ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

Webpage Flipper with Duplicate Elimination and Clone Mining

Ratti Geeta R[1,] R.D.Kadu[2]
[1]PG Student, [2] Associate Professor SITRC Nasik, Maharashtra, India

***Abstract:** Web page Flipper with duplicate elimination and clone mining is a supervised web-scale forum crawler. The goal of this flipper is to rummage relevant forum content from the web with minimal overhead and to eliminate the duplicates. The information content in forum threads is the target of forum crawlers. Although every forum have different layouts or styles and are motorized by different forum software packages, they always have alike implicit navigation paths connected by specific URL types to lead users from entry pages to thread pages. Page type classifiers that are robust can be trained from as few as 5 annotated forums and applied to a large set of unseen forums. The expansion of Internet has also resulted in the duplication of numerous copies of web pages. The main objective to be achieved is to eliminate the duplicate web pages by using an appropriate algorithm. Duplicated web pages that consist of identical structure but have different data can be regarded as clone or cloning occurs. So the algorithm that will be used here is CloneMiner algorithm.*

***INDEX TERMS**:* **Cloning, Forum Crawling, Information Search and Retrieval, URL Pattern.**

I. INTRODUCTION

Internet forums are important platforms where users can request and exchange information with others. For example, the Trip Advisor Travel Board is a place where people can ask and share travel tips. As the information in forums is rich, researchers are more interested in mining knowledge from them. There is a need of contents to be downloaded first if one needs to gather information from forums. Generic crawlers, implement a breadth- first traversal (BFT) strategy, are usually ineffective and inefficient for forum crawling. The reason behind this is two non-crawler-friendly characteristics of forums
(1) Duplicate links & uninformative pages and
(2) page-flipping links.
A forum generally has many duplicate links that point to a common page but with different URLs. A generic crawler unknowingly follows these links will trawl many duplicate pages that make it inefficient.

II. LITERATURE SURVEY

Duplicate URLs are omnipresent in web sites, as web server software often uses aliases and redirections, and dynamically generates the same pages from various different URL requests [1]. A novel algorithm Dust Buster is proposed for exposure of DUST that is to discover rules that transform a given URL to others that are likely to have similar contents. Google a prototype of a large scale search engine which makes heavy use if the structure present in hypertext [2]. Google is being designed to crawl and index the Web efficiently to produce much more satisfactory search results than existing systems. They provided in depth description of large scale Web search engine. Google is a scalable search engine. The most important goal is to provide high quality search results over a rapidly growing WWW (World Wide Web). The analysis of numerous Web sites and Web applications is performed to evaluate and identify duplicate web pages[3]. The experiments illustrated that the proposed method detected clones among static web pages and the efficiency of the method was proved by a manual authentication. iRobot is a prototype of an intelligent forum crawler[4]. It has the intelligence to understand the content and the structure of a forum site, and then decide how to choose traversal paths among different kinds of pages. Pre sampled pages used and then decide how to select an optimal traversal path to avoid duplicates and invalids. How to design a repository for forum archiving is still an open problem.

Conventionally the de-duping problem has been addressed by fetching and examining the content of the URL [5]. Given a set of URLs partitioned into equivalence classes based on the content by addressing the problem of mining the set and learning URL rewrite rules that transform all URLs of an equivalence class to the same canonical form. Rewrite rules can be applied to eliminate duplicates among URLs that are encountered for the first time during crawling even without fetching their content. Here the use of fixed delimiters is made so this needs to be changed in future.

Internet Search Engines are posed with challenges owing to the growth of the Internet which over flow more copies of Web documents over search results making them less significant to users [6]. They recommended a method of "descriptive words" for definition of near- duplicates of documents, which was on the basis of the choice of N words from the index to determine a "signature" of a document. Marketing Intelligence is derived through an interactive analysis framework uniquely configured to leverage that connectivity and content of annotated online discussion [7]. The system delivers both qualitative and quantitative accounts of features derived from online messages. A machine learning technique was used to generalize the set of rules, which reduced the resource footprint to be usable at web-scale [8]. Rules included false positives due to the approximate similarity measures. There is a need to explore ways of handling this in a robust fashion. Generalization was performed separately for source and target and also there is a need to explore the feasibility of generalizing both in a iterative fashion.

Partial alignment [9] means that align only those data fields in a pair of data records that can be aligned with certainty. This approach results in precise alignment of multiple data records. Empirical results using a large number of Web pages show that the new two step technique can segment data records and extract data from them very accurately. A focused crawler traverses the Web to collect documents related to a particular area, and can be used to build area specific collection of documents for use in digital libraries and domain specific search[10]. Breath first search (BFS) method is being used by general crawlers to traverse the Web for as much amount of information required. Focused crawler help the search indexer to index all documents present on the World Wide Web linked to a specific domain which in turn provides search engine's users complete and fresher most information.

III. IMPLEMENTATION DETAILS

A Design

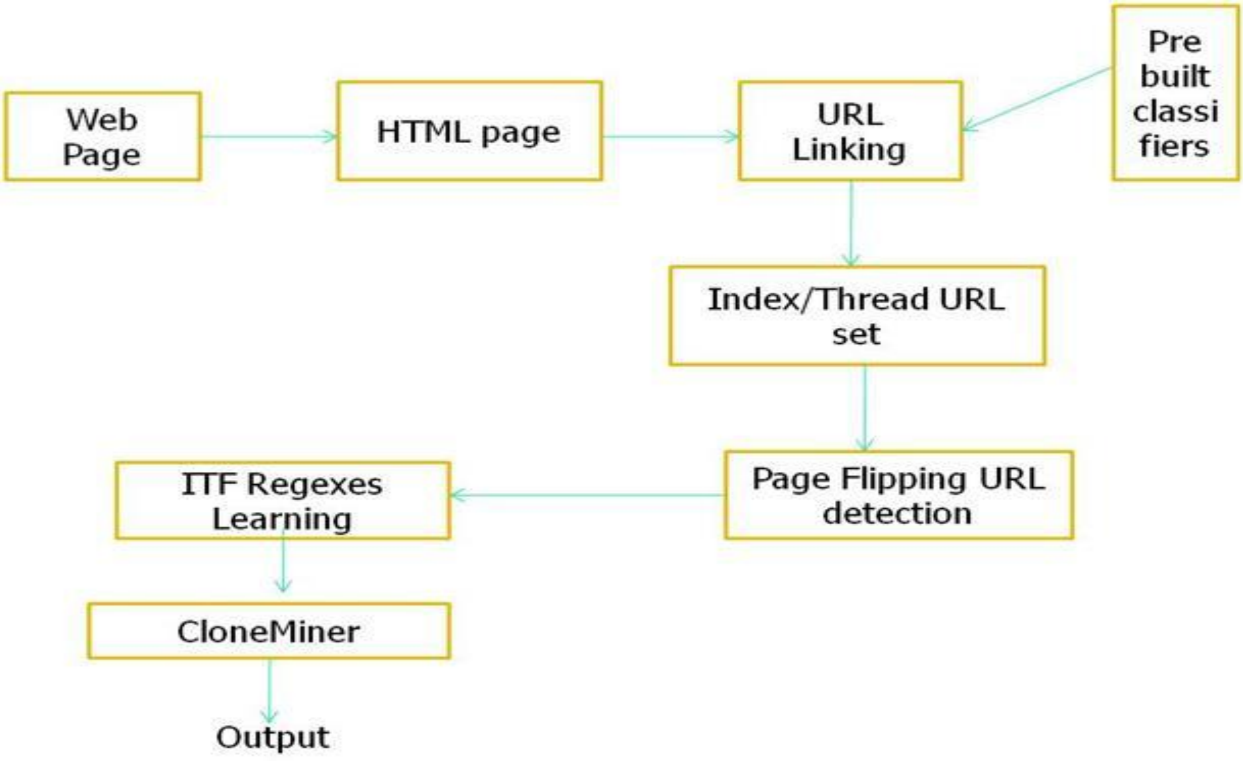


Fig1: A Block diagram of Web page flipper with duplicate elimination and clone mining

B. Block Diagram

Figure A shows the block diagram of Web page flipper with duplicate elimination and clone mining. It consists of two major parts: the learning part and the online crawling part. The learning part learns ITF regexes of a given forum from automatically constructed URL examples. The online crawling part applies learned ITF regexes to crawl all threads efficiently. Given any page of a forum, Web page flipper with duplicate elimination and clone mining first finds its entry URL using *Entry URL Discovery* module. Then, it uses the *Index/ThreadURL Detection* module to detect index URLs and thread URLs on the entry page; the detected index URLs and thread URLs are saved to the URL training set. Next, the destination pages of the detected index URLs are feed to this module again to detect more index URLs and thread URLs until no more index URL detected. After that, the *Page-Flipping URL Detection* module tries to find page-flipping URLs in both index pages and thread pages and saves them to the training set. Finally, the *ITF RegexesLearning* module learns a set of ITF regexes from the URL training set. Web page flipper with duplicate elimination and clone mining performs online crawling as follows: it first pushes the entry URL into a URL queue; next it fetches a URL from the queue and downloads its page, and then pushes the outgoing URLs that are matched with any learned ITF regex into the URL queue. This step is repeated until the URL queue is empty.

ITF Regexes Learning

To learn ITF regexes, Web page flipper with duplicate elimination and clone mining adopts a two-step supervised training procedure. The foremost step is training set construction. The next step is regex learning.

Constructing URL Training Set

The goal of training set construction is to automatically create sets of highly precise index URL, thread URL, and page-flipping URL string samples for regex learning. We use a similar procedure to construct index URL and thread URL training sets since they have very similar properties except the types of their destination pages; we present this part first. Page-flipping URL strings have their own specific properties which are different from properties of index URL and thread URL strings; we present this part later.

Index and Thread URL String Training Sets (I/T URL STS)

Recall that an index URL is a URL that is on an entry page or index page; and its destination page is another index page; while a thread URL is a URL that is on an index page; and its destination page is a thread page. It is to be noted that the only way to distinguish index URLs from thread URLs is the type of their destination pages. Therefore, method is needed to decide page type of a destination page.

The index page and thread page have their own typical layouts. Usually, an index page has many thin records, relatively long anchor text and short plain text; while a thread page has a few large records. Each post has a very long text block and relatively short anchor text. In addition, each record in an index page or a thread page usually has a link pointing to a user profile page.

Page-Flipping URL String Training Set

Page-flipping URLs point to index pages or thread pages but they are very different from index URLs or thread URLs. URLs have following properties:
1) Their anchor text is either a sequence of digits such as 1, 2, 3, or special text such as “last”;
2) Their appearance is at the same location in the DOM tree of their destination pages as in their source page;
3) Their destination pages have similar layout with their source page. Tree similarity is used to determine whether the layouts of two pages are similar or not; Our page-flipping URL detection module works based on above properties.

ITF Regexes Learning

It is shown how to create index URL, thread URL, and page flipping URL string training sets; next explained is how to learn ITF regexes from these training sets.
Each pattern matches a separation of URLs. These patterns are refined recursively until no more specific patterns could be generated. These patterns are final output as they cannot be refined further.

Online Crawling

Given a forum, web page flipper with duplicate elimination and clone mining first learns a set of ITF regexes it performs online crawling using a breadth-first strategy. It first pushes the entry URL into a URL queue; next it fetches a URL from the URL queue and downloads its page; and then it pushes the outgoing URLs that are matched with any learned regex into the URL queue. Web page flipper with duplicate elimination and clone mining repeats this step until the URL queue is empty or other conditions are satisfied.

Entry URL Discovery

An entry page needs to be specific to start the crawling process. In practice, especially in web-scale crawling, manual forum entry page annotation is not practical. Forum entry page discovery is not a trivial task since entry pages vary from forums to forums.

C. Algorithm

1. Index /Thread URL detection:

This algorithm is used to detect the index or thread URLs. The input to this algorithm is a web page. This page may be any HTML page and will be generated by using DOM tree. URL linking will be done with the help of pre-built classifiers. Then this algorithm will be used to identify the index URL and the thread URL.

2. Page Flipping URL detection:

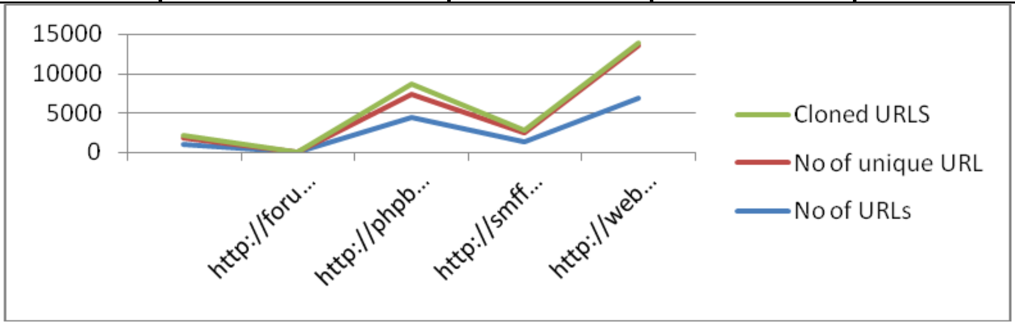
The input to this algorithm is a index URL or thread URL. The page flipping algorithm works with the help of ITF regex that we get the regular expression so that we can jump from one link to the other.

3. Clone Miner Algorithm:-

This algorithm is used to detect structural clones. Larger granularity similarities or duplicates are known as structural clones. This algorithm has its own token based uncomplicated clone detector. The process starts from uncomplicated clones. Similar to the uncomplicated clone sets (SimpleCSets), method clone sets (MCSets at level 3), file clone sets (FCSets at level 5) and directory clone sets (DCSets at level 7), which consist of groups of cloned entities at successively higher levels of abstraction. (SimpleCSets). The output from certain uncomplicated clone detectors is in the form of clone pairs. However, we can easily form SCSets by grouping clone pairs in such a way that every member in a set is a clone of every other member. Structural clone detection technique works with the information of uncomplicated clone. Clone Miner uses Repeated Tokens Finder (RTF) , a token based uncomplicated clone detector.RTF tokenizes the input URL into a token string and then search is made to find whether it is duplicate or not i.e. clone already available.

IV. RESULTS

Sr. no	URL Name	No. of URL's	No. of unique URL	Cloned URL'S
1	http://forums.asp.net/	1026	822	204
2	http://phpbb.com/blog/438	4389	2960	1369
3	http://smfforum.com/	1398	1114	284
4	http://webhostingblog.com/	6932	6652	280



Here the result is analysed by considering thread URL in the forum web page. It is done by constructing URL training set. Training set is constantly called on arrived thread URL string to check to clone matching. The exactness is upto 75% using thread URL .Later 2 more training set for index URL and page flipping URL will be added to get 99% accuracy.

V. CONCLUSION AND FUTURE SCOPE

A. Conclusion

The crawling algorithm implemented using Web page flipper with duplicate elimination and clone mining have estimated to get results faster as compared to generic way. A data duplicate from URL is avoided and it saves web page users time avoiding clicking on same URL multiple times.

Clones are detected by clone miner algorithm and it will give extra accuracy in removing duplicates and it will also save space required for forum web page or article contents on the server.

B. Future Scope

After completion of the system the integration with search engine results will reduce the duplicates in the result. Output compatible with mobile web browser will be effective for smart phone browsers which are used now a day.

ACKNOWLEDGEMENT

I am grateful to our Head of the Department for valuable support and guidance given timely. Also a grateful thanks and appreciation to faculty members for extending their kind cooperation towards us. I am also deeply indebted to our technical staff for their constant support and valuable suggestions. I owe an obligation of gratitude and sincere thanks to all my friends for useful evocation. Parent’s moral support, confidence nourished as with the determination to work hard leading my project work to gigantic triumph.

REFERENCES

[1] Z. Bar-Yossef, I. Keidar, and U. Schonfeld.” Do not crawl in the DUST: different URLs with similar text”. In Proc. of 16thWWW, pages 111-120, 2007.

[2] S. Brin and L. Page. The Anatomy of a Large-Scale Hyper textual Web Search Engine. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998.

[3] Di Lucca, G.A., Di Penta , M., Fasolino, A.R., 2002.” An Approach to Identify Duplicated Web Pages,” Proceedings of the 26th Annual International Computer Software and Applications Conference, pp:481-486.

[4] R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang.iRobot: An Intelligent Crawler for Web Forums. In Proc. of 17thWWW, pages 447-456, 2008.

[5] A. Dasgupta, R. Kumar, and A. Sasturkar.De-duping URLs via rewrite rules. In Proc. of 14th KDD, pages 186-194, 2008.

[6] Ilyinsky,S., Kuzmin,M., Melkov, A., Segalovich,I., 2002. “ An Efficient method to detect duplicates of Web Documents with the use of inverted index”, Proceedings of the Eleventh International World Wide Web Conference.

[7] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo.Deriving Marketing Intelligence from Online Discussion. In Proc. 11th SIGKDD, pages 419-428, 2005.

[8] H. S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg and A. Sasturkar. Learning URL Patterns for Webpage De-duplication. In Proc. of 3rd WSDM, pages 381-390, 2010.

[9] Y. Zhai and B. Liu. Structured Data Extraction from the Web based on Partial Tree Alignment. IEEE Trans. Knowl. Data Eng., 18(12):1614−1628, 2006.

[10] Mukesh Kumar, RenuVig .Focused Crawling Based Upon Tf-Idf Semantics and Hub Score Learning Journal Of Emerging Technologies in Web Intelligence, Vol. 5, No. 1, February 2013.