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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

A Graph Based Approach for Eliminating DUST Using Normalization Rules

Jayashri Waman, Prof. Pankaj Agarkar

M.E. Student, Dept. of Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohgaon, Pune, India

Assistant Professor, Dept. of Computer Engineering, Dr.D.Y. Patil School of Engineering, Lohgaon, Pune, India

ABSTRACT: Duplicate content means search engines have to waste time in crawling all the different duplicate versions of a page, and you're relying on them to do it in the way you want them to. Duplicate content generally refers to substantive blocks of content within or across domains that either completely matches with other content or is appreciably similar. Mostly, this is not deceptive in origin. Use of such duplicated data is a waste of resources which results in poor user experiences.
We focus on removing links of duplicate contents by address of the website i.e. URL. We will convert URL into graphical format of sequences and perform the operations. Also URL tokenizer is used to understand the web protocol and top -level domain. This approach will help in a healthy way to remove same content from a set of web pages. So web crawlers can easily accept this approach and can make better indexing possible. The proposed method achieved larger reductions in the number of duplicate URLs than the existing approach.

KEYWORDS: URL normalization, De-duping, Consensus sequences, Canonical Form.

I. INTRODUCTION

The URLs which are having similar content are called as DUST (Duplicate URLs with Similar Text). Syntactically these URLs are different but having similar content. For example, in order to facilitate the user's navigation, many web sites define links or alternative paths to access a document. In addition, webmasters usually mirror content to balance web request 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 accessing the content.

Detecting DUST is an extremely important task for search engines since crawling this redundant content leads to waste of resources such as Internet bandwidth and disk storage. The DUST creates disturbance in results of link analysis algorithms and also results in poor user experience due to duplicate results. To resolve these problems, several authors have proposed methods for detecting and removing DUST from search engines. Initially efforts were focused on comparing document content to remove DUST, which was again a resource consuming process. However more recent studies proposed methods 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, including ones that are found for the first time during the crawling. The main challenge for these methods is to derive general rules with a reasonable cost from the available training sets. As observed in [6], many methods derive rules from pairs of duplicate URLs. Thus the quality of these rules is affected by the criterion used to select these pairs and the availability of specific examples in the training sets. To avoid processing large numbers of URLs, most of the methods employ techniques such as random sampling or by looking for DUST only within sites, preventing the generation of rules involving multiple DNS names. Because of these issues, current methods are very susceptible to noise and, in many cases, derive rules that are very specific. Thus, an ideal method should learn general rules from few training examples, taking maximum advantage, without sacrificing the detection of DUST across different sites.

People use search engines for searching information. But retrieved documents contains a large volume of duplicate documents. Hence there is need to improve the search results. Data filtering algorithms used by some of search engines which eliminate duplicate and partial duplicate documents to save time and effort.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Search engines in response to a user's query typically produces the list of documents ranked according to closest to the user's request. These documents are presented to the user for examination and evaluation. Web users have to go through the long list and inspect the titles, and snippets sequentially to recognize the required results. Filtering the search engines' results consumes the users' effort and time especially when a lot of near duplicate. [2]In this paper we present new approach based on graph formation using all urls from training set.

II. RELATED WORK

In this section we are presenting the different methods those are presented to eliminate duplicate URLs.

Current research on duplicate URL detection can be classified in two different methodologies: content based and URL based. In content based methods, it is necessary to download all contents of the URL, inspect it and then match it. Thus this method consumes lots of resources and to avoid such a waste of resource several URL based methods has been proposed. In the paper we are focusing on URL based methods.

Kaio Rodrigues et el [1] presented a new and novel method called DUSTER, which uses multiple sequence alignment to find similarities and differences in URLs. These similarities can be used to determine fixed and mutable substrings in URLs. These substrings helps to generated normalization rules. Thus the multiple sequence alignment is able to find more general rules than the pairwise rule generation.

Yossefet el. [3] presents DustBuster – the algorithm for discovering site-specific dust rules. It has four phases. The first phase uses the URL list alone to generate a short list of likely dust rules. The second phase removes redundancies from this list. The next phase generates likely parameter substitution rules. The last phase validates or refutes each of the rules in the list, by fetching a small sample of pages.The method discovers rules that transform a URL to other URL which is having similar content. DustBuster scans previous crawl logs or web server logs to find dust, without accessing actual page contents. However to verify these rules, the method uses sampling process which fetches few actual web pages.But it has some limitations such as, by using simple string-substitution they do not capture all possible rules and the rules generated are not generalized.

Dasgupta et al. [4] proposes Mining and learning URL Rewrite rules. Based on the contents, the proposed system uses a set of URLs to partition it into classes. So the URLs having similar content are considered in one class. The proposed system generates URL rewrite rules by mining of URLs with content-similarity data. A fixed set of delimiters are used; which are useful to study the effect of more flexible tokenization. These rewrite rules can then be applied to eliminate duplicates among URLs are found first time in web crawling. A simple framework is proposed which captures the most common URL rewrite patterns occurring on the web. It provides an efficient algorithm for mining and learning URL rewrite rules. However to the system need to be enhanced to capture a wider set of rules, while still being efficiently learnable. However this method doesn't use all samples efficiently to form general rule. Thus it is sensitive to noise.

Koppula et al. [5] extends the representation of URL and Rule presented in [4]. In this paper, authors proposed a set of techniques to mine rules from URLs and use these rules for de-duplication using URL strings. The technique is composed of mining the crawl logs and utilizing clusters of similar pages to extract specific rules from URLs which are belonged to each cluster. It presents tokenization of URLs to extract all possible tokens from URLs which are mined by rule generation techniques. Thus, to reduce the number of pair-wise rules, the generated rules are processed by decision tree algorithm which generates precise generalized rules. Since preserving each mined rules for de-duplication is not efficient due to the large number of specific rules, the proposed system presents a machine learning technique to generalize the set of rules. These generalize set of rules reduces the resource footprint. The rule extraction techniques are robust against web-site specific URL conventions.

Thus the extensions result in better utilization of the information encoded in the URLs to generate precise Rules with higher coverage. Koppula et al. propose a technique for extracting host specific delimiters and tokens from URLs. They extend the pairwise Rule generation to perform source and target URL selection. The machine learning based generalization technique generates better precision of Rules. Collectively, these techniques form a robust solution to the de-duplication problem. They presents MapReduce adaptation of the proposed techniques. The experiments shows that the proposed techniques produce 2 times more reduction in duplicates with half the number of Rules compared to [4].

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Finally, via large scale experimental evaluation on a 3-billion URL corpus, they show that the techniques are robust and scalable.

Tao Lei et al. [6] presents Pattern tree based URL normalization. In this paper, a pattern tree-based approach is proposed to learning URL normalization rules. With the pattern tree, the statistical information from all the training samples is used to make the learning process, robust and reliable. First a pattern tree is prepared based on the training set, and then normalization rules are generated by identifying duplicate nodes on the pattern tree. The learning process is accelerated as rules are directly generated based on pattern tree nodes. Thus, with the help of statistical information, the learning process is made more robust for noise and incompleteness of the training samples. The computational cost is also low as rules are directly induced on patterns, rather than on every duplicate URL pair. A graph-based strategy is used to select a subset of deployable normalization rules. The system significantly reduces the running time of the learning process.As the auto-generated training sets may contain some fake duplicate information, the proposed algorithms need to be improved.

The proposed algorithm need to be improved for scalability and precision. Also for better performance, need to use more efficient graph based algorithms.

III. PROPOSED ALGORITHM

System module having three main parts consensus sequence generation, rule generation, select valid rule.

1. Consensus sequence generation

Generating consensus sequence from each clusters of duplicate urls is done by performing URLtokenization. First we need to take a cluster and perform the tokenization of all urls available in that particular cluster. After tokenization we need to count the number of tokens from each url individually. Then arrange urls in the descending order. By taking longest i.e. first url from sorted list Graph formation get started. Once a Graph formed by containg all urls our next step is generation of consensus sequence. By reading whole graph using breadth first search we get resulting consensus sequence. As shown in fig. (1).

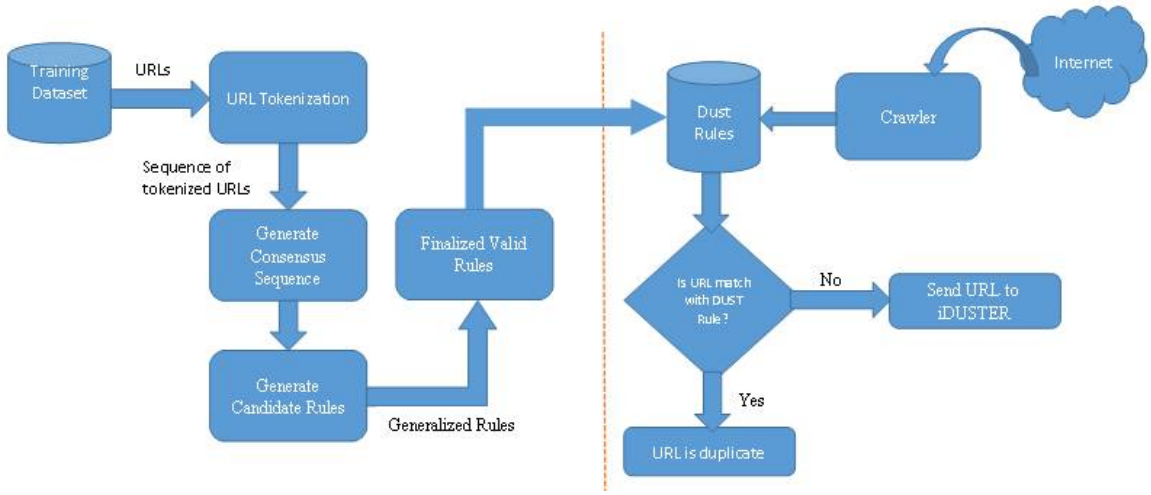

Fig.1 Graph Based approach framework (only left side)

2. Rules generation

In this part output of consensus sequence generation algorithm used to generate rules.While generating rules the tokens of consensus sequence take into consideration. According to that more generic normalization rules get formed.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

3.Selection of valid rules

In this part the rules get filtered out according to their performance in a validation set. By considering support andfalse-positive rate. Fig.1 Only left side of dotted part shows our proposed approach, while right side gives the idea of working process.

IV. PSEUDO CODE

Pseudo code for Consensus sequence generation is as follows –

Input: Cluster $c_1 = \{u_1, \ldots, u_n\}$ having n duplicate URLs.
Output: Consensus sequence for c_1.
Step 1: T = tokenize$\{u_1, \ldots, u_n\}$.
Step 2: Count number of tokens in each URL.
Step 3: Select largest tokenized URL.
Step 4: Generate a graph for all URLs.
Step 5: While generating graph consider all URL in descending order.
Step 6: V = vertices and E = edges.
Step 7: Put null for unmatched node.
Step 8: Insert node if not present previously by incrementing levels accordingly.
Step 9: Once graph is constructed read it by using breadth first search.
Step 10: Final consensus sequence get generated.

V.RESULT ANALYSIS

Normalized rules are basically used to reduce duplicate urls. By using Graph Based approach we can make use of large urls due to it's efficiency. Following graph shows the result by compairing some clusters with pairwise method. In fig. 2 it shows cluster distribution among them. And fig. 3 shows cluster reduction and false positive rate.Here for result analysis we distribute urls as training urls and testing urls. On training urls we apply algorithms and then generate rules. And test that rules on testing urls to create results.

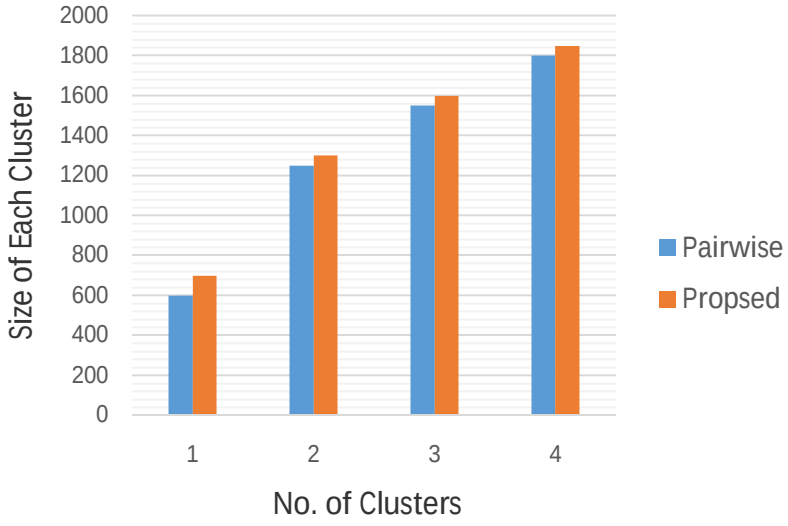

Fig 2. Graph for cluster distribution

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

As shown in fig 2 we are distributing the dup-clusters containing similar or near similar contents. Each cluster is having set of duplicate urls. Here we distribute four cluster between pairwise method and our proposed approach based on graph. Due to time efficiency we can distribute large no. of urls at proposed system. Now by applying normalization rules we find results as shown in fig. 3.Clusters get reduced efficiently by our proposed approach. As within less time we can make use of maximum urls to generate rules. So the false positive rate is reduced in proposed system. For fpr_{max}= 0% the reduction achieved by proposed method is about 29% and existing method is about 23% and so on.

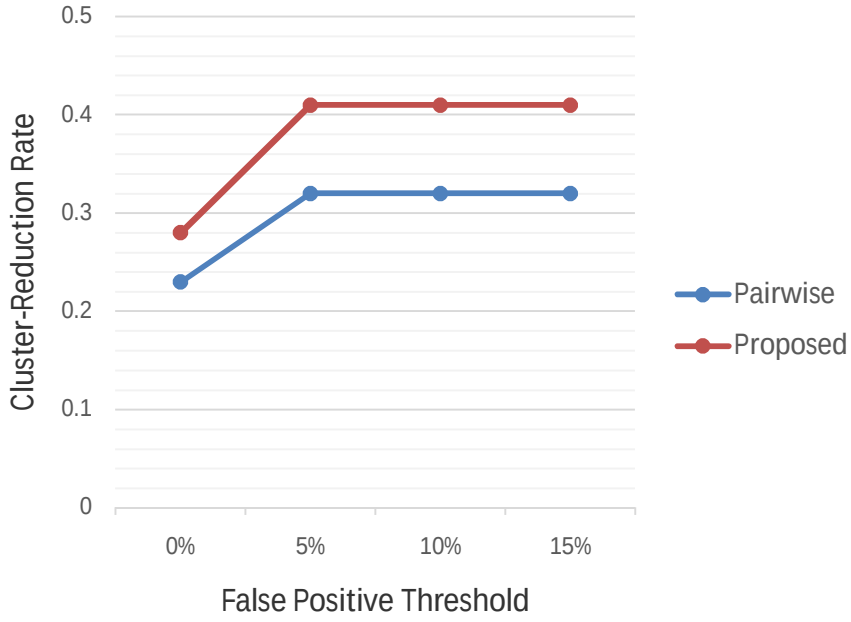

Fig. 3. Graph for cluster reduction and False Positive Threshold

Thus our both the fig.(2)&(3) shows the results which are based on the dataset taken from common crawl web site. And distributed as training set and test set within available urls. Training set urls are used to train or generate rules. After that rules are applied on test set so that duplicate urls goes into same canonical form. Thus achieved url normalization.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a new method called as Graph based approach, to address the DUST related problems. This method detects, distinct URLs that correspond to pages with duplicate or near-duplicate content. The method learns normalization rules that are very precise in converting distinct URLs which refer the same content to a common canonical form, which makes it easy to detect DUST. DUSTER uses an algorithm based on a full multi-sequence alignment of training URLs with duplicate content. It analyzes the alignments obtained and generates accurate and general normalization rules. The authors evaluated the method on two different sample sets and found a reduction in the number of duplicate URLs, generating gain of 82% and 140.74%.As future work, the proposed system can be improved for scalability and precision.

REFERENCES

1. Kaio Rodrigues, Marco Cristo, Edleno S. de Moura, Altigran da Silva, 'Removing DUST using Multiple Alignment of Sequences' IEEE Transactions on Knowledge and Data Engineering, DOI 10.1109/TKDE,2015,2407354.
2. B. S. Alsulami, M. F. Abulkhair, and F. E. Eassa, 'Near duplicate document detection survey',the Proceedings of International Journal of Computer Science and Communications Networks, 2(2):147–151, 2012.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

3. Z. Bar-Yossef, I. Keidar, and U. Schonfeld, 'Do not crawl in the dust: Different urls with similar text', ACM Trans. Web, 3(1):3:1–3:31, jan 2009.
4. A. Dasgupta, R. Kumar, and A. Sasturkar, 'De-duping urls via rewrite rules', In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08, pages 186–194, New York, NY, USA, 2008.
5. H. S. Koppula, K. P. Leela, A. Agarwal, K. P. Chitrapura, S. Garg, and A. Sasturkar, 'Learning url patterns for webpage deduplication',InProceedings of the third ACM international conferenceon Web search and data mining, WSDM '10, pages 381–390, New York, NY, USA, 2010.
6. T. Lei, R. Cai, J.-M. Yang, Y. Ke, X. Fan, and L. Zhang, ' A pattern tree-based approach to learning url normalization rules',InProceedings of the 19th international conference on World wide web,10, pages 611–620, New York, NY, USA, 2010.
7. Rahul Mahajan, Dr. S.K. Gupta, Rajeev Bedi, ' Reducing Duplicate Content Using XXHASH Algorithm', International Journal of Science and Research (IJSR) ,Volume 3 ,Issue 7, July 2014.
8. V. Rajapriya, 'A LITERATURE SURVEY ON WEB CRAWLERS', International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 5, pg. 36-44, May 2014
9. Ruchi Gupta, Dr. Pankaj Agarwal, Dr. A. K. Soni. Genetic Algorithm Based Approach for Obtaining Alignment of Multiple Sequences. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 12, 2012.

BIOGRAPHY

Ms Jayashri Wamanis a student of master's degree in Computer Science, Computer Engineering Department, Dr. D. Y. Patil School of Engineering, Pune, India. She is pursuing a Master of Computer Science (ME) degree.

Prof. Pankaj Agarkar is an assistant professor and PG coordinator at Computer Engineering Department, Dr. D. Y. Patil School of Engineering, Pune, India.He has 20 years of experience in teaching.