is to learn, by mining these dup-clusters, rulesthat transform duplicate URLs to the same canonical form. In Table1, U= {u0, u1, u2, u3, u4} is partitioned in dup-clusters C1 and C2. The canonical form of C1 and C2 are given by n1, n2 respectively. This process, called as URL normalization, identifies, at crawling time. The normalization of URLs. U shows the static URL and the u0, u1, u2, u3, u4 are the duplicated URLs. URLs u0,u1,u2 having the same content but different URLs which shows the dynamic web pages to optimized this different URLs showing the same content normalization technique is used by taking the canonical form. The jaccard similarity function is used for compare the contents between two web pages. The function is depends on the threshold value which is calculated according to the contents present in the web pages. The scoring function is the Jaccard similarity coefficient [1] which is commonly used to measure the overlap between two sets. For two sets, it is denoted as the cardinality of their intersection divided by the cardinality of their union.[1]

Equation:

$$sf(Xi,Yi)= \left\{\frac{(Xi \cap Yi)}{(Xi \cup Yi)} \; if \; \exists (XiY)\right\} \ldots\ldots\ldots\ldots\ldots\ldots [1]$$

As shown in the equation Xi is the web page and Yi is another web page to compare the content between two web pages the intersection upon union of the contents presents in the web page is calculated. The intersection is the common or similar words presents in two web pages and the union is the total words presents in these web pages. After calculating the threshold value which helps to discarding the relevant URLs. The threshold value 0.9 is having 90% of same content between two web pages. Hence, web pages having less than 0.9 threshold value will displayed on the search engine and we pages having 0.0 threshold value will be discarded. According to this function more relevant URLs are served on the search engine and URLs are optimized.

REFERENCES

[1] Kayo Rodrigues, Marco Cristo, Deleon S. de Moure, and Antiguan da Silva, “Removing DUST Using Multiple Alignment of Sequences" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 8, AUGUST 2015.

[2] A. Agarwal, H. S. Copula, K. P. Lela, K. P. Chitrapura, S. Garg, P. Kumar GM, C. Haty, A. Roy, and A. Sasturkar, “Url normalizationfor de-duplication Of web pages," in Proc. 18th ACM Conf. Inf.knowl. Manage., 2009, pp. 1987-1990.

[3] M. Theobald, J. Siddharth, and A.Paepcke, “Spotsigs: Robust and effcient near duplicate detection in large web

collections," in Proc. 31stAnnu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 563-570

[4] X. Mao, X. Liu, N. Di, X. Li, and H. Yan, "Sizespotsigs: An effective deduplicate algorithm considering the size of page content," in Proc. 15th Pacic-Asia Conf. Adv. Knowl. Discovery Data Mining, 2011, pp. 537-548.

[5] V.A.Narayana P. Premchand Dr. A. Govardhan, "A Novel and Effcient Approach For Near DuplicatePage Detection in Web Crawling" 2009 IEEE International Advance Computing Conference (IACC 2009)Patiala, India, 6-7 March 2009

[6] Lei Xiang XinMeng, "A data mining approach to topic-specific web resource discovery"2009 Second International Conference on Intelligent Computation Technology and Automation, 978-0-7695-3804-4/09 26.00 ➞ 2009 Crown Copyright DOI 10.1109/ICICTA.2009.378

[7] LingXia Hu, "Intelligent Crawling based on Rough Set for Web Resource Discovery", 978-1-4244-5265-1/10/26.00 ➞2010 IEEE

[8] DuyguTaylan, MitatPoyraz, SelimAkyokusand Murat Can Ganiz, "Intelligent Focused Crawler: Learning which Links to Crawl" 978-1-61284-922-5/11/26.00 ➞2011 IEEE.

[9] Ashish Kumar Sultania, "Algorithm for Detecting Dynamic Webpage and its Importance"2012 International Conference on Radar, Communication and Computing (ICRCC),SKP Engineering College, Tiruvannamalai, TN., India. 21 - 22 December, 2012. pp.257-259.

[10] H. S. Koppula, K. P. Leela, A. Agarwal, K. P. Chitrapura, S. Garg, and A. Sasturkar, "Learning url patterns for webpage deduplication," in Proc. 3rd ACM Int. Conf. Web Search Data Mining, 2010, pp. 381-390.

[11] A. Dasgupta, R. Kumar, and A. Sasturkar, "De-duping urls via rewrite rules," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp.186-194.

[12] T. Lei, R. Cai, J.-M. Yang, Y. Ke, X. Fan, and L. Zhang, "A pattern tree-based approach to learning URL normalization rules," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 611-620

[13] Z. Bar-Yossef, I. Keidar, and U. Schonfeld, "Do not crawl in the dust: Different urls with similar text," ACM Trans. Web, vol. 3, no. 1, pp. 3:1-3:31, Jan. 2009.

[14] Reihaneh Emamdadi, Mohsen Kahani, Fattane Zarrinkalam "A Focused Linked Data Crawler based on HTML Link Analysis," 2014 4th International Conference on Computer and Knowledge Engineering(ICCKE), 978-1-4799-5487-2/14/$31.00©2014/IEEE.

[15] Ruchika Patel Pooja Bhatt, "A Survey on Semantic Focused Web Crawler for Information Discovery Using Data Mining Technique" IJIRST {International Journal for Innovative Research in Science and Technology| Volume 1 | Issue 7 | December 2014ISSN (online): 2349-6010