International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Special Issue No. 01, November 2015
www.ijates.com
ijates
ISSN 2348 - 7550

COMPARATIVE STUDY OF URL BASED APPROACHES FOR THE IDENTIFICATION OF NEAR DUPLICATE WEB PAGES.

Kavita Goyal[1], Dr. Saba Hilal[2], Dr. Jay Shankar Prasad[3]

[1]Faculty of DAVIM, Researcher at School of Computer and Information Sciences
MVN University, Palwal (India)
[2]Dept. of Computer Science, Jamia Millia Islamia, Delhi (India)
[3]School of Computer and Information Sciences, MVN University, Palwal (India)

ABSTRACT

Web is the solution of every question. When the huge amount of information is available online, it becomes easy for the user to find answer to every question just at one click. On one hand it becomes easy, but on other hand it creates problem .when web is searched, it generates duplicate links for the same query. To remove the duplicate links is required to improve the performance of search engine and to provide better results to end user.This paper studies identifies various approaches for near duplicate identification and studies one of the approach based on URL normalization.

***Keywords**: Near Duplicate Document; URL Normalization.*

I. IDENTIFICATION OF NEAR DUPLICATE WEB PAGES

a) Various Approaches

There ae various approaches for the identification and removal of near duplicate web pages This research paper studies the various approaches based on URL.

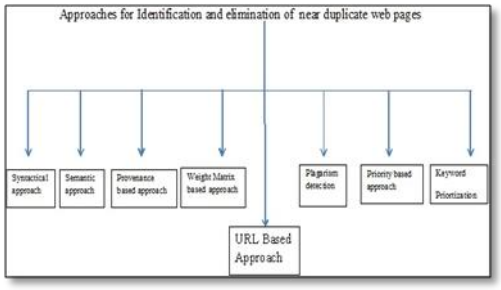

Fig 1.Various Approaches for Identification of Near Duplicate Web Pages

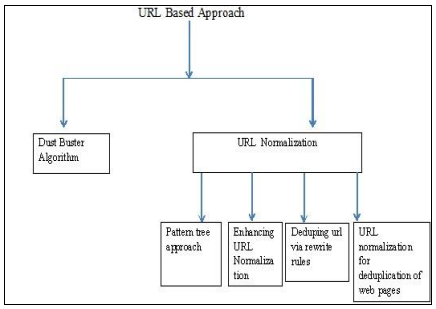

Fig 2.Various Approaches Based on URL of Apage

II. URL BASED APPROACH

URL based approach defines algorithms and approaches which defines rules to identify duplicate URLs and techniques to remove those duplicate URLs**.**The approaches it defines are:

a. Different URL with similar text

b. URL normalization.

In first different URL with similar text are identified and the latter oneis to transform duplicate URLs to a canonical form using a set of rewrite rules.

a. Different URL with similar text

In [9],Ziv BarYossef, describes a approach for finding near duplicate documents based on duplicate URLs.It is usually observed that when a query is given to search engine for searching then it generates different URLs links but those links contain same text.DUST algorithm defines rules which identifies such URLs.It improves crawling process,reduces indexing overhead.It defines two rules: One is substring substitution and the other is parameter substitution.In substring substitution replaces the **occurrence** of string in URL by other substring.In parameter substitution replaces the value of parameter in URL by some other value.

b. URL normalization

- **Pattern tree approach**
- **Enhancing metadata for url normalization**
- **Deduping URL via rewrite Rule.**
- **URL normalization for deduplication of web pages.**
- **Pattern tree approach**

In [11], a more robust and reliable technique is introduced for URL normalization. This technique is based on generating automatically a pattern tree of the website.Pattern tree removes conflicts and reduces redundancies.A pattern tree is a group

of hierarchically organized URL patterns. Each node on a pattern tree represents a group of URLs sharing the same syntax structure formation; and the pattern of a parent node characterizes all its children. In this way, a pattern tree actually provides a statistic to the syntax scheme ofa website.

- **Enhancing metadata for url normalization**

In [12] Lay-Ki Soon,suggest a technique to identify equivalent URLs by using metadata of web page in addition to standardized URL normalization method.The metadata used are page size and body text of web pages.

- **Deduping URL via rewrite Rule.**

In [13], it is stated that deduping is a serious problem because it affects whole process of crawling, indexing and searching.usually deduping is done by examining the content of URL.Here it presents different way of deduping URL.A set of URL are partitioned into a set of equivalence class on the basis of content of URL and rewrite rules are used to transform URL of same class to same canonical form. Rewrite rules are applied to the set of URL to eliminate URL that appears first time during crawling without fetching their content.

- **URL normalization for deduplication of web pages**

Deduplication mean compression technique for eliminating duplicate copies of URL.In [10], approach is described based on deduplication of URLs. It is based on fetching the crawl logs and it extracts rules from the set of URLs which generate the same content. It uses machine learning technique to generalize these rules and it also reduces resource consumption at web

III. ONLINE TOOLS FOR URL NORMALIZATION

There are number of ways in which URL can be normalize. Normalize URL by using online tools, normalize URL by some given code. Online tools which normalize URL are given below:

Given by	Tool name	Available at
NPM	url-tools	https://www.npmjs.com/package/url-tools
CPAN	URL::Normalize - Normalize/optimize URLs.	http://search.cpan.org/~toreau/URL-Normalize-0.04/lib/URL/Normalize.pm

Fig 3: Online Tools for URL Normalization

Code for URL normalization is given by oracle at

http://docs.oracle.com/javase/7/docs/api/java/net/URI.html#normalize%28%29

Google has given code to normalize URL at

https://code.google.com/p/url-normalize/

Other links are

https://en.wikipedia.org/wiki/URL_normalization

IV. RELATED RESEARCHES

The area of near duplicate detection with the problem of different URL with similar text has studied by many authors.In [5] Amit Agarwal

Et al. classified it into two areas given byBar-Yossef et al. who gave the DUST algorithm and another approach given by Dasgupta who extended the work of previous algorithm.

In [1], approaches related to different URL with same text are studied and work is done in this area only.It studied approach based on host pair with replicated content, copy detection mechanism for digital documents, syntactic clustering of web, and finding duplicated content across multiple databases.

In [2], it studies approaches for deduping which classify deduping into two main categories: Content based deduping and URL based deduping.For URL based deduping it studies DUST algorithm given by Bar Yossef and rewrite rules given by A. Dasgupta.

In [3], it studies DUST buster algorithm given by Bar Yossef. It also studies various algorithms for URL normalization given by Berners-Leethat divide normalization into three categories, syntax based, scheme based and protocol based.

V. CONCLUSION

When user search a query over search engine it gets number of results. All the results are not dissimilar and 2 to 3 links are almost similar. These links are duplicate links and many algorithms have been developed to reduce similarity. This paper has studied those approach and classified approaches in different way. The future work will concentrate on reducing similar links to achieve better results.

VI. FUTURE DIRECTIONS

This paper discussed various approaches of near duplicate detection based on URL based approach. In future a framework can be designed which automatically stores the generated URL from query. After that it compares the URL using sorting algorithm and eliminates identical URL .After that it will store Different URL in the file. Now it will compare URL based on summary content. Those URL which are different in address but generates similar summary content are discarded and remaining URL is stored in final file as a result and will be displayed to user.

REFERENCES

[1] Z.B.Yossef, I.Keidar, U.Schonfeld, “Do Not Crawl in the DUST: Different URLs with Similar Text”, in Proceedings of the 16th international conference on World Wide Web (WWW)’07, pages: 111-120, ISBN: 978-1-59593-654-7 doi>10.1145/1242572.1242588

[2] T.Lei,, R.Cai,J.Yang, Y.Ke,X.Fan,L.Zhang(2010),“A Pattern Tree-based Approach to Learning URL Normalization Rules”, in proceedings of the 19th international conference on World wide web (WWW’09), pages 611-620, ISBN:978-1-60558-799-8 doi>10.1145/1772690.1772753.

[3] L-Ki Soon, S. H. Lee (2008, Dec, 20-22), "Enhancing URL Normalization Using Metadata of Web Pages," in Computer and Electrical Engineering, 2008. (ICCEE). International Conference on, vol., no., pp.331-335, doi: 10.1109/ICCEE.2008.112

[4] A. Dasgupta, R.Kumar, A. Sasturkar (2008), “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,ISBN: 978-1-60558-193 4 doi>10.1145/1401890.1401917

[5] A.Agarwal, H. S. Koppula, K.P. Leela K. P. Chitrapura, S.Garg P.K. GM(2009), “URL Normalization for De-duplication of Web Pages”,in proceeding of 18th ACM conference on Information and knowledge management(CIKM)’09, Pages 1987-1990, ISBN: 978-1-60558-512-3 doi>10.1145/1645953.1646283