A SURVEY ON DUST (DUPLICATE URLs WITH SIMILAR TEXT) USING MULTIPLE ALIGNMENT OF SEQUENCES

Bhagwat K.P. and Chaudhari M.S.

Abstract: The use of World Wide Web is the largest collection of data today and it is continuously increasing day by day. While crawling the duplicate contents, search engines have to waste time for different versions of a page. To deal with this, several studies have been proposed to detect and remove duplicate documents without fetching their contents. If we make use of such documents with duplicate contents then it will result in building of low quality rankings, a waste of resources and poor user experiences. Normalization rules are proposed to transform all duplicate URLs into the same canonical form. We are going to make use of DUSTER that will derive quality rule. We will convert URL into sequences by making use of multiple sequences Alignment algorithm by performing the operations. So web crawlers can help in making the indexing possible.

Keywords: URL (Uniform Resource Locator), Search engine, web technology, normalization rules, web crawling, DUST (Duplicate URLs with Similar Text)

I. INTRODUCTION

In the search engine, the results which come after crawling the web pages contains duplicate or near-duplicate contents. So these duplicate URLs are known as DUST. If we check, URLs such as http://google.com/news and http://news.google.com can result the same content. When we search a data, hundreds of results appear. To find the most accurate result search engines have a huge job of providing the information on the page. Therefore, to make easy the user’s navigation, many web sites developers define links or redirection as alternative path to find a document. Also it can be used by the webmasters for the balancing of the load and to ensure fault tolerance. Thus to detect the duplicate contents is an extremely important task because crawling the redundant contents tends to several drawbacks. To overcome the drawbacks several studies have been focused on comparing URLs rather than the whole document content. By the analysis of this technique we can say that URL based DUST detection is efficient and is also known as URL based de-duping.

These URL based de-duping mine crawl logs, and make use of clusters for referencing the near duplicate contents to learn normalization rules that transform duplicate URLs into a unified canonical form. Then this information can be used by the web crawler to avoid fetching DUST including the ones that are found when we crawl the content for the first time. To obtain a more general set of Normalized rule the Multiple Sequence Alignment (MSA) algorithm is used.

This survey paper presents an up-to-date review of existing literature survey in duplicate and near duplicate detection in Web. In response to users query Search engines typically produces the list of documents ranked according to closest of it. These documents are presented to users for analysis and evaluation. Web users have to go through a long list of documents and have to inspect the titles and then contents sequentially to recognize the required results. Filtering the search engines result consumes the users effort and time especially when there are lot of near duplicate content.

II. LITERATURE SURVEY

Prior to the detection of duplicate content there are two methods such as content based and another is URL based. But in content based we have to read the entire content of any document and compare with the entire content of another document. So it leads to the waste of time and resources as well. URL based does not need the entire content so it is more effective means for the detection of duplicate document. Dust Buster mines from the past web server logs, without looking at the page details.
Kaio Rodrigues, M. Cristo, E. S. de Moura and A. S. Silva [1] and [5] proposed Dust-Buster algorithm which is used for mining DUST from several URLs. When the web crawlers collect a large number of URLs it corresponds to the duplicate or near duplicate content. This uses multiple sequence alignment to find similarities and differences in URLs. The Dust detection is a problem of finding normalization rules which should transform a given URL to another likely to have similar content. But the substitution rules were not able to capture huge common duplicate URL transformations on the web, so rewrite rules were introduced for the formalization of URL. This helps in reducing the duplicate URLs.

A. Agrawal, H. S. Koppula and B. S. Alsulami [2] and [3] proposed that without explicitly knowing the entire contents of the web page the clusters are divided according to the content of similar pages. These pages are then applied with specific rules from URLs belonging to similar cluster. Due to these rules it gets easy to get the exact URL asked by the user.

Z. Yossef [4] describes the approach for the duplicate URLs when we crawl the web pages. So DUST algorithm is used for detection of such duplicate URLs. Two rules are imposed: One is string substitution which replaces the occurrence of string in URL by other substring and other is parameter substitution which replaces the value of parameter in URL by other value.

A pattern tree based URL normalization technique has been proposed by Tao Lei et[8] in which statistics from the entire training sets are calculated. The tree pattern is generated for a website which removes conflicts and reduces redundancies. Each node in a tree pattern represents a group of URLs which have the same syntax.

A. Dasgupta [6] proposed for the large number of URLs which can be partitioned into equivalence classes and each equivalence class containing URLs should have the similar content. It address the problem of mining the set and learning URL rewrite rules that transforms all the URLs of an equivalence class to the same canonical form and then these rewrite rules can be applied to eliminate duplicate URLs without fetching the content that are encountered during crawling the web page for the first time.

X. Mao, X. Liu [7] analysed that the short web pages cannot be handled by algorithms. So by taking the size of the page contents into account the various algorithms can be applied on short as well as long web pages. When the documents especially contain noisy data such as Ads and banners then its presentation becomes poor. So as to improve the effectiveness of short web pages de-duplication is taken care of considering the size of core content.

M. Theobald and J. P. Kumar [9] [10] proposed the near duplicate contents de-duplication without fetching the entire content of the document.

### III. CONCLUSION

In this paper DUSTER is introduced to remove the duplicate URLs with similar text. This method detects, distinct URLs that correspond to pages with duplicate or near duplicate content. Also the method learns normalization rules that convert distinct URLs which refer the same content to the canonical form. DUSTER makes use of Multiple Sequence Alignment (MSA) algorithm of training URLs with different contents.

### REFERENCES


