IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Issue-03 E-ISSN: 2347-2693

A Review on Duplicate and Near Duplicate Documents Detection Technique

Patil Deepali E.[1*], Ghatage Trupti B.[2], Takmare Sachin B.[3], Patil Sushama A.[4]

[1*,2,3]*Department of Computer Science and Engineering,*
Bharati Vidyapeeth's College of Engineering, Kolhapur, Maharashtra, India
[4]*DC Branch, Dept of Digital Communication, SSSIST Sehore.*

www.ijcseonline.org

Received: Feb/25/2016 Revised: Mar/10/2016 Accepted: Mar/23/2016 Published: Mar/31/ 2016

***Abstract*—** Duplicated web pages in consist of identical structure but regarded as clones regarded as clones different data. The identification of similar and near-duplicate pairs in a large collection is a significant the problem with the twide-spread application. The problem deliberated for diverse data types in diverse settings. The contemporary materialization is efficient of the problem identification of the near duplicate Web pages. This is challenging in the web scale to the voluminous data and the high dimensionalities of documents. This review has a fundamental intention to present an up-to-date review of the existing of literature in duplicate and near duplicate detection of general documents and web documents in web crawling. The classification of the existing literature in duplicate and the near duplicate detection techniques and a detailed description of same are the presented so as to make the review more comprehensible.

***Keywords*—** Web crawling, web pages, web mining, web content mining, and duplicate document, near duplicate detection.

I. INTRODUCTION

The portion of data mining that revolves around the assessment of World Wide Web is known as Web mining. Internet technology, Data Mining, World Wide Web as well as the Semantic Web, are incorporated in Web mining. The web usage mining, web content mining and web structure mining, are the three areas into which web mining has been into classified. Process of the information detection from millions of sources across the World Wide Web is known as the Web content mining. Web crawling is employed by the search engines to populate a local indexed repository of the web pages which is in turn utilized to answer the user search queries. Business has become more proficient and fruitful owing the ability to access the contents of the interest amidst huge heaps of data. The web crawling forms an integral component for search engines. A program or the automated script that traverses the World Wide Web in the systematic and automated manner is known as web crawler or web spider or web robot.

The web crawlers are also known by other names like ants, bots, automatic indexers, and worms. Web crawlers aid in creation of the web pages that proffer input for the systems that index, mine or else analyze pages. Documents and links related to the diverse topics are crawled by the Generic crawlers while precise knowledge is use to restrict the focused crawlers to crawl only specific topics. On the other hand the issue of near- duplicate web document elimination in generic crawl still remains unaddressed [9].Near duplicates posses minute difference and so are not regarded are exact duplicates. The typographical errors, versioned or the plagiarized documents, multiple representations of the same physical object, spam emails the generated from same template and the like are the some of chief causes for the prevalence of near duplicate pages [9]. Such similar content near duplicates contain and vary only in minimal areas of the document like the advertisements, counters and timestamps. The web searches consider these differences as inappropriate. Various studies have identified a substantial portion of the web pages as near duplicates [2].The duplicate detection has been recently studied in order to apply same in the web search tasks like the provision of efficient and effective web crawling, document ranking and document archiving. The duplicate detection methodologies proposed vary from manually coded rules to applications of the cutting edge techniques to the machine learning. A small number of authors have projected the methodologies for near duplicate detection recently. The computational minimization of storage resources was the intention of those systems. This present an extensive review of modern researches the associated with the problems that prevail in Detection of duplicates and near duplicates both in the web documents and general documents obtained by web crawling.

A. Duplicate and Near Duplicate Documents Detection

The duplicate documents are often found in large databases of digital documents like those found in the digital libraries or in the government declassification effort. The duplicate document detection is a scheme employed to avert the search by results from the constituting number of documents with the same or the same content nearly. The possibility for the search quality being degraded as a result of various copies of

the same documents being listed in the results search. Duplicate document analysis is carried out by only when the both of following conditions true:

• Collection of employs the link-based ranking model. This applicable for the crawlers that Crawl Web sites like that Web Sphere Portal crawler or Web crawler.
• Collection-security is disabled.
The duration of the global analysis, the indexing processes identify duplicates by scanning the document content for every document. When the two documents comprise identical the document content, they are regarded as duplicates. Files that bear small dissimilarities and not identified as being "exact duplicates" of each other but are identical to a remarkable extent are known as the near-duplicates. The examples of near duplicate documents:
• Files with a few different words - widespread form of the near-duplicates
• Files with the same content but different formatting – for instance, documents might contain the same text, but dissimilar fonts, bold type or italics
• Files with the same content but different file type – for Microsoft Word and PDF versions of the same file.

Administration of the large, unstructured document the repositories are carried out with the aid of the Near Duplicate Detection Technology (NDD). The NDD reduces conserves time, costs and diminishes the risk of errors, building a compelling ROI in all the circumstances where it is necessary for the people to make sense of large sets of documents.

II. LITERATURE REVIEW

Recently, the detection of near duplicate and duplicate web documents has gained popularity in web mining research community. This review merges and extents a wide range of works related to detection of duplicate and web documents and near duplicate documents. Detection techniques for identification of the duplicate and detection algorithms, near duplicate documents, Web based tools and other researchers of the duplicate and near duplicate documents are reviewed in corresponding the sub- sections.

A. Detection Techniques for Identification of Duplicate and Near Duplicate Documents

A technique for the estimation of degree of the similarity among pairs of documents was presented in 1997 by Broder et al. [2], which was known as a shingling, does not rely on any linguistic knowledge other than the ability to tokenize documents into a list of words. It is merely syntactic. Shingling, all word sequences of adjacent words are extracted. If two documents contain same set of shingles they are the considered equivalent and if their sets of the shingles appreciably by overlap, they are exceedingly the similar. The reduce in set to order off the shingles to a small, however the representative, subset they authors employed an unbiased the deterministic sampling technique that reduces by the storage requirements for the retaining information about the each document, and also computational effort of the comparing documents.

Grainy hash vector (GHV) representation by deployed in cooperative DIR systems for efficient and accurate merge-time the duplicate detection was introduced by the Bernstein et al. [3]. GHVs have the ability to detect exact duplicates and near-duplicates. They have mathematical properties that are well-defined. They conducted experiments on TREC AP collection and the demonstrated that GHVs identify the duplicate and near-duplicate document pairs at merge time effectively and efficiently. The duplication of management in cooperative DIR can be excellently performed by GHVs. Internet Search Engines is posed with challenges owing to by the growth of Internet that flood more copies of Web documents over search the results making them less relevant to users. The issues of computational efficiency and duplicate document detection effectiveness while relying on the "collection statistics" to consistently recognize document replicas in full-text the collections was their focus in the recent works [5]. Instead of using just one I- Match signature, they employed numerous I-Match the signatures all of which were derived from the randomized versions of original lexicon, in their proposed the solution.

The scheme proposed does not involve the direct computation of signature overlap regardless of employing the multiple fingerprints. Hence, in comparison with the case of fingerprints single valued, the signature comparison is just slightly slower. It can be the observed that addition of one extra signature component can improve by signature stability, i.e. further signature components of addition provide can more gains. The successful of derivation of lexicons for I-Match from a collection the different from target one, which is most preferred when the target collection is noisy, was demonstrated. The near-duplicate detection for personal contents like emails, web pages visited and documents was efficient. The solution provided for the duplicate detection by hybrid method was effective and in addition scalable. Precisely, the initially method conducts offline processing for the popular queries.

Additionally unpopular queries for improve the performance; it does additional work at run time. The scalability problem of traditional offline the methods could be effectively dealt by such a strategy, provided that performance problem of the traditional online methods is avoided. Their copy detection approach determines the similar web documents, similar graphically captures and sentences the similar sentences in any two web documents. Besides handling wide range of the documents, their copy detection approach is applicable to the web documents in different subject areas as it does not require static word lists. The problem of near-duplicate and

duplicate text has become increasingly the important owing to the growth of the text collection in size and multiple sources from which it is gathered. In the duplicate document detection, various works have been by performed. Their techniques utilized by the many applications.

The investigation on the scalability and performance of the duplicate document detection (DDD) is modest. Ye et al. [10] performed a systematic study on the parameter correlations in evaluated numerous and DDD most important parameters of DDD. The results illustrate that particularly for small the documents consisting of a major fraction of the whole Web; the precision of DDD is badly affected by small sampling the ratio. In order to make DDD feasible to the deal with large scale documents of collections, they proposed an adaptive the sampling strategy on the basis of their observation, which minimizes the sampling ratio of the documents with constraint of given the precision thresholds. The observations in their work were intended to aid in guiding the future DDD work.

B. Detection Algorithms

A method that can eliminate the near-duplicate documents from a collection of hundreds of millions documents by the computing independently for the each document by vector of features less than 50 bytes long and comparing only vectors rather than entire the documents, has been presented by Andrei Z. Broder [1]. Provided that m is the size of collection, entire the processing takes time O (mlogm). The algorithm illustrated has been the successfully implemented and is employed in context of the AltaVista search engine, currently.

Their results illustrate that besides increase his satisfaction and reducing the user time; there was a good increase in precision of retrieval the system. Intended for the identification of near-duplicate web pages, Broder et al.'s [2] shingling algorithm and Charikar's [4] the random projection based approach were considered state-of-the-art" algorithms. Addition, a new algorithm for efficiently identifying similar the collections that form what they call a similar the cluster was the proposed by authors. They tradeoffs made between generality of the similar cluster concept and the cost of identifying the collections that meet by the criteria, during the development of their definitions and algorithm.

The specific definition of what a human would a similar consider the cluster cannot be captured by the definition of similarity as it is certain that more than one human would the probably not agree any. However, their definition and cluster growing algorithm improve crawling and result displaying. Chowdhury et al. [5] proposed a novel similar document detection algorithm called I-Match. They utilized by the multiple data collections to evaluate their performance. The employed document collections were different in terms of degree of expected document duplication, size, and document lengths. It was illustrated that the I- Match, operates on the number of basis of the documents and it deals with the documents of all sizes efficiently.

The comparison with the state of the art, their method proved to have improved accuracy of detection. Deng et al. [6] established by a simple algorithm known as Stable Bloom Filter (SBF), which is based on the following idea: Given that there was no way to the store whole history of the stream, the stale information is removed by SBF in order to provide the space for those more recent elements. They systematically identified the some properties of SBF and the consequently illustrated a guaranteed tight upper bound of false positive rates. The authors conducted by experiments to compare with the SBF alternative methods. Provided that an acceptable false positive rate and fixed small space and were given, the outcome illustrated that their method was superior in terms of both time efficiency and accuracy. An elegant and efficient the probabilistic algorithm to approximate the number of the near-duplicate pairs was proposed by Deng et al. [7]. The algorithm scans the input data set once and uses only for the small constant space, independent of the number of objects in the data set, to provide a provably accurate estimate with the high probability.

They performed by the theoretical analysis and also the experimental evaluation on real and synthetic data. They illustrated that in reasonably small the dimensionality, the algorithm significantly outperforms the alternative the random sampling method. Their results illustrate that the combination of the multiple text-based signals and its computation over both rendered bodies and fetched significantly improve by the accuracy of duplicate detection algorithms. The multiple applications such as duplicate Web page detection on the Web provided with the efficient solutions.

C. Web Based Tools

A system for rapidly the determining document similarity among a set of documents obtained from an information retrieval (IR) system. The two documents are determined by the similar if the number of terms found to not be contained in the both documents is less than the some predetermined threshold compared to the total number of terms in the document. A tool, known as SIF that intends to identify all the similar files in a large file system has been presented by Udi Manber [10]. Files having significant number of the common pieces, though they are very different otherwise, are considered to be the similar. Additionally, SIF utilizes a preprocessed index to swiftly identify all the similar files to a query file. The applications such as the file management, information collecting, data compression, file synchronization, program reuse, and maybe even the plagiarism detection employ SIF.

III. CONCLUSIONS

The explosive growth of information the source available on the World Wide Web has the necessitated users to make use of automated tools to locate the desired information resources and to asses and follow the usage patterns. Web contains mirrored web pages and duplicate pages in abundance. The identification efficient of duplicate and near duplicates is a vital issue that has arose from escalating the amount of data and necessity to integrate the data from diverse sources and needs to be addressed. This presented a comprehensive review of up-to-date researches of Duplicate and near duplicate document detection both in general and web crawling.

In addition, a short introduction about web mining, web Crawling and duplicate document detection have also been presented. This felt necessary when the work on developing Duplicate/Near document duplicate detection is very hopeful, and is still in promising status. This review intends to aid upcoming researchers in the field of Duplicate/Near duplicate document detection in the web crawling to understand the available methods and help to perform their research in further direction. This presented a comprehensive review of up-to-date researches of Duplicate and near duplicate document detection both in general and web crawling. The result of classification the literature existing in duplicate document and the near duplicate detection techniques and a detailed description of same are the presented so as to make the review more comprehensible.

ACKNOWLEDGMENT

The authors are grateful to sincere thanks express and gratitude to Computer Department of Engineering, BVCOEK for encouragement and facilities that were offered to us for carrying out this project. The authors would like to thank Prof. Chougule A. B. (Head of computer science and Engineering (BVCOEK)) & Prof. S. B. Takmare.

REFERENCES

[1] 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.

[2] 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.

[3] Bernstein, Y., Shokouhi, M., and Zobel, J., "Compact Features for Detection of Near- Duplicates in Distributed Retrieval", in 'Proceedings of String Processing and Information Retrieval Symposium (to appear)', Glasgow, Schotland, 2006.

[4] Charikar, M.,“Similarity estimation techniques from rounding algorithms”, In Proc. 34th Annual Symposium on Theory of Computing (STOC 2002), pp. 380-388, 2002.

[5] Chowdhury, A., Frieder, O., Grossman, D., and Catherine Mccabe, M., “Collection Statistics for Fast Duplicate Document Detection", In. ACM Transactions on Information Systems (TOIS), Vol. 20, No. 2, 2002.

[6] Deng, F., Rafiei, D., "Approximately detecting duplicates for streaming data using stable bloom filters" ,Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 25-36, 2006.

[7] Deng, F., Rafiei, D., "Estimating the Number of Near Duplicate Document Pairs for Massive Data Sets using Small Space", University of Alberta, Canada, 2007.

[8] Manku, G. S., Jain, A., Sarma, A. D., "Detecting near-duplicates for web crawling", Proceedings of the 16th international conference on World Wide Web, pp: 141 – 150, 2007.

[9] Udi Manber., "Finding Similar Files In A Large File System", Proceedings of the USENIX Winter 1994 Technical Conference on USENIX Winter 1994 Technical Conference, San Francisco, California, pp. 2-2, 1994.

[10] Ye, S., Wen, J., R., and Ma, W.Y., "A systematic study of parameter correlations in large scale duplicate document detection", Text and Document Mining, 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, pp. 275-284, 2006.

AUTHORS PROFILE

Ms. Patil Deepali Eknath is a M.E. student in Bharati Vidyapeeth’s College of Engineering, Kolhapur. Maharashtra, India. Her research interest lies in Networking and Network security. She has published one international paper and presented two paper in National Level Conference.

Ms. Ghatage Trupti Babasaheb is a M.E. student in Bharati Vidyapeeth’s College of Engineering, Kolhapur, Maharashtra, India. She has worked as Lecturer in Dr. D.Y. Patil Polytechnic, Kasaba Bawada, Kolhapur, Maharashtra, India. Her research interest lies in Data Mining, Database.She has published one international paper and presented two paper in National Level Conference.

Mr. Takmare Sachin Balawant is working as assistant professor in Computer Science and Engineering department of Bharati Vidyapeeth’s College of Engineering, Kolhapur with Teaching experience of about 10 years. He has published about 3 International Papers and 5 National Papers.

Ms. Patil Sushama Arjun is a Student of M. Tech, DC Branch, Dept of Digital Communication, and SSSIST Sehore.