Near Duplicate Document Detection Survey

Bassma S. Alsulami, Maysoon F. Abulkhair, Fathy E. Eassa
Faculty of Computing and Information Technology
King AbdulAziz University
Jeddah, Saudi Arabia

Abstract—Search engines are the major breakthrough on the web for retrieving the information. But List of retrieved documents contains a high percentage of duplicated and near document result. So there is the need to improve the performance of search results. Some of current search engine use data filtering algorithm which can eliminate duplicate and near duplicate documents to save the users' time and effort. The identification of similar or near-duplicate pairs in a large collection is a significant problem with wide-spread applications. In this paper survey present an up-to-date review of the existing literature in duplicate and near duplicate detection in Web.

Keyword—Duplicate document, near duplicate pages, near duplicate detection, Detection approaches

1. INTRODUCTION

Information on the Web is very huge in size. There is a need to use this big volume of information efficiently for effectively satisfying the information need of the user on the Web. Search engines become the major breakthrough on the web for retrieving the information. Where, among users looking for information on the Web, 85% submit information requests to various Internet search engines. Search engines are critically important to help users find relevant information on the Web.

Search engines in response to a user's query typically produces the list of documents ranked according to closest to the user's request. These documents are presented to the user for examination and evaluation. Web users have to go through the long list and inspect the titles, and snippets sequentially to recognize the required results. Filtering the search engines' results consumes the users' effort and time especially when a lot of near duplicate.

The efficient identification of near duplicates is an important in a many applications especially at that has a large amount of data and the necessity to save data from diverse sources and needs to be addressed. Though near duplicate documents display striking similarities, they are not bit wise similar. Web search engines considerable problems due to duplicate and near duplicate web pages. These pages increase 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 [1].

The identification of similar or near-duplicate document in a large collection is a significant problem with wide-spread applications. The problem has been deliberated for different data types (e.g. textual documents, spatial points and relational records) in a variety of settings. Another contemporary materialization of the problem is the efficient identification of near-duplicate Web pages. This is certainly challenging in the web-scale due to the voluminous data and high dimensionalities of the documents [2]. 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 [3], ranking [4], clustering [5] and archiving caching [6].

The paper is organized as follows. Overview of Near Duplicate Document is introduced in Section 2. The goal of Near Duplicate Detection is defined in Section 3. Section 4, we describe main Near Duplicate approaches. In Application of Duplicate Document Detection are presented in Section 5. Finally we conclude the paper in Section 6.

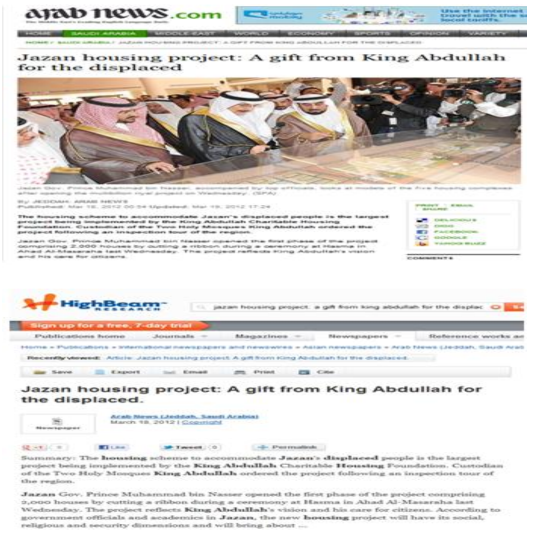
2. NEAR DUPLICATE DOCUMENT:

Two documents are regarded as duplicates if they comprise identical document content. Documents 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 [7]. Web contains duplicate pages and mirrored web pages in abundance. Standard check summing techniques can facilitate the easy recognition of documents that are duplicates of each other (as a result of mirroring and plagiarism). A more difficult problem is the identification of near-duplicate documents. Two such documents are identical in terms of content but differ in a small portion of the document such as advertisements, counters and timestamps. Following are some of the examples of near duplicate documents [1]:

- Files with a few different words - widespread form of near-duplicates
- Files with the same content but different formatting – for instance, the documents might contain the same text, but dissimilar fonts, bold type or italics
- Files with the same content but different file type – for instance, Microsoft Word and PDF versions of the same file.

The most challenging among all the above, from the technical perspective, is the first situation - small differences in content. The application of a near de-duplication technology can provide the capacity to recognize these files.

In the Web, there are two types of near duplicates [8]. Figure 1 shows a pair of same-core Web pages that only differs in the framing, advertisements, and navigational banners added each by the San Francisco Chronicle and New York Times. Both articles exhibit almost identical core contents, reporting on Jazan housing project, a gift from King Abdullah for the displaced.



Jazan housing project: A gift from King Abdullah for the displaced

Home » Publications » International newspapers and newswires » Asian newspapers » Arab News (Jeddah, Saudi Arab

Recently viewed: Article: Jazan housing project. A gift from King Abdullah for the displaced.

Jazan housing project: A gift from King Abdullah for the displaced.

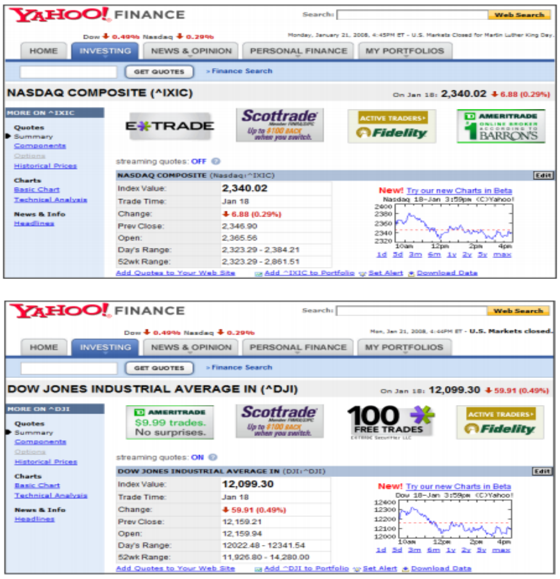
Arab News (Jeddah, Saudi Arabia)
March 18, 2012 | Copyright

Summary: The **housing** scheme to accommodate **Jazan**'s **displaced** people is the largest project being implemented by the **King Abdullah** Charitable **Housing** Foundation. Custodian of the Two Holy Mosques **King Abdullah** ordered the project following an inspection tour of the region.

Jazan Gov. Prince Muhammad bin Nasser opened the first phase of the project comprising 2,000 houses by cutting a ribbon during a ceremony at Hasma in Ahad Al-Masaraha last Wednesday. The project reflects **King Abdullah**'s vision and his care for citizens. According to government officials and academics in **Jazan**, the new **housing** project will have its social, religious and security dimensions and will bring about ...

Fig. 1 Example of same-core Web pages

Figure 2 is an example of the opposite case from Yahoo! Finance, showing two daily summaries of the NASDAQ and Dow Jones indexes. In particular for domains like stock markets, news sites often use very uniform layouts and the actual contents-of-interest only constitute a fraction of the page. Hence, though visually even more similar than the pair in Figure 1, the pair in Figure 2 should not be identified as near duplicates. Typically, our sociologists would only consider Figure 1's same-core pair to be near duplicates, since the core articles are their focus. Near duplicates would not be discarded but could be collected into a common set which is then tagged in batch.



NASDAQ COMPOSITE (^IXIC) On Jan 18: 2,340.02 6.88 (0.29%)

NASDAQ COMPOSITE (Nasdaq:^IXIC)

Index Value:	2,340.02
Trade Time:	Jan 18
Change:	6.88 (0.29%)
Prev Close:	2,346.90
Open:	2,365.56
Day's Range:	2,323.29 - 2,384.21
52wk Range:	2,323.29 - 2,861.51

DOW JONES INDUSTRIAL AVERAGE IN (^DJI) On Jan 18: 12,099.30 59.91 (0.49%)

DOW JONES INDUSTRIAL AVERAGE IN (DJI:^DJI)

Index Value:	12,099.30
Trade Time:	Jan 18
Change:	59.91 (0.49%)
Prev Close:	12,159.21
Open:	12,159.94
Day's Range:	12022.48 - 12341.54
52wk Range:	11,926.80 - 14,280.00

Fig. 2 Example of same-frame Web pages

3. NEAR DUPLICATE DOCUMENT DETECTION:

Detection of Near Duplicate Document (NDD) is the problem of finding all documents rapidly whose similarities are equal to or greater than a given threshold. Near Duplicate document detection became an interesting problem in late 1990s with the growth of Internet [9]. Most existing techniques for identifying near duplicates are divided into two categories:

- Near duplicate prevention.
- Near duplicate detection.

Near duplicate prevention techniques include physical isolation of the information and use of special hardware for authorization. Related work about copy prevention techniques will not be given because it is beyond of the scope of this paper. . Related work about near duplicate detectiontechniques will be given in next section.

4. Near Duplicate Document Detection Approaches:

The problem of near duplicate detection of documents in general, and Web pages in particular, has been well studied, and a variety of approaches have been proposed. Approaches on near-duplicates detection and elimination are many in the history. In general these approaches may be broadly classified as shown in Figure 3 into Syntactic, URL based and Semantic based approaches [7].

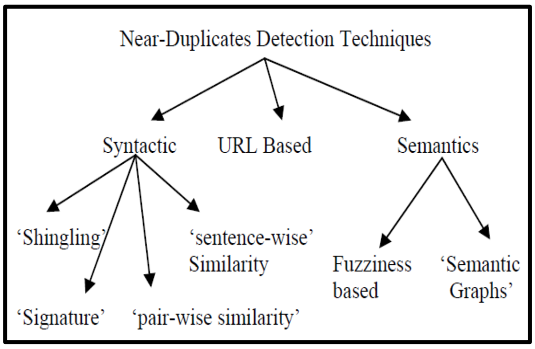


Fig. 3 Near-duplicates detection techniques

4.1 Syntactical Approaches:

One of the earliest was by Broder et al [11], proposed a technique to compute the resemblance of two documents, each is broken into overlapping fragments called shingles. Shingles does not rely on any linguistic knowledge other than the ability to tokenize documents into a list of words, i.e., it is merely syntactic. In this, all word sequences (shingles) of adjacent words are extracted. If two documents contain the same set of shingles they are considered equivalent and can be termed as near-duplicates. Broder et al. used an unbiased deterministic sampling technique to reduce the set of shingles to a small, yet representative, subset. This sampling reduces the storage requirements for retaining information about each document, and it reduces the computational effort of comparing documents. The problem of finding similarity of text documents was investigated and a new similarity measure was proposed to compute the pair-wise similarity of the documents using a given series of terms of the words in the documents.

Pair-wise similarity computation deals with finding pairs of objects in a large dataset that are similar according to some measure. This problem is frequently encountered in text processing applications, for example, clustering for unsupervised learning. In [12], the near duplicate document completes the pair-wise similarity comparisons in two steps: inverted index building and then similarity computations with it.

Sentences-wise similarity proposed in [13], similarity measure can be acquired by comparing the exterior tokens of inter-sentences, but relevance measure can be obtained only by comparing the interior meaning of the sentences. A method to explore the Quantified Conceptual Relations of word-pairs by using the definition of a lexical item was described, and a practical approach was proposed to measure the inter-sentence relevance.

In determining which k-grams in a document should be used for creating signatures, Theobald et al.'s SpotSigs method is perhaps the most creative and interesting one [8]. When developing near duplicate detection methods for clustering news articles shown on various Web sites, they observe that stop words seldom occur in the unimportant template blocks such as navigation sidebar or links shown at the bottom of the page. Based on this observation, they first scan the document to find stop words in it as anchors. K tokens right after an anchor excluding stop words are grouped as a special k-gram, or so called a "spot signature" in their terminology. The raw representation of each target document is therefore a set of spot signatures. To some extent, the construction of spot signatures can be viewed as a simple and efficient heuristic to filter terms in template blocks so that the k-grams are extracted from the main 420 content block only. Once the spot signatures have been extracted, the same techniques of using hash functions as seen in other NDD methods can be directly applied to reduce the length of the spot signature vectors.

The method based on shingles and the signature method when compared, the signature method in the presence of inverted index was more efficient. As a result, the above stated syntactic approaches carry out only a text based comparison. And these approaches did not involve the URLs or any link structure techniques in identification of near-duplicates. The following subsection discusses the impact of URL based approaches on near duplicates detection.

4.2 URL Based Approaches:

A novel algorithm, Dust Buster, for uncovering DUST (Different URLs with Similar Text) was 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. 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 [14].

Reference [15] shows another approach where detecting process was divided into three steps.

1. Removal according to URLs. First, remove pages with the same URL in the initial set of pages to avoid the same page been download repeated due to repeat links.
2. Remove miscellaneous information in the pages and extract the texts. Pretreatment the pages, remove the navigation information, advertising information, html tags, and other miscellaneous information on the pages, extract the text content and get a set of texts.
3. Detect with DDW algorithm. Use the DDW algorithm to detect similar pages.

The combination of such URL based approaches along with syntactic approaches is still not sufficient as they do not have semantic in identifying near-duplicates. The following subsection discusses briefly a few semantic based approaches.

4.3 Semantic Approaches:

A method on plagiarism detection using fuzzy semantic-based string similarity approach was proposed. The algorithm was developed through four main stages. First is pre-processing which includes tokenization, stemming and stop words removing. Second is retrieving a list of candidate documents for each suspicious document using shingling and Jaccard coefficient. Suspicious documents are then compared sentence-wise with the associated candidate documents. This stage entails the computation of fuzzy degree of similarity that ranges between two edges: 0 for completely different sentences and 1 for exactly identical sentences. Two sentences are marked as similar (i.e. plagiarized) if they gain a fuzzy similarity score above a certain threshold. The last step is post-processing hereby consecutive sentences are joined to form single paragraphs/sections [16]. Recognizing that two Semantic Web documents or graphs are similar, and characterizing their differences is useful in many tasks, including retrieval, updating, version control and knowledge base editing. A number of text based similarity metrics are discussed as in [17] that characterize the relation between Semantic Web graphs and evaluate metrics for three specific cases of similarity that have been identified: similarity in classes and properties used while differing only in literal content, difference only in base-URI, and versioning relationship.

5. Application of Duplicate Document Detection:

Identifying NDDs has a much wider range of applications. Some of these applications are as following [18]:

- Technical support document management: Many companies have millions of technical support documents which are frequently merged and groomed. In this process it is very important to identify NDDs.
- Plagiarism detection: Modern electronic technologies have made it extremely easy to plagiarize. In order to tackle this problem NDD detection mechanisms can be used.
- Web crawling: The drastic growth of the World Wide Web requires modern web crawlers to be more efficient. NDD detection algorithms are one of the means that can be used in this regard.
- Digital libraries and electronic publishing: Effectively organizing large digital libraries, which include several large electronically published collections and news archives with some overlap, requires NDD detection algorithms.
- Database cleaning: In database systems an essential step for data cleaning and data integration is the identification of NDDs.
- Files in a file system: near-duplicate detection to reduce storage for files.
- E-mails: identify near-duplicates for spam detection.

6. CONCLUSION

In this paper, we investigated the problem of how to eliminate near duplicate document. 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 survey of up-to-date researches of Duplicate/Near duplicate document detection. We review the main near duplicates document approaches.

REFERENCES

[1] J. P. Kumar and P. Govindarajulu. Duplicate and near duplicate documents detection: A review. European Journal of Scientic Research, 32(4):514{527, 2009.

[2] Ranjna Gupta et. al. Query Based Duplicate Data Detection on WWW (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 1395-1400

[3] G. S. Manku, A. Jain, and A. D. Sarma. Detecting near-duplicates for web crawling. In In ACM WWW'07, pages 141–150, NY, USA, 2007. ACM

[4] Yi, L., Liu, B., Li, X., 2003. "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

[5] Fetterly, D., Manasse, M., Najork, M., 2004. "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

[6] Hung-Chi Chang and Jenq-Haur Wang. Organizing news archives by near-duplicate copy detection in digital libraries. In Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, volume 4822 of Lecture Notes in Computer Science, pages 410{419. Springer Berlin / Heidelberg, 2007.

[7] 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

[8] Theobald, M., Siddharth, J., and Paepcke, A. 2008. Spotsigs: robust and efficient near duplicate detection in large web collections. In SIGIR. 563–570

[9] Udi Manber. Finding similar files in a large file system. In WTEC'94: Proceedings of the USENIX Winter 1994 Technical Conference on USENIX Winter 1994 Technical Conference, pages 2{2, Berkeley, CA, USA, 1994. USENIX Association

[10] E. Uyar. Near-duplicate News Detection Using Named Entities. M.S. Thesis, Department of Computer Engineering, Bilkent University, Ankara, Turkey, 2009.

[11] A. Broder, S. Glassman, M. Manasse and G. Zweig. Syntactic clustering of the web. In Proc. of the 6th International World Wide Web Conference, Apr. 1997.

[12] Wu, Y. et al (26, 3 2012). Efficient near-duplicate detection for q&a forum . Retrieved from http://aclweb.org/anthology-new/I/I11/I11-1112.pdf

[13] Maosheng Zhong, Yi Hu, Lei Liu and Ruzhan Lu, A Practical Approach for Relevance Measure of Inter-Sentence, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, pp: 140-144, 2008

[14] 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, pp: 111 – 120, 2007.

[15] Junping Qiu and Qian Zeng, Detection and Optimized Disposal of NearDuplicate Pages, 2nd International Conference on Future Computer and Communication, Vol.2, pp: 604-607, 2010.

[16] Salha Alzahrani and Naomie Salim, Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection, 2010.

[17] Krishnamurthy Koduvayur Viswanathan and Tim Finin, Text Based Similarity Metrics and Delta for Semantic Web Graphs, pp: 17-20, 2010

[18] Nikkhoo , H. K. (2010). The impact of near-duplicate documents on information retrieval evaluation. (Master's thesis, University of Waterloo, Ontario, Canada)Retrieved from http://uwspace.uwaterloo.ca/bitstream/10012/5750/1/Khoshdel%20Nikkhoo_Hani.pdf