ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Detection and Removal of DUST: Duplicate URLs with Similar Text Using DUSTER

Jyoti G. Langhi, Prof. Shailaja Jadhav

Department of Computer, Marathwada Mitra Mandal's College of Engineering, Pune, India

Department of Computer, Marathwada Mitra Mandal's College of Engineering, Pune, India

ABSTRACT:World Wide Web is a medium commonly used to search information using Web crawlers. Some pages collected by the web crawlers contains duplicate content. Different URLs with Similar Text are generally known as DUST. The proposed method is to detect and remove duplicate documents without fetching their contents. Here the normalization rules are used to transform all duplicate URLs into the same canonical form. To improve the performance of search engines, a new method called DUSTER is used. DUSTER converts all the URLs into multiple sequence of alignments which generates candidate rules and rules validation. Here, DUSTER filters out candidate rules according to their performance in a validation set and finally removes the duplicate URLs. Using this method reduction of large number of duplicate URLs is achieved. Our contribution work is, we intend to improve the scalability and precision of our method, and to evaluate it using other datasets. For its scalability, we intend to provide a comprehensive comparison among strategies to cope with very large dup-clusters, which includes (a) to better understand the impact of using split dup-clusters instead of the original ones, (b) to propose distributed algorithms for the task and (c) to use more efficient multiple sequence alignment algorithms. Distributed processing is an effective way to improve scalability, reliability and performance of a database system. Distributed database is to be used.

KEYWORDS-Crawling, Dup-Cluster, DUSTER, Distributed Database, URL Normalization, Web Technology

I. INTRODUCTION

On the web there are different URLs that have similar content. These similar URLs are known as DUST. Duplicate URLs occur for many reasons. Detection of DUST is important task for search engine because Crawling these duplicate URLs is a waste of resources. DUST results in poor user experience. The existing system focused on document content to remove Duplicate URLs. Generation of Dynamic web pages leads to Duplication of contents. InDUSTER method, these duplicate URLs are converted into same canonical form which can be used by web crawlers to avoid DUST. The existing system uses random sampling instead of processing all URLs. In the proposed system, to avoid duplicate URLs, multiple sequence alignment is used to obtain asmaller and general set of rules. Multiple sequence alignment is used in scientifictechnologies to identify identical patterns. This Multiplesequence alignment can be used to identifysimilar strings,which can be used for deriving normalization rules. Moregeneral rules can be generatedusing multiple sequencealignment algorithm to remove the duplicate URLs with similar text. In the proposed method, the fragmentation is applied on the set of known DUST.To improve the scalability we can use the scalable web crawler. By scalable we mean that a crawler is designed to scale up the entire Web. It has been used to fetch tens of millions of Web documents. To achieve the scalability we can implement the data structures so that they use bounded amount of memory, regardless of the size of the crawl. Hence vast majority of data structures are stored on disk and small parts of them are stored in memory for efficiency.

II. RELATED WORK

DUST detection is divided into two methods, one is content based method and another one is URL based method. In content based method the whole page has to be fetched and full content has to inspect by comparing it using syntactic

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

or semantic evidence. In URL based method the duplicate URLs can be find out without examining the content of the page. In the following paragraphs some URL-based methods are focused on which reports the best results in literature.

In the base paper [1] the DUSTER framework is proposed. DUSTER detects duplicate URLs which refers to the pages with duplicate or near duplicate content. This method uses normalization rules that converts distinct URLs which refer to same content to a common canonical form. This makes easy to detect them. The scalability and precision can be improved using other data sets.

The work done in [2] focuses that the basic and deep tokenization of URLs to extract all possible tokens from URLs which are mined by Rule generation techniques proposed by them for generating normalization Rules. Proposed system implements for giving output to the user efficiently and large-scale de-duplication of documents.

A new technique SizeSpotSigs [3] is used for effective near duplicate detection algorithm considering the size of page content in mining. Proposed system implements noise-content ratio to work better.

Authors in [4] proposes top-down approach. In this paper, a new technique pattern tree based approach is used for learning URL normalization rules. In this training data set is created first. Then a pattern tree is generated on basis of training data set. After that the duplicate nodes are identified from the pattern tree. Finally,the normalization rules are generated. As these normalization rules are directly applied on pattern rather than on every URL pair, the respective computational cost is low. The proposed system is help to user select deployable rules by removing conflicts and redundancies.

The list of retrieved document contains duplicated and near duplicate results. B. S. Alsulami and other authors [5] have done a survey on Near Duplicate Document Detection. The detection of Near Duplicate Document is the problem of finding all documents fast whose similarities are equal to or greater than the threshold which is given. There are two techniques: Near duplicate prevention and Near duplicate detection. The proposed system is used in to Technical support doc management, Plagiarism Detection, Web Crawling, Digital libraries and electronic publishing, Database cleaning, Files in a file system, E-mails applications.

The first URL-based method is DustBuster [6]. This technique is used to avoid duplicate urls in crawling. A dynamic forum site, an academic site, a large news site, a small news site, the proposed system DustBuster mines dust effectively. It can reduce crawling overhead, increases crawl efficiency and reduces indexing overhead.

The authors in [7] presented machine learning technique to generalize the set of rules. These rules reduces resource footprint to be usable at web scale. The basic and deep tokenization of URLs to ex-tract all possible tokens from URLs which are mined by their Rule generation techniques for generating normalization rules. The proposed system is used to measure the performance of URL dataset on key metrics.

The SpotSigs [8], a new algorithm for extracting and matching signatures for near duplicate detection in large Web crawls. The proposed system is used to filter natural-language text passages out of noisy Web page components.

III. PROPOSED SYSTEM ARCHITECTURE

A. DESIGN CONSIDERATIONS:

1. Web crawler firstly fetches the URLs from the application server.
2. To improve the scalability we can use parallel crawler.
3. New Set of URLs are merged with already known URLs to form new set of known URLs. By following canonical tags, a new set of known DUST is also available.
4. Using Fragmentation we can split the set of known DUST.
5. Use cluster normalization rules that transform duplicate URLs into a unified canonical form and optimizes the URLs.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

6. 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.
7. Web crawler firstly fetches the URLs from the application server.
8. Similar words from the duplicated contents are extracted only if the similar words are cross the threshold value which represents the similarity.
9. Finally the duplicate URLs can be detected and removed. So that the set of original URLs be left out.

B. DESCRIPTION OF PROPOSED SYSTEM:

We first proposed the use of multiple alignments as a way to avoid the problems of simple pair-wise rule extraction. The main differences between this work and the previous one are (1) the handling of large dup-clusters; (2) the adoption of new methods for intra-cluster generalization and alignment penalization; (3) the elimination of a hierarchical clustering step with the reduction of the number of generated rules; and (4) the simplification of the algorithm, by supporting fewer kinds of tokens.Existing methods have the limitations of de-duping. In this paper, new method, called as DUSTER is proposed to overcome the limitations it uses multiple sequence alignment to obtain a smaller and general set of normalization rules. Even when crawling in large scale scenarios, this method can generate rules with an acceptable computational cost. Its complexity is proportional to the number of URLs to be aligned.

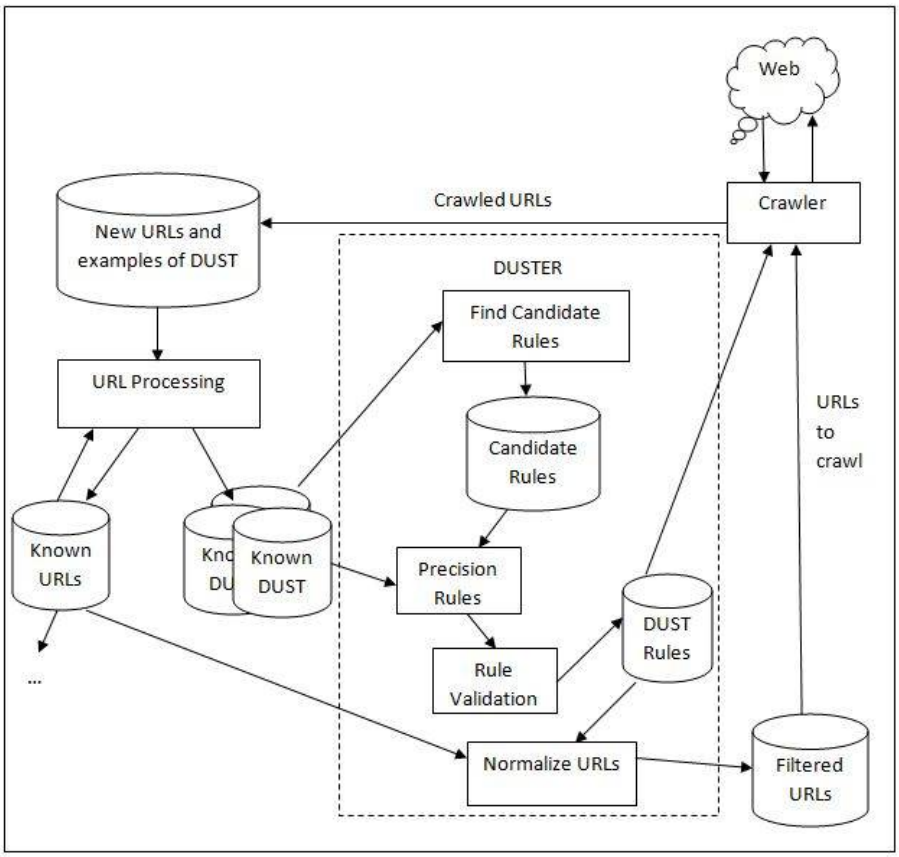

Fig. 1. Block Diagram of Proposed System (DUSTER Framework)

In this figure, first the new set of URLs is crawled then it is merged with the already known URLs, which forms a new set of known URLs. During crawling, by following canonical tags, the crawler is also able to identify examples of DUST. As a result, a new set of known DUST is also available. For the scalability we have divided the set of known

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

DUST into multiple sets using fragmentation. These sets can be still enriched by processes such as those based on content signature, followed by manual inspection. The set of known DUST is taken by DUSTER and use it to find and validate rules, by splitting it in training and validating sets. Precision Rules are also applied for precision improvement. The resulting rules are then used to normalize the known URLs yielding a new (and reduced) set of URLs to be crawled. By using this set and the set of DUST rules, the crawler can gather new URLs and close the cycle.

1) **Multiple Sequence Alignment-** Multiple sequence alignment is used to identify similarities and differences among strings/sequences. These similarities and differences can be used to determine fixed and variable substrings in URLs. It helps to derive normalization rules. As multiple sequence alignment methods find patterns involving all the available strings, the method can find more general rules and avoids problems related to pair-wise rule generation and finding rules across sites. Thus, a full multi-sequence alignment of duplicate URLs can make the learning process more robust and less susceptible to noise.

2) **Phases of DUSTER-** The proposed method DUSTER is divided in two main phases as mentioned below,

   Phase 1:Candidate rules generation: In this phase, The multi-sequence alignment algorithm is applied first in dup-clusters to align all the URLs and obtains consensus sequences for each dup-cluster. Then the candidate rules are generated from these consensus sequences. A heuristic is used to ensure the efficiency of the method for large clusters.
   Phase 2:Validating candidate rules: In this phase according to performance the candidate rules get filtered out in a validation set.

Advantages of Proposed System

1. Less time consuming: Effective crawler can avoid spending too much time crawling unproductive sites.

2. High relevant sites are prioritized: Crawler finds more relevant deep websites and consistently harvests more relevant forms.

3. Quickly discover relevant content sources: Crawler can achieve higher accuracy on finding searchable forms which has relevant content by using multiword crawling technique.

4. Optimize the search results links and give accurate results to the users.

5. Help the users to minimize the search results and provide unique results.

IV. SYSTEM ANALYSIS

The dataset used to analyze the system is WBR10. This data set is a collection of over 150 millionweb pages crawledfrom the Brazilian domain using an actual Brazilian crawlingsystem. Thiscrawling was performed from September toOctober, 2010. No restrictions were thereregardingcontentduplication or quality. Two different baseline methods areused for the comparison. Firstbaseline is the work done byDasgupta [9]. It is implemented usingR_{fanout} . The secondbaseline is themethod proposed in [4], which we refer toas R_{tree}.Table I presents the total number of rules learnedby the three methods after the training (candidate rules) andthe number of the rules ready to be usedin the test (validrules). Thus, in Table I, we compare, for $fpr_{max} = 0$, how manycandidate rules aregenerated and, out of them, how many arevalid.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

TABLE I

Number of Candidates and Valid Rules Generated by Different Methods in WBR10 dataset ($fpr_{max} = 0$)

Data Set	Method	#Candidates	#Valid	Rate
WBR10	$R_{fanout-10}$	31565	1985	6.29%
	R_{tree}	6974	1575	22.58%
	DUSTER	786	577	85.48%

Fig. 2 compares the duplicate cluster distributions after applying rules generated by all methods in both collections. In WBR10 our method reduced 449,405 clusters (63.14 percent) against 187,180 (26.30 percent) from $R_{fanout-10}$

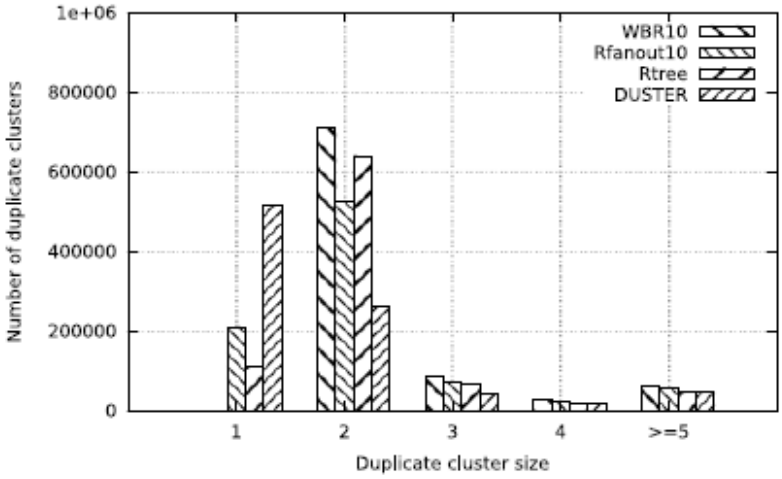

Fig.2. Duplicate cluster distribution in WBR10 using different methods and $fpr_{max} = 0$

In Fig. 3, the rules generated by DUSTER can reduce more clusters than those generated by $R_{fanout-10}$ and R_{tree}at all false-positive levels in given collection.

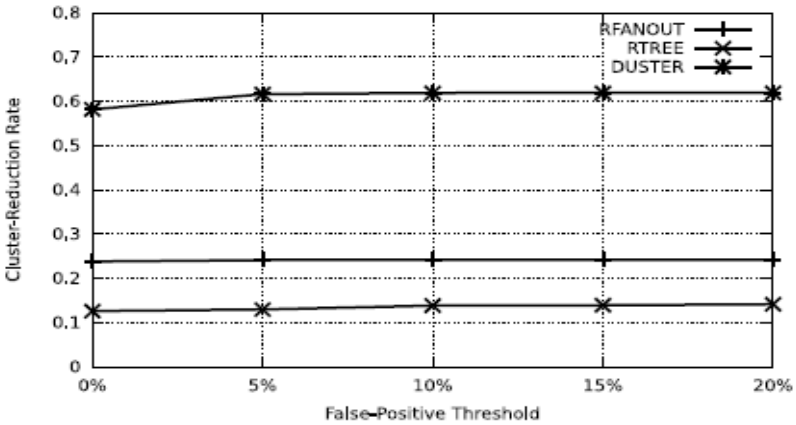

Fig.3. Comparison of Cluster-Reduction Rate in WBR10 using different methods and false-positives rate

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

V. EXPECTED OUTCOME

The proposed system DUSTER which detects the duplicate URLs with similar text (DUST) and removes duplicate URLs from the dup-cluster. Efficiency of the system will depend upon multiple sequence alignment algorithms which provide high quality and able to align n sequences in time. It is dependent on DUSTER learns normalization rules that are very precise in converting distinct URLs which refer the same content to a common canonical form, making it easy to detect them. The performance is superior of the DUSTER system. The performance is better when DUSTER generates smallest number of candidate rules and highest rate of valid rules. The method in a set of duplicate URLs extracted from the dataset collection and found a reduction in the number of duplicate URLs that is larger than the one achieved by our best baseline.The learning process more robust and less susceptible to noise.

VI. CONCLUSION

In this Project, we make an attempt to use DUSTERmethod to address the DUST problem, that is, thedetection ofdistinct URLs that correspond to pages with duplicate or nearduplicatecontent. DUSTERlearns the normalization rules forconverting distinct URLs by easily detected. The proposedsystem willbe improved for scalability and precision. Alsothe approach which we have presented, prioritizes headtrafficand we would like to explore the feasibility of rule generationfor the rest of tail traffic. Clusterswhich are the ground truthfor generating rules may include false positives due to theapproximatesimilarity measures.

REFERENCES

[1] Kaio Rodrigues, Marco Cristo, Edleno S. de Moura, and Altigran da Silva, “Removing DUST Using Multiple Alignment of Sequences.”, IEEE Transactions On Knowledge and Data Engineering,VOL. 27, NO. 8, AUGUST 2015.
[2] 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.
[3] 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] 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.
[5] 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.
[6] 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.
[7] Agarwal, H. S. Koppula, K. P. Leela, K. P. Chitrapura, S. Garg, P. Kumar GM, C. Haty, A. Roy, and A. Sasturkar, “Url normalization for de-duplication of web pages,” in Proc. 18th ACM Conf. Inf. knowl. Manage., 2009, pp. 1987–1990.
[8] 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.
[9] A. Dasgupta, R. Kumar, and A. Sasturkar, “De-duping urls via rewrite rules”, ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 186194.