IJASRM International Journal of Advanced Scientific Research and Management, Volume 4 Issue 5, May 2019
www.ijasrm.com
ISSN 2455-6378

Near Duplicate Web Page Detection for Efficient Web Crawling: A Survey

S. S. Bhamare

School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University Jalgaon (M.S) India.

Abstract

The immense quantity of information in the World Wide Web, content mining gives lists to the search engines for the sake of the relevance to the keywords. Web content mining is used to discover the knowledge from Web page content. Major search engines takes keywords as input for searching web page information form web. To increase the performance of search engines for searching relevant web pages only, it is essential to detect and eliminates duplicates and near duplicates web pages from web. Duplicate and near duplicate content is content that appears in more than one place on the Internet. This paper presents an extensive survey on different existing duplicates and near duplicates page detection algorithms and methods that helps efficient web crawling.

Keywords: *Web crawling, Search engine, Web Page, Near duplicate page.*

1. Introduction

World Wide Web is the global network that provides huge number of web pages contains quality information. Exploiting the information resources and turning them into useful knowledge available to concerned people is a great challenge, Currently, Web mining has become important for people to collect, analyze and spread information with the fast expansion of Internet. Due to the enormous size of WWW, manual browsing is difficult for Web users. To overcome this problem, Web mining is proposed to automatically locate and retrieve web information from WWW and determine implicit knowledge for Web users. Performing the task of Web Mining, search engines plays important role to provide required relevant web pages from web through web crawling. In this huge size of web, duplicates and near duplicates web pages are presents and that harm web crawling performance to crawling of relevant web pages as per user request and affect the performance of web mining task. This paper discuss major existing duplicates and near duplicates web page detection and elimination algorithms and methods.

2. Duplicate and Near Duplicates Web Pages

In rapid expansion of internet, World Wide Web contains multiple copies of similar content. As per the survey near about 30 to 35 % of the web pages are duplicates and near duplicates on web and this redundancy of web pages are increased day by day. Due to this redundancy of web pages major search engines are going to index the redundant information of user requests and it harm the performance of searching. So it is necessary that search engines must try to avoid similar content of multiple copies in indexing and help to perform efficient web mining tasks. Web has web pages in different form that increased redundancy such as mirror copy of site or web pages, old and new versions of web pages of their content are identical etc. To improve the performance of search engines web crawler, it is must to detect near duplicates web pages before processing. Duplicates and near duplicates web page detection is a preprocessing task to detect and eliminates duplicates and near duplicates web pages.

3. Major Techniques and Methods Adopted

Near duplicates and duplicates web page detection and elimination is a major challenge to the researcher for effective web crawling by search engines. Many researchers worked on this topic and implements different techniques and methods that helps near duplicates and duplicates web page detection and elimination.

Di Lucca & et al (2003) proposed an approach Web Applications (WAs) to detect duplicated web pages. This proposed approach is worked based on similarity metrics, and addresses the detection of

clones made up of client or server static pages of the WA. It was implemented with HTML language and ASP technology. Special features of HTML and ASP are being used in the computation of metrics, and for defining a distance measure between WA pages. To determine the similarity degree of the pages the distance measure is used: two pages will be considered as clones if they are characterized by the same values of the defined metrics (i.e., their distance is zero). This approach requires the WA pages to be statically analyzed, and the pre-defined features of the pages to be extracted. To perform the experiments, a prototype tool has been developed that automatically compute the distance between pages. [3]

SimFinder algorithm was proposed by Gong C & et al (2008), this algorithm work fast to detect all near-duplicates in large-scale short text databases. The weighting scheme ad-hoc term is used by SimFinder algorithm to measure each term's discriminative capability. A certain number of terms with higher weights are selected as features for each short text. For each text SimFinder algorithm generates several fingerprints, and only texts with at least one fingerprint in common are compared with each other. An optimization method is hired in SimFinder to make it more effective. This experiment shows that SimFinder is an effective solution for small text duplicate detection with almost linear time and storage complexity. [4]

A hybrid approach is introduced by Cihan Varol & et al (2015) to improve near-duplicate detection algorithm, called as shingling algorithm. In this, they combine pattern matching technique and statistical results on the frequency of usage of words to improve the performance of the shingling algorithm for improve near-duplicate detection algorithm. [5]

Two approaches are proposed by Shiva KN & et al (1998) to compute near-duplicates simultaneously between all web documents. These two approaches assume that two documents i.e. Ti and Tj can be near-duplicates only when documents Ti and Tj share more than 'k' fingerprints, where 'k' is a predefined threshold. [6]

Shingling algorithm is proposed by Broder & et al (2000), this algorithm keeps a sketch of shingles of separate document to compute the similarity between two documents. Any documents with at least one common shingle are examined and checked to see if they exceed the threshold for similarity. Broder's shingling method is implemented in the AltaVista search engine for duplicate document detection. [7]

Chowdhury A. & et al (2002) was proposed IIT-Match or I-Match technique in which the I-Match algorithm maps each individual document into a single hash value. Two documents are considered near-duplicates if their hash values are exactly the same. I-Match does not ignore small documents and places each document into at most one duplicate set. I-Match algorithm increases accuracy and usability. [8]

Charikar M. & et al (2002) focus on sketching algorithms for estimating similarity, i.e. the construction of functions that produce concise sketches of objects in a collection, such that the similarity of objects can be estimated efficiently from their sketches. In this, sim(x, y) is a similarity function that maps objects pairs x, y to a number in [0, 1], the degree of similarity is measured between x and y. sim(x, y) = 1 corresponds to objects x, y that are identical while sim(x, y) = 0 corresponds to objects that are not identical. [9]

G. S. Manku & et al (2007) shows that Charikar's simhash [9] is practically useful for identifying near-duplicates in web documents belonging to a multi-billion page repository. Fingerprinting technique simhash likes the property that fingerprints of near-duplicates differ in a small number of bit positions. This experiment validate that for 8 billion webpages repository, 64-bit simhash fingerprints and k = 3 are reasonable. They also develop a method to solve the Hamming Distance Problem. This technique is useful for both online queries (single fingerprints) and batch queries (multiple fingerprints). [10]

V.A.Narayana & et al (2009), introduced a new and efficient approach to identify near duplicate Web pages in web crawling of search engines. It perform the task detection of near duplicate Web pages first before storing the crawled Web pages in to search engine index or repositories. It calculate first similarity score between two pages based on keywords extracted from crawled web pages. Those web page document having similarity scores is more than a threshold value are considered as near duplicates web document. This detection of near duplicates web pages has reduced storage memory for search engine index or repositories and improved the performance of search engine. [11]

Ahmad M. Hasnah (2006) was proposed a new algorithm of duplicate data removal that can detect and eliminate redundant or duplicated data that are display on the search screen as a result of user's queries. This algorithm helps to reduce the size of storage data. It keeps only relevant data to discover the knowledge. In addition, this algorithm also applied on expert systems to reduce the number of production rules that stored in the database of system knowledge with the capacity to regenerate the original set of rules whenever needed. This algorithm is based on formal concept analysis, which has been developed by many research scholars in the world for different scientific applications. Generally near duplicate document detection techniques are partitioned into three main categories: shingling techniques, similarity measures calculations and document images. [12]

Midhun Mathew & et al (2011) proposes a new approach for finding near duplicates pages of an

input web-pages, from a huge database. It works on the semantic structure and content and context of a web page rather than the only content. The weighting scheme recommended in [19] is used for creating a term document-weight matrix (TDW) which plays an important role in the proposed algorithm. Here researcher present a three-stage algorithm. This algorithm accepts an input record and a threshold value and returns an optimal set of near-duplicates. In first phase, i.e. rendering phase, in this all pre-processing are done and applied a weighting scheme. Then a global ordering is done to form a term-document weight matrix. In second phase, i.e. filtering phase, in this two well-known filtering mechanisms are applied, i.e. prefix filtering and positional filtering [14], to minimize the size of competing record set and hence to minimize the number of comparisons. In third phase, i.e. verification phase, in this they applied singular value decomposition and a similarity checking is performed based on the threshold value and lastly to get an optimal number of near-duplicate records. [13]

To detect near duplicates data Xiao C. & et al (2008) proposes a new algorithm i.e. similarity join algorithms with applications. They suggest a positional filtering principle, this filtering principle exploits the ordering of tokens in a record and leads to upper bound estimates of similarity scores. They also demonstrate that, the existing prefix filtering method and can work on tokens both in the prefixes and the suffixes as complementary. To conduct a broad experimental study using several real datasets, and show that the proposed algorithms outperform by previous ones. This new algorithm also show that, algorithm can be adapted or joint with existing approaches to produce better results or also improve the runtime efficiency in detecting near duplicate Web pages.[14]

Lavanya Pamulaparty & et al (2013), was also discussed major algorithms are used detection and elimination of near duplicates web pages such as Shingling, SPEX, Simhash, I-Match and Fuzzy Fingerprinting.[15]

An effective method is proposed by J. Prasanna Kumar & et al (2013) for detection of near duplicate web page which is based on a similarity measure using fingerprint. The main aim of this research work is to successfully avoid the near duplicates pages for better conservation of network bandwidth and minimize the storage costs. In this approach, at first the preprocessing tasks is performed. Preprocessed of crawled web pages using a parsing technique to remove the web page HTML tags, only java scripts and other irrelevant data present in the web pages, then after to remove the stop words from the crawled web pages. Finally, the stemming process can be carried out, in stemming words are converted to base form and keywords are extracted from the crawled Web pages for more processing. Then, cascade filtering is used for sentence level extraction and fingerprint as the filtering techniques. Finally, to calculated a bit-by-bit difference measure between the fingerprint of the crawled Web page and other web pages. If the computed bit-by-bit difference measure value is less than a predefined threshold value then the crawled web page is called as a near-duplicate web page. Else, the crawled web page is added to the repositories or database. [16]

4. Research Issues and Challenges

Near duplicates and duplicates web page detection is not an independent research topic because the near duplicates and duplicates web page detection is preprocessing task of web crawling of search engines to reduced redundant web data and provide relevant web data and perform efficient web mining tasks. As we see the size of the web is very large and is still increasing day by day. The web data has also increased. The different existing techniques and methods was proposed for near duplicated web page detection. Researchers are faced many difficulties while developing such techniques due to change in nature of web such as,

- The web pages do not have similar structure.
- The web information is updated very fast.
- The web information or data is in heterogeneous form.
- The web page information is linked. (Web pages are links within a site and also across different sites.)
- The web information is redundant

We see web is dynamic and information on the web changes continually and tracking the changes is the most important issue and challenge to the researcher. Researcher must consider these issues and challenges to develop a new improved techniques and methods for near duplicates web page detection and elimination.

5. Conclusion

Web crawling performance of search engines is slow down due to serious problem of duplicate and near duplicate web pages. Search engine crawled web pages, which includes near duplicates web pages and store in repositories or database. This web crawled information is less relevant to the user search request. It is essential to detect duplicate and near duplicate web pages for efficient web crawling, there has been number of algorithms or techniques was developed by researchers. This paper presented a broad survey on important existing near-duplicate document detection algorithms for effective web crawling. There is also significant scope for future work and experimentation for research scholars to design and implement improved techniques or algorithms for near-duplicate web page detection.

References

[1] Bing Liu, Web Data Mining (Exploring Hyperlinks, Contents, and Usage Data), Springer.
[2] Yi, L., Liu, B., Li, X., Eliminating noisy information in web pages for data mining, In Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2003, pp. 296 – 305.
[3] Di Lucca, G. A., Di Penta, M., Fasolino, A. R., An Approach to Identify Duplicated Web Pages, In Proceedings of the 26th Annual International Computer Software and Applications Conference, 2002, pp: 481- 486.
[4] Gong C., Huang Y., Cheng X., Bai S., Detecting Near-Duplicates in Large-Scale Short Text Databases, In Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 5012, 2008, pp. 877-883.
[5] Cihan Varol, Sairam Hari., Detecting near-duplicate text documents with a hybrid approach, Journal of Information Science 2015, Vol. 41(4) 405–414.
[6] Shiva KN and Garcia-Molina H., Finding near-replicas of documents on the web. Proceedings of workshop on Web databases, Valencia, Spain, March 1998, pp. 204–212.
[7] Broder A., Identifying and filtering near-duplicate documents, In Proceedings of the 11th annual symposium on combinatorial pattern matching, Montreal, Canada, June 2000, pp. 1–10.
[8] Chowdhury A, Frieder O, Grossman D and McCabe MC., Collection statistics for fast duplicate document detection. ACM Transactions on Information Systems 2002; 20: 171–191.
[9] M. Charikar., Similarity estimation techniques from rounding algorithms. Proceedings 34th Annual Symposium on Theory of Computing (STOC 2002), pages 380–388, 2002.
[10] G. S. Manku, A. Jain, and A. D. Sarma., Detecting near-duplicates for web crawling, ACM WWW'07, 2007, pages 141–150, NY, USA.
[11] V.A.Narayana, P. Premchand and A. Govardhan, A Novel and Efficient Approach For Near Duplicate Page Detection in Web Crawling., IEEE International Advance Computing Conference, March 6-7, 2009.
[12] Ahmad M. Hasnah., A New Filtering Algorithm for Duplicate Document Based on Concept Analysis, Journal of Computer Science, Vol. 2, No. 5, 2006, pp. 434-440.
[13] Midhun Mathew, Shine N Das, Pramod K.Vijayaraghavan., A Novel Approach for Near-Duplicate Detection of Web Pages using TDW Matrix., International Journal of Computer Applications, April 2011, pp :16-21.
[14] Xiao C., Wang W., Lin X., et al., Efficient Similarity Joins for Near Duplicate Detection., In Proceeding of the 17th international conference on World Wide Web, 2008, pp: 131-140.
[15] Lavanya Pamulaparty, Dr. M. Sreenivasa Rao, Dr. C. V. Guru Rao., A Survey on Near Duplicate Web Pages for Web Crawling, International Journal of Engineering Research & Technology (IJERT) , September – 2013, ISSN: 2278-0181 Vol. 2 Issue 9.
[16] J. Prasanna Kumar, P. Govindarajulu., Near-Duplicate Web Page Detection: An Efficient Approach Using Clustering, Sentence Feature and Fingerprinting., International Journal of Computational Intelligence Systems, Vol. 6, No. 1, January, 2013, pp. 1-13
[17] Y. Syed Mudhasi et al., Near-Duplicates Detection and Elimination Based on Web Provenance for Effective Web Search., International Journal on Internet and Distributed Computing Systems (IJIDCS) Vol: 1 No: 1, 2011 http://www.ijidcs.org/issues/v1n1/ijidcs-4.pdf
[18] Ranjna Gupta, Neelam Duhan, A.K. Sharma and Neha Aggarwal., Query Based Duplicate Data Detection on WWW., International Journal on Computer Science and Engineering, Vol. 2, No. 4, 2010.pp: 1395-1400.
[19] Shine N Das, Midhun Mathew, Pramod K.Vijayaraghavan., An Approach for Optimal Feature Subset Selection using a New Term Weighting Scheme and Mutual Information., In Proceeding of the International Conference on Advanced Science, Engineering and Information Technology, Malaysia, January 2011, pp 273-278.
[20] Akansha Singh., Faster and Efficient Web Crawling with Parallel Migrating Web Crawler., IJCSI International Journal of Computer Science Issues.