International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

Browser Extension TO Removing Dust Using Sequence Alignment and Content Matching

Priyanka Khopkar[1], D.S.Bhosal

[1]PG Student, Ashokrao Mane group of institution, Vathar

[2]Associate Professor, Ashokrao Mane group of institution, Vathar

---

Abstract - ***If documents of two URLs are similar, then they are called DUST. Similarly, detection of near duplicate documents is complex. The duplicate documents content will be similar but there will be small differences in the content. Different URLs with same content are the source of multiple problems. Most of the existing methods generate very specific rules. So more number of rules are required to increase detection of duplicate URLs. Existing methods cannot detect duplicate url's across different sites where as candidate rules are derived from URL pairs within the dup-cluster. Existing Methods Complexity is proportional to the number of specific rules generated from all clusters. In the proposed system, the URL normalization process is used which identifies DUST with fetching the content of the URLs. In Proposed system, a new method is present, which obtains a smaller and more general set of normalization rules using multiple sequence alignment. The proposed method is used to generate rules with an acceptable computational cost even when crawling in large scale scenarios. The valid URL'S contents can be fetched from its web.***

Key Words*: *Crawlers, Dust, Uniform Resource Locator (URL).

1.INTRODUCTION

There are multiples URLs that have similar content on the web. Different URLs with Similar content are known as DUST. Since Crawling this duplicate URLs is results in poor user experience, a waste of resources, so the task of detecting DUST is important for a search engine. In the proposed method, these duplicate URLs are converted into same canonical form which can be used by web crawlers to avoid and removing DUST. In proposed system, multiple sequence alignment and URLs content matching methods are used. multiple sequence alignment is considered as a natural generalization of the pair wise alignment problem, for any given set of sequences with more than two sequences. This problem requires that all sequences to be of the same length and thus spaces are inserted at the appropriate places.

In the proposed system, to obtain a smaller and general set of rules to avoid duplicate URLs multiple sequence alignment is used. Multiple sequence alignment is used for identifying identical patterns. This Multiple sequence alignment can be used to identifying similar strings, which can be used for deriving normalization rules. More general rules can be generated using this multiple sequence alignment algorithm to remove the duplicate URLs with similar contents. After URLs normalization process, normalized URLs are sends towards URL Content Matching for further comparison, Where first efforts focused on comparing document content that inspect the URLs with fetching the corresponding page contents.

After removing duplicate urls , there is another algorithm is proposed which presents a novel and interesting problem of extracting top-k lists from the web. Compared to other structured data, top-k lists are cleaner, easier to understand and more interesting for human consumption, and therefore are an important source for data mining and knowledge discovery. We demonstrate a algorithm that automatically extracts over 1.7 million such lists from the a web snapshot and also discovers the structure of each list.

The goal of the proposed system is to detect dust and remove duplicate URLs. It is achieved by using two algorithms, the algorithm that detects the dust rules from a list and the algorithm that is used to convert these URLs into same canonical form. The contents of a valid URL can be fetched from its web server. If the documents of two URLs are similar, they are called DUST. Similarly identification of near duplicate documents is complex. The content of the near duplicate documents will be similar but there will be small differences in the content. The web pages with same content, but different URLs is the source of multiple problems. A list consists of an URL and a HTTP code. The list type can be obtained from web server logs. The algorithm to detect DUST rules is used to create an ordered list of DUST rules from a website. Canonization is the process of converting every URL into a single canonical form. There is a possibility of efficient canonization of DUST rules detected.

2.RELATED WORK

A. Agarwal, H. S. Koppula, K. P. Leela, K. P. Chitrapura, S. Garg, P. Kumar GM, C.Haty,A.Royand A.Sasturkar [1], a set of techniques have to proposed to mine rules from URLs and utilize these learnt rules for de-duplication using just URL strings without fetching the content explicitly. The rule

extraction techniques are robust as compare to web-site specific URL conventions.

Bassma S. Alsulami, Maysoon F. Abulkhair, Fathy E. Eassa[3] , The identification of similar or near-duplicate pairs in a large collection is a source of problem with wide-spread applications. Survey present a review of the existing literature in duplicate and near duplicate detection in Web.

A. Dasgupta, R. Kumar, and A. Sasturkar[4], preserving each rules for de-duplication is not enough due to the great number of rule extraction. The rule extraction methods are strong against website URL rules. The proposed system uses a set of URLs which is divided into equivalence classes that are content based. The URLs with similar content are considered in the same class. The proposed system addresses the problem of mining URL set and learning URL rewrite rules which transforms all URLs with an equivalence class to the same canonical form. The proposed system automatically generates URL rewrite rules by mining a given collection of URLs with content-based information. These rewrite rules can be useful to predict duplicate URLs that are encountered for the initial time during web crawling, without fetching their content. Such transformation rules can be suggested by a simple framework which is enough to get the most ordinary URL rewrite patterns happening on the web. The proposed system, addresses an algorithm for extracting and learning URL rewrite rules and show that under some assumptions, it is complete.

Kaio Rodrigues, M. Cristo, E. S. de Moura, and A. S. da[10] ,present DUSTER, a new approach to detect and eliminate redundant content when crawling the web. DUSTER takes advantage of a multi-sequence alignment strategy to learn rewriting rules able to transform URLs to other likely to have similar content, when it is the case.

Zhixian Zhang , Kenny Q. Zhu , Haixun Wang , Hongsong Li[5], it is concerned with information extraction from top-K web pages, which are web pages that describe top K instances of a topic which is of general interest.

H. S. Koppula, K. P. Leela, A. Agarwal, K. P. Chitrapura, S. Garg, and A. Sasturkar [9], authors proposed a set of methods which extract conventions from URLs and use these rules for de-duplication by just URL strings without fetching the content clearly. The technique is made by extracting the crawl logs and using clusters of related pages to mine detailed rules from URLs which belongs to each cluster. It presents basic and deep tokenization of URLs to mine all possible tokens from URLs which are extracted by rule generation techniques for generating normalization rules.The generated pair-wise rules are consumed by decision tree algorithm ,to reduce the number of pair-wise rules to generate precise generalized rules.The rule extraction methods are strong as compare to website URL rules.

3.PROPOSED WORK

Here designed system uses multiple sequence alignment which obtains a smaller and more general set of normalization rules. Multiple sequence alignment is a tool to identify similarities and differences among string. After removing duplicate urls , there is another algorithm is proposed which presents a novel and interesting problem of extracting top-k lists from the web. Compared to other structured data, top-k lists are cleaner, easier to understand and more interesting for human consumption, and therefore are an important source for data mining and knowledge discovery.

A. System Architecture

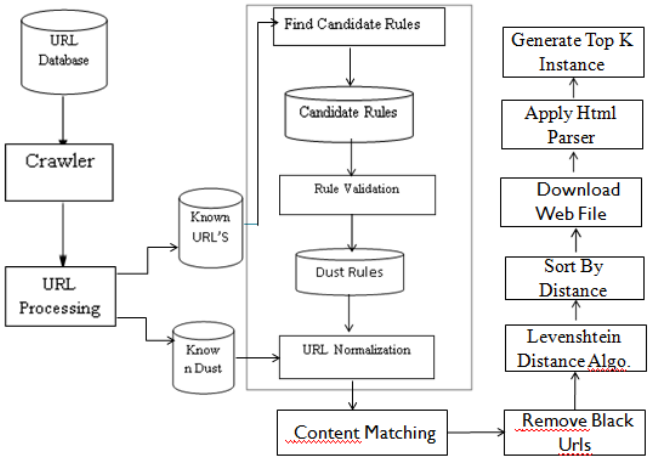

Fig 1: Architecture for Browser Extension to Remove Duplicate Url's.

The proposed architecture consists of four modules.

1. URL Dataset Visualization

The work included two sets of data to achieve data deduplication. The feasibility of data sets, both small and large data sets are used for experimentation are known. These datasets contain either the URLs of many websites or the URLs of many web pages. The pre-requisite to select these small and big data set is that it should contain at least 2 sized duplicate clusters in both data sets. Also it is characterized by the number of hosts, URLs and the duplicate clusters present in them. The collection Mixer is constructed with the data sets containing web pages and clusters. Based on the human judgment, the core content is collected.

2. Tokenization and Clustering

Basic tokenization is the process of parsing URL to extract tokens. The protocols, hostnames, query arguments and the path components are also extracted from the specified standard delimiters. Firstly, clusters are formed with the URLs in the datasets. Then, anchors are selected from the URL clusters formed in the previous step. The selected anchors are validated and if the anchors are found to be valid, then the child pattern is generated. The process of cluster formation with the URLs are known as Clustering. It is the basic step in which the cluster is formed and is produced to the rule generalization module. The URLs which consists of more similarity in the web page content is termed as a duplicate cluster. The rules are generated for all the URL pair present in the duplicate clusters.

3. Pairwise Rule Generation

Pairwise rule generation module is designed for generating pairwise rules from the URL pairs of the duplicate clusters. The transformational rules are framed in this module. This is the critical part of the work which decides the efficient working of de-duplication process. Here target URLs are used for generating transformation.

4. Levenshtein Distance

Levenshtein distance (LD) is a measure of the similarity between two Url's, which we will refer to as the source url (s) and the target url (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t.

4.SCOPE OF THE WORK

The main goal of the proposed work is to remove duplicate url,s from web.

1.To remove duplicate url,s from web using multiple sequence alignment.

2.Presents an approach that improve performance of search result and improve the speed of server.

3.Present a system to improve the speed and accuracy of recommendation system in big data application.

5. CONCLUSION

In this paper, we study Multiple Sequence Alignment and proposed a new system that uses Pairwise Rule Generation which decides the efficient working of de-duplication process. Multiple Sequence Alignment improves the performance of search engine.

6.REFERENCES

[1] A. Agarwal, H. S. Koppula, K. P. Leela, K. P. Chitrapura, S. Garg,P. Kumar GM, C.Haty, A. Roy, and A. Sasturkar, "Url normalizationfor de-duplication of web pages," inProc. 18th ACM Conf. Inf. knowl.Manage., 2009, pp. 1987–1990.

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

[3] B.S.Alsulami, M. F. Abulkhair, and F. E. Eassa, "Near duplicate document detection survey," Int. J. Comput. Sci. Commun. Netw.,vol. 2, no. 2, pp. 147–151, 2012.

4] 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.

[5] Zhixian Zhang , Kenny Q. Zhu , Haixun Wang , Hongsong Li . Automatic Extraction of Top-k Lists from the Web.

[6] G. Blackshields, F. Sievers, W. Shi, A. Wilm, and D. G. Higgins.(2010). Sequenceembedding for fast construction of guide trees for multiple sequence alignment. Algorithms Mol. Biol. [Online]. 5,p. 21.

[7] D. F. Feng and R. F. Doolittle. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. [Online]. 25(4), pp. 351–360.

[8] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform," Nucleic Acids Res. [Online].30(14),pp. 3059–3066.

[9] 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.

.[10] Kaio Rodrigues, Marco Cristo, Edleno S. de Moura, and Altigran da Silva"Removing DUST Using Multiple Alignment of Sequences"

[11] C. L. A. Clarke, N. Craswell, and I. Soboroff, "Overview of the TREC 2004 terabyte track," in Proc. 13th Text Retrieval Conf., 2004, pp. 2–3.