A Survey on Design and Implementation of Clever Crawler based on DUST Removal

Kanchan S. Khedkar[1] P. L. Ramteke[2]
[1,2]Department of Computer Science & Information Technology
[1,2]HVPM COET, Amravati, India

Abstract— Now days, World Wide Web has become a popular medium to search information, business, trading and so on. A well know problem face by web crawler is the existence of large fraction of distinct URL that correspond to page with duplicate or nearby duplicate contents. In fact as estimated about 29% of web page are duplicates. Such URL commonly named as dust represent an important problem in search engines. To deal with this problem, the first efforts is focus on comparing document content to detect and remove duplicate document without fetching their contents .To accomplish this, the proposed methods learn normalization rules to transform all duplicate URLs into the same canonical form. A challenging aspect of this strategy is deriving a set of general and precise rules. The new approach to detect and eliminate redundant content is DUSTER .When crawling the web duster take advantage of a multi sequence alignment strategy to learn rewriting rules able to transform to other URL which likely to have same content . Alignment strategy that can lead to reduction of 54% larger in the number of duplicate URL.

Key words: DUSTER, web crawling, Dust Buster

I. INTRODUCTION

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 web sites define links or redirections as alternative paths to reach a document. In addition, webmasters usually mirror content to balance load and ensure fault tolerance. Other common reasons for the occurrence of duplicate content are the use of parameters placed in distinct positions in the URLs and the use of parameters that have no impact on the page content, such as the session id attribute, used to identify a user connection or the cookie information being stored in the URLs [1]. 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 de-duplication of documents. Web pages which have the same content but are referenced by different URLs, are known to cause a host of problems.

II. FUNDAMENTALS OF WEB CRAWLING

Crawlers have bots that fetches new and recently changed websites, and indexes them. By this process billions of websites are crawled and indexed using algorithms (which are usually well guarded secrets) depending on a number of factors. Several commercial search engines changes the factors often to improve the search engines process. It generally starts with a set of URLs from the previous crawl, visits each of these websites, detects links and adds it to the list of links to crawl. It also notes whether there is any new website or website that has been recently changed (updated), websites that are no more in use and accordingly index is updated. Indexer compiles the list of words it sees and its location on each page for future consultation. The information compiled are mostly because crawlers are majority text based. When an user initiates a search, the key words are extracted and searches the index for the websites which are most relevant. Relevancy is determined by a number of factors and also it different for the different search engines.

III. PROBLEM STATEMENT

The problem addressed in this project is related to de-duplication of web pages. More specifically, it is related to the existence of syntactically different URLs linking to the same content. These URLs, generically known as DUST, usually have specific patterns that can be learned and used by URL based de-duping methods. The input to this problem consists of a set of URLs U (i.e., a training set) partitioned into groups of similar pages (referred to as dup-cluster) from one or more domains. The strategy of the URL-based de-duping methods is to learn, by mining these dup-clusters, rules that transform duplicate URLs to the same canonical form. In Table1. U={ u1; u2; u3; u4; u5} is partitioned in dup-clusters C1 and C2. The canonical form of the URLs in C1 and C2 are given by n1 and n2, respectively. Note that the URLs of a same dup-cluster point to the same or similar content where URLs from different dupclusters likely correspond to different content. This process, called as URL normalization, identifies, at crawling time, whether two or more URLs are DUST without fetching their contents. As crawlers have resources constraints, the best methods are those that achieve larger reductions with smaller false positive rates using the minimum number of normalization rules.

IV. SURVEY DETAILS

Current research on DUST detection can be classified in two main families of methods: content-based and URL based . In content-based DUST detection, the similarity of two URLs is determined by comparing their contents using syntactic or semantic evidence as shingles, text signatures, pair-wise similarities, sentence-wise similarities, and semantic graphs. Thus, in content-based DUST detection, to infer if two distinct URLs correspond duplicates, or near duplicates, it is necessary to fetch and inspect the whole content of their corresponding pages. In order to avoid such a waste of resources, several URL-based methods have been proposed to determine duplicate URLs without examining the associated contents. For a comprehensive review of the literature, we refer the reader to that describe both and URL-based methods. In the following paragraphs, we focus on URL-based methods including the ones that, as far as we know, reported the best results in the literature.

The first URL-based method proposed was Dust Buster. In their work, the authors addressed the DUST detection problem as a problem of finding normalization rules able to transform a given URL to another likely to

have similar content. The rules consist of substring substitutions learned from crawl logs or web logs. Rules are selected if they have large support, they do not come from large groups and URLs matched by them have similar sketches or compatible sizes in the training log. Redundant rules are eliminated based on their support information. By evaluating their method in four websites, the authors found that up to 90 percent of the top 10 rules were valid, 47 percent of the duplicate URLs were recognized and the crawl was reduced by up to 26 percent. Since substitution rules were not able to capture many common duplicate URL transformations on the web, Dasgupta et. al. presented a new formalization of URL rewrite rules. The new formulation was expressive enough to capture all previous substitution rules as well as more general patterns, such as the presence of irrelevant substrings, complex URL token transpositions and session-id parameters. The authors use some heuristics to generalize the generated rules. In particular, they attempt to infer the false positive rate of the rules in order to select the most precise ones. To accomplish this, they verify if the set of values that a certain URL component assumes is greater than a threshold value N, a heuristic which they call fanout-N. Their best results were obtained with N ¼ 10. In this work, we refer to this method as Rfanout_10. By applying the set of rules found by Rfanout_10 to a number of large scale experiments on real data, the authors were able to reduce the number of duplicate URLs by 60 percent, whereas only substitution rules achieved 22 percent reduction. The authors in extended the work in to make their use feasible at web scale. They observed that the quadratic complexity of the rule extraction performed in is prohibitive for very large dup-clusters. Thus, they proposed a method for deriving rules from samples of URLs. In addition, they used a decision tree algorithm to learn a small number of higher precision rules to minimize the number of rules deployed to the crawler. The main differences between this work and the previous one are the handling of large dup-clusters; the adoption of new methods for intra cluster generalization and alignment penalization the elimination of a hierarchical clustering step with the reduction of the number of generated rules; and the simplification of the algorithm, by supporting fewer kinds of tokens.

V. PROPOSED WORK AND OBJECTIVES

Different URLs that have similar content is known as DUST. Detecting such a duplicate results is an extremely important task for search engines since crawling this redundant content leads to several drawbacks such as waste of resources. Crawler resources are wasted in fetching duplicate pages, indexing requires larger storage and relevance of results are diluted for a query[1]. To overcome these problems, several authors have proposed methods for detecting and removing such duplicated content from search engines.

The previous papers, focused only comparing document contents, or strategies that inspect only the URLs without fetching the corresponding page content . These methods, known as URL-based de-duping, mine crawl logs and use clusters of URLs referring to (near) duplicate content to learn normalization rules that transform duplicate URLs into a unified canonical form. This information can be then used by a web crawler to avoid fetching DUST. The main challenge for these methods is to derive general rules with a reasonable cost from the available training sets. Thus, an ideal method should learn general rules from few training examples, taking maximum advantage, without sacrificing the detection of DUST across different sites. DUSTER a new method that takes advantage of multiple sequence alignment in order to obtain a smaller and more general set of normalization rules .Following are steps for detecting and removing duplicated content from search engines:

1) Web crawler firstly fetches the URLs from the application server.
2) Large number of URLs are serve by the web crawler to optimize this URLs, first focus on forming the clusters of the similar content are formed by mine crawl logs.
3) Use cluster normalization rules that transform duplicate URLs into a unified canonical form and optimizes the URLs.
4) After that for further optimization of each cluster, comparing document content by using Jaccard similarity coefficient which is commonly used to measure the overlap between two sets.
5) Similar words from the duplicated contents are extracted only if the similar words are cross the threshold value which represents the similarity.
6) In this way relevant links are reduced.

VI. AIMS AND OBJECTIVES

The main aim is to reduce the relevant links from the result of search and display the Optimized result by applying optimization algorithms and techniques.

A. Objectives:

1) To optimize the search results links and give accurate result.
2) Help to minimize the search.
3) Implementation of proposed work on data sets

VII. CONCULSION

The main objective of the review paper was to throw some light Design and Implementation of Clever Crawler base on dust removal .We also discussed the various methods and the researches related to respective project and their strengths and weaknesses associated. We believe that all the surveyed in this paper are effective for web search, but the advantages favors more for clever crawler due to reducing the number of duplicated links during web search.

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 efficient near duplicate detection in large web collections," in Proc. 31stAnnu. Int. ACM SIGIR Conf. Res. Develop. Inf.
Retrieval, 2008, pp. 563–570.
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 Pacific-Asia Conf. Adv. Knowl. Discovery
Data Mining, 2011, pp. 537–548.

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

[5] 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.

