A Survey of Duplicate And Near Duplicate Techniques

Rahul Mahajan, Dr. S.K. Gupta, Mr. Rajeev Bedi

ABSTRACT ---- World Wide Web consists of more than 50 billion pages online. The advent of the World Wide Web caused a dramatic increase in the usage of the Internet. The World Wide Web is a broadcast medium where a wide range of information can be obtained at a low cost. A great deal of the Web is replicate or near- replicate content. Documents may be served in different formats: HTML, PDF, and Text for different audiences. Documents may get mirrored to avoid delays or to provide fault tolerance. The problem of finding relevant documents has become much more prominent due to the presence of duplicate data on the WWW. This redundancy in results increases the users' seek time to find the desired information within the search results, while in general most users just want to cull through tens of result pages to find new/different results. This survey paper has a fundamental intention to present an review of the existing literature in duplicate and near duplicate detection of general documents and web documents in web crawling.

Index Terms -- Duplicate Content, De-duplication, Near Duplicate , Replicate, Search Engine, Web Crawling, Web Mining

1. Introduction

With the drastic development of World Wide Web (WWW) information is being accessible at the finger tip anytime anywhere through the massive web repository. The Web is now the primary source of information for many people [7].Over 80% of Web searchers uses Web search engines to locate online information or services [17]. The voluminous amount of web documents has resulted in problems for search engines leading to the fact that the search results are of less relevance to the user. In addition to this, the presence of duplicate and near-duplicate web documents has created an additional overhead for the search engines critically affecting their performance [9]. Standard check summing techniques can facilitate the easy recognition of documents that are duplicates of each other (as a result of mirroring and plagiarism).

- ***Rahul Mahajan*** *is currently pursing M.tech from Beant College of Engineering and Technology, Gurdaspur, Punjab, India Email: rmahajan19@gmail.com*
- ***Dr. S.K.Gupta****, Head & Associate Professor (Computer Science Engineering Department), at Beant College of Engineering and Technology Gurdaspur, Punjab, India. Email: skgbcetgsp@gmail.com*
- ***Mr. Rajeev Bedi*** *, Assistant Professor(Computer Science & Engineering Department) at Beant College of Engineering and Technology, Gurdaspur, Punjab, India. Email: rajeevbedi@rediffmail.com*

The efficient identification of near duplicates is a vital issue that has arose from the escalating amount of data and the necessity to integrate data from diverse sources and needs to be addressed. Though near duplicate documents display striking similarities, they are not bit wise similar [20]. Web search engines face considerable problems due to duplicate and near duplicate web pages. These pages enlarge the space required to store the index, either decelerate or amplify the cost of serving results and so exasperate users. Thus algorithms for recognition of these pages become inevitable [17]. Due to high rate of duplication in Web document the need for detection of duplicated and nearly duplicated documents is high in diverse applications like crawling [10], ranking [21], [24], clustering [26], [7], [11], archiving and caching.

2. DUPLICATE DOCUMENT AND NEAR DUPLICATE

In any web search, it is fairly likely that some documents that are returned are very similar. For example, the same documents may exist on several servers or in several users' directories. One very frequent example of this is the Java API documents from Sun which are found on almost every Java developer's machine. Since these are very well known and described, they are very easy to eliminate using any of the existing techniques. However, more difficult cases occur when there are several versions of the same document on various servers. The same document is found in several forms, such as HTML and PDF. Duplicate document analysis is carried out only when both of the following conditions are true:

• The collection employs the link-based ranking model. This model is applicable to crawlers that crawl Web sites like the Web crawler or Web Sphere Portal crawler.

• Collection-security is disabled.

When two documents comprise identical document content, they are regarded as duplicates. Files that bear small dissimilarities and are not identified as being "exact duplicates" of each other but are identical to a remarkable extent are known as near-duplicates. More precisely, finding Web pages that are almost, but not exactly the same for billions of documents is a very time-consuming task. In practice, the task can easily take days (if not weeks depending on the data set size), even with powerful distributed computing infrastructures and after trading accuracy for efficiency (e.g. by reducing document dimensionality).

3. Related Work

Recently, the detection of duplicate and near duplicate web documents has gained popularity in web mining research community. Very few research papers have suggested methodologies for duplicate near duplicate detection both in general documents and the web documents obtained by web crawling. Broder et al. [4] have suggested a technique for the estimation of the degree of similarity among pairs of documents is known as *shingling*. He has suggested a technique, in which all sequences of adjacent words are extracted. If two documents contain the same shingles set they are treated as equivalent and if the shingles set overlaps, they are considered as exact similar. Fetterly et al. [9] use five-gram as a shingle and sample 84 shingles for each document. Then the 84 shingles are built into six *super shingles*. The documents having two *super shingles* in common are considered as nearly duplicate documents. Yun Ling [25] developed a method to detect and delete near duplicated web pages; priority-based on text information. By this method, an algorithm to extract text information of web pages by DOM tree and a priority based algorithm for detecting near duplicated text information are implemented, so as to reduce the noise of web pages and hence to improve the efficiency of detecting near duplicated text information. Narayana et al. [19] presented an approach for the detection of near duplicate web pages in web crawling. Near duplicate web pages are detected followed by the storage of crawled web pages in to repositories. The keywords are extracted from crawled pages and on the basis of these keywords; the similarity score between two pages is calculated. The documents are considered as near duplicates if its similarity score satisfies a threshold value. Midhun Mathew et al. [16] offered a novel idea for the detection of near duplicate web pages. It uses a three stage algorithm in which the similarity verification is based on Singular Value Decomposition (SVD) [18] using angle threshold . But SVD requires more complicated Mathematical operations on TDW matrix along with the conversion of Jaccard threshold t into angle threshold which increases the algorithm complexity as well as the practical difficulties in measuring the angle. He introduce a new technique called MWO for similarity verification which directly works on Jaccard threshold and it reduces the complexity of the algorithm. Yerra and Yiu Kai Ng [23] presented new approach that performs copy detection on web documents .Their copy detection approach determines the similar web documents, similar sentences and graphically captures the similar sentences in any two web documents. Besides handling wide range of documents, their copy detection approach is applicable to web documents in different subject areas as it does not require static word lists. Jalbert and Weimer [15] proposed a system that automatically classified duplicate bug reports as they arrived to save developer time .Their system predicted duplicate status by utilizing surface features, textual semantics, and graph clustering. Gong et al. [12] proposed the SimFinder, an effective and efficient algorithm to identify all near duplicates in large-scale short text databases. The three techniques, namely, the ad hoc term weighting technique, the discriminative-term selection technique, the optimization technique are included in this SimFinder algorithm. It was illustrated that the SimFinder was an effective solution for short text duplicate detection with almost linear time and storage complexity by the experiments conducted. Hui Yang et al. [22] proposed an algorithm DURIAN which explored the use of simple text clustering and retrieval algorithms for identifying near-duplicate public comments. DURIAN identifies form letters and their edited copies in public comment collections by employing a traditional bag-of-words document representation, document attributes ("metadata"), and document content structure.

4. Duplicate Document Detection Algorithms

Andrei Z. Broder [2] proposed a method that can eliminate near-duplicate documents from a collection of hundreds of millions of documents by computing independently for each document a vector of features less than 50 bytes long and comparing only the vectors rather than entire documents. Provided that m is the size of the collection, the entire processing takes time O (m log m). The algorithm illustrated has been successfully implemented and is

employed in the context of the AltaVista search engine, currently.

Ahmad M. Hasnah [1] presented a novel data reduction algorithm employing the concept analysis which can be used as a filter in retrieval systems like search engines to eliminate redundant references to the similar documents. A study was performed on the application of the algorithm in automatic reasoning which effected in minimizing the number of stored facts without loosing of knowledge, by the authors. Their results illustrate that besides reducing the user time and increase his satisfaction; there was a good increase in precision of the retrieval system

Bar Yossef et al. [3] presented a novel algorithm, Dust Buster, for uncovering DUST (Different URLs with Similar Text) . They intended to discover rules that transform a given URL to others that are likely to have similar content. Dust Buster employs previous crawl logs or web server logs instead of probing the page contents to mine the dust efficiently. It is necessary to fetch few actual web pages to verify the rules via sampling. Search engines can increase the effectiveness of crawling, reduce indexing overhead, and improve the quality of popularity statistics such as Page Rank, which are the benefits provided by the information about the DUST.

Cho et al. [5] presented a new algorithm for efficiently identifying similar collections that form what they call a similar cluster. They made tradeoffs between the generality of the similar cluster concept and the cost of identifying collections that meet the criteria, during the development of their definitions and algorithm. The specific definition of what a human would consider a similar cluster cannot be captured by any definition of similarity as it is certain that more than one human would probably not agree any. However, their definition and cluster growing algorithm improved crawling and result displaying. They illustrated that in case of large web graphs: the work of a crawler can be reduced by 40%, and results can be much better organized when presented to a user, thus proving the high utility of their definition and algorithm.

Manku et al. [13] made two research contributions in developing a near-duplicate detection system intended for a multi-billion page repository. Initially, they demonstrated the appropriateness of Charikar's fingerprinting technique [5] for the objective. Subsequently, they presented an algorithmic technique Simhash to identify the existing f-bit fingerprints that varies from a given fingerprint in at most k bit positions, provided that value of k is small. They added the concept of feature weight to random projection .Features are computed using standard IR(Information Retrieval) techniques like tokenization , case folding, stop-word removal stemming and phrase detection. With simhash high dimensional vectors are transformed into f -bit fingerprint where f is small-sized fingerprints.

Y. Bernstein, J. Zobel [25] introduced a SPEX algorithm for efficiently identifying shared chunks in a collection. The fundamental observation behind the operation of SPEX is that if any sub chunk of a given chunk can be shown to be unique, then the chunk is its entirely must be unique. For example, if the chunk 'quick brown' occurs only once in the collection then there is no possibility that the chunk 'quick brown fox' is repeated. The algorithm can be extended to any desired chunk size l by iteration, at each phase incrementing the chunk size by one. It is able to provide an accurate representation of duplicate chunks of length u in a time proportional to (uv), where v is the length of the document collection

5. Conclusion

Web contains duplicate pages and mirrored web pages in abundance. The problem of finding relevant documents has become much more prominent due to the presence of duplicate and near duplicate data on the WWW. This redundancy in results increases the users' seek time to find the desired information within the search results, while in general most users just want to cull through tens of result pages to find new/different results. The efficient identification of duplicate and near duplicates is a vital issue that has arose from the escalating amount of data and the necessity to integrate data from diverse sources and needs to be addressed. In this paper, we have presented a comprehensive review of researches of Duplicate/Near duplicate document detection both in general and web crawling.

References:

[1] Ahmad M. Hasnah, "A New Filtering Algorithm for Duplicate Document Based on Concept Analysis", Journal of Computer Science, Vol. 2, No. 5, pp. 434-440, 2006

[2] Andrei Z. Broder ,"Identifying and Filtering Near-Duplicate Documents", Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. UK: Springer-Verlag, pp. 1-10,2000

[3] BarYossef, Z., Keidar, I., Schonfeld, U "Do Not Crawl in the DUST: Different URLs with Similar Text", 16th International world Wide Web conference, Alberta, Canada, Data Mining Track, 8-12 May,2007

[4] Broder, A., Glassman, S., Manasse, M., and Zweig G. "Syntactic Clustering of the Web, In 6th International World Wide Web Conference", pp: 393-404, 1997

[5] Charikar, M "Similarity estimation techniques from rounding algorithms", In Proc. 34th Annual Symposium on Theory of Computing pp. 380-388. 2002

[6] Cho, J., Shivakumar, N., Garcia-Molina, H.,. "Finding replicated web collections", ACM SIGMOD Record, June Vol. 29, No. 2, pp. 355 – 366,2000

[7] Cole, J. I., Suman, M., Schramm, P., Lunn, R., & Aquino, J. S. "The ucla internet report surveying the digital future year three UCLA Center for Communication Policy", 2003

[8] Dean, J., Henzinger, M. R. "Finding related pages in the World Wide Web", In: Proceeding of the 8th International World Wide Web Conference (WWW), pp. 1467-1479, 1999

[9] Fetterly D, Manasse M, Najork M, "On the evolution of clusters of near-duplicate Web pages", In Proceedings of the First Latin American Web Congress, pp.37- 45 ,Nov 2003

[10] Fetterly, D., Manasse, M., Najork, M., Wiener, J " A large-scale study of the evolution of web pages." In: Proceedings of the 12th International World Wide Web Conference (WWW),pp.669-678,2003

[11] Fetterly, D., Manasse, M., Najork, M.,. "Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages", in: Proceedings of the 7th International Workshop on the Web and Databases (WebDB), pp. 1-6,2004

[12] Gong, C., Huang, Y., Cheng, X., Bai, S."Detecting Near-Duplicates in Large-Scale Short Text Databases", Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol.5012, pp. 877-883,2008

[13] Gurmeet Singh Manku, Arvind Jain and Anish Das Sarma, "Detecting near-duplicates for web crawling", In Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, pp 141-150, 2007

[14] Henzinger, M.,. "Finding near-duplicate web pages: a large-scale evaluation of algorithms," in SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press , pp. 284-291,2006

[15] Jalbert, N., Weimer, W. "Automated Duplicate Detection for Bug Tracking Systems",IEEE International Conference on Dependable Systems and Networks With FTCS and DCC, DSN , pp. 52-61,24-27 June 2008

[16] 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, pp :16-21,April 2011

[17] Nielsen Media "Search engines most popular method of surfing the web" http://www.commerce.net/news/press/0416.html

[18] Shine N Das, K. V. Pramod, "Relevancy based Re-ranking of Search Engine Result", Proceedings of International Conference on Mathematical Computing and Management, Kerala, India, June 2010

[19] V.A. Narayana, P. Premchand and A. Govardhan, "Effective Detection of Near Duplicate Web Documents in Web Crawling", International Journal of Computational Intelligence Research, Volume 5, Number 1, pp. 83–96, 2009

[20] Xiao, C., Wang, W., Lin, X., Xu Yu, J., "Efficient Similarity Joins for Near Duplicate Detection", Proceeding of the 17th international conference on World Wide Web, pp: 131-140,2008

[21] Wang, Y., Kitsuregawa, M., "Evaluating contents-link coupled web page clustering for web search results", in: Proceedings of the 11th ACM International Conference on Information and Knowledge Management (CIKM), pp. 499-506, 2002

[22] Yang, H., Callan, J., Shulman, S., "Next Steps in Near-Duplicate Detection for eRulemaking", Proceedings of the 2006 international conference on Digital government research, Vol. 151, pp: 239 – 248, 2006

[23] Yerra, R., and Yiu Kai, NG.,. "A sentence-Based Copy Detection Approach for Web Documents", Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 3613, pp. 557-570,2005

[24] 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), pp. 296 – 305, 2003

[25] Y. Bernstein, J. Zobel —Accurate discovery of co-derivative documents via duplicate text detection‖ Information Systems pp: 595–609,2006

[26] Zamir, O., Etzioni, O., "Web document clustering: A feasibility demonstration", In: Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR),pp.46-54,1999