A NEAR-DUPLICATE DETECTION ALGORITHM TO FACILITATE DOCUMENT CLUSTERING

Lavanya Pamulaparty[1], Dr. C.V Guru Rao[2] and Dr. M. Sreenivasa Rao[3]

[1]Department of CSE, Methodist college of Engg. & Tech., OU, Hyderabad
[2]Department of CSE, S R Engineering College, JNT University, Warangal
[3]Department of CSE, School of IT, JNT University, Hyderabad

ABSTRACT

Web Ming faces huge problems due to Duplicate and Near Duplicate Web pages. Detecting Near Duplicates is very difficult in large collection of data like "internet". The presence of these web pages plays an important role in the performance degradation while integrating data from heterogeneous sources. These pages either increase the index storage space or increase the serving costs. Detecting these pages has many potential applications for example may indicate plagiarism or copyright infringement. This paper concerns detecting, and optionally removing duplicate and near duplicate documents which are used to perform clustering of documents .We demonstrated our approach in web news articles domain. The experimental results show that our algorithm outperforms in terms of similarity measures. The near duplicate and duplicate document identification has resulted reduced memory in repositories.

KEYWORDS

Web Content Mining, Information Retrieval, document clustering, duplicate, near-duplicate detection, similarity, web documents

1. INTRODUCTION

The WWW is a popular and interactive medium to disseminate information today. Holocene epoch has detected the enormous emergence of Internet document in the World Wide Web. The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. Due to these situations, we are currently drowning in information and facing information overload [1]. 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 [2]. The demand for integrating data from heterogeneous sources leads to the problem of near duplicate web pages. Near duplicate data bear high similarity to each other, yet they are not bitwise identical [3] [4] but strikingly similar. They are pages with minute differences and are not regarded as exactly similar pages. Two documents that are identical in content but differ in small portion of the document such as advertisement, counters and timestamps. These differences are irrelevant for web search. So if a newly-crawled page Pduplicate is deemed a near-duplicate of an already-crawled page P, the crawl engine should ignore Pduplicate and its entire out-going links (intuition suggests that these are probably near-duplicates of pages reachable from P) [4, 35]. Near Duplicate web pages from different mirrored sites may only differ in the header or footnote zones that denote the site URL and update time [5].

DOI : 10.5121/ijdkp.2014.4604

Duplicates and Near Duplicate Web pages are creating large problems for web search engines like they increase the space needed to store the index, either slow down or increase the COST of saving results and annoy the users [34]. Elimination of near-duplicates saves network bandwidth, reduces storage costs and improves the quality of search indexes. It also reduces the load on the remote host that is serving such web pages [10].

The determination of the near duplicate web pages [28-29] [21] aids the focused crawling, enhanced quality and diversity of the query results and identification on spam. The near duplicate and duplicate web page identification helps in Web mining applications for instance, community mining in a social network site [20], plagiarism detection [24], document clustering [29], collaborative filtering [30], detection of replicated web collections [31] and discovering large dense graphs [34]. NDD removal is required in Data Cleaning, Data integration, Digital libraries and electronic published collections of news archives. So in this paper, we propose a novel idea for finding near duplicate web pages from a huge repository.

2. RELATED WORK

The proposed research has been motivated by numerous existing works on and near duplicate documents detection. Duplicate and near-duplicate web pages are creating large problems for web search engines: they increase the space needed to store the index, either slow down or increase the cost of serving results, and annoy the users. This requires the creation of efficient algorithms for computing clusters of duplicates [4, 6, 7, 16, 21, 22, 29, 30]. The first algorithms for detecting near-duplicate documents with a reduced number of comparisons were proposed by Manber [25] and Heintze [17]. Both algorithms work on sequences of adjacent characters. Brin [40] started to use word sequences to detect copyright violations. Shiva Kumar and Garcia-Molina [31] continued this research and focused on scaling it up to multi-gigabyte databases. Broder et al [5] defined two concepts resemblance and containment to measure the similarity of degree of two documents. He used word sequences to efficiently find near-duplicate web pages.

The dimensionality reduction technique proposed by Charikar's Simhash [34] is to identify near duplicate documents which maps high dimensional vectors to small-sized fingerprints. They developed an approach based on random projections of the words in a document. Henzinger [9] compared Broder et al.'s [7] shingling algorithm and Charikar's [34] random projection based approach on a very large scale, specifically on a set of 1.6B distinct web pages.

In the syntactical approach we define binary attributes that correspond to each fixed length substring of words (or characters). These substrings are a framework for near-duplicate detection called shingles. We can say that a shingle is a sequence of words. A shingle has two parameters: the length and the offset. The length of the shingle is the number of the words in a shingle and the offset is the distance between the beginnings of the shingles. We assign a hash code to each shingle, so equal shingles have the same hash code and it is improbable that different shingles would have the same hash codes (this depends on the hashing algorithm we use). After this we randomly choose a subset of shingles for a concise image of the document [6, 8, and 9]. M. Henzinger [32] uses like this approach AltaVista search engine .There are several methods for selecting the shingles for the image: a fixed number of shingles, a logarithmic number of shingles, a linear number of shingle (every nth shingle), etc. In lexical methods, representative words are chosen according to their significance. Usually these values are based on frequencies. Those words whose frequencies are in an interval (except for stop- words from a special list of about 30 stop-words with articles, prepositions and pronouns) are taken. The words with high frequency can be non informative and words with low frequencies can be misprints or occasional words.

In lexical methods, like I-Match [11], a large text corpus is used for generating the lexicon. The words that appear in the lexicon represent the document. When the lexicon is generated the words with the lowest and highest frequencies are deleted. I-Match generates a signature and a hash code of the document. If two documents get the same hash code it is likely that the similarity measures of these documents are equal as well. I-Match is sometimes instable to changes in texts [22]. Jun Fan et al. [16] introduced the idea of fusing algorithms (shingling, I-Match, simhash) and presented the experiments. The random lexicons based multi fingerprints generations are imported into shingling based simhash algorithm and named it "shingling based multi fingerprints simhash algorithm". The combination performance was much better than original Simhash.

3. ARCHITECTURE OF PROPOSED WORK

The paper proposed the novel task for detecting and eliminating near duplicate and duplicate web pages to increase the efficiency of web crawling. So, the technique proposed aims at helping document classification in web content mining by eliminating the near-duplicate documents and in document clustering. For this, a novel Algorithm has been proposed to evaluate the similarity content of two documents.

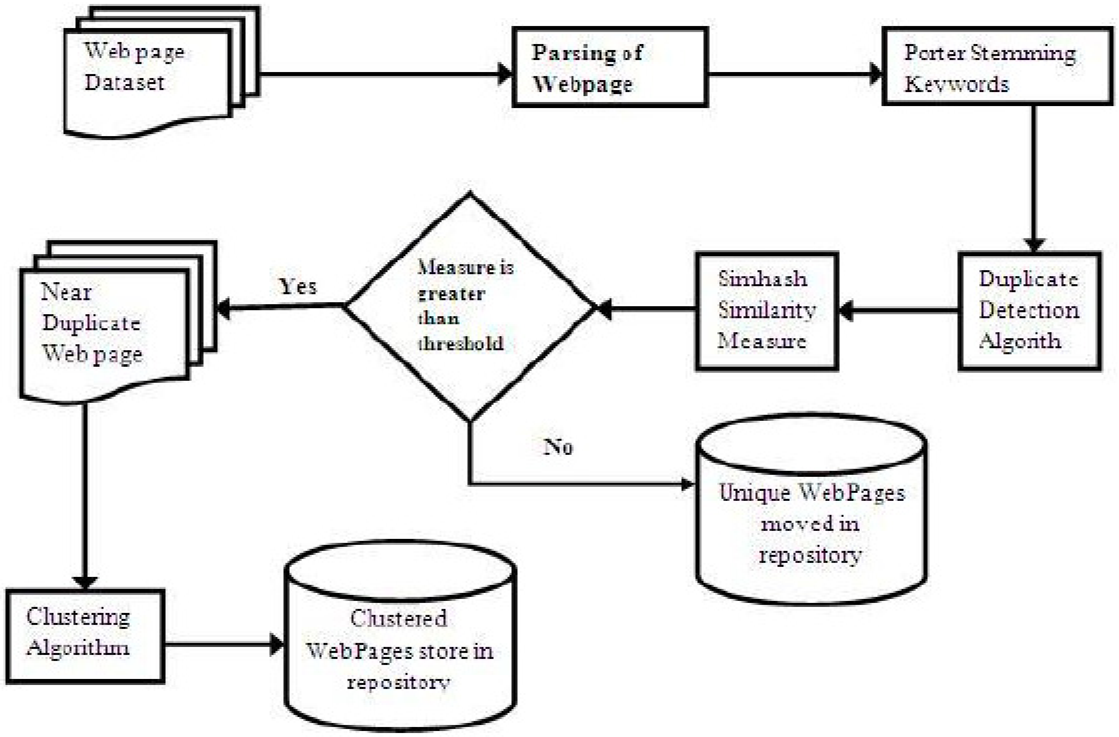

Web page Dataset
Parsing of Webpage
Porter Stemming Keywords
Duplicate Detection Algorith
Simhash Similarity Measure
Measure is greater than threshold
Yes
No
Near Duplicate Web page
Unique WebPages moved in repository
Clustering Algorithm
Clustered WebPages store in repository

Figure 1. Proposed Architecture

3.1. Architectural Steps

Architectural steps which includes:(i)Web page Dataset Collection (ii) Pre-processing (iii)Store in database (iv) Rendering DD Algorithm (v)Verifying the similarity content (vi) Filtering Near Duplicates (vii) Refined Results

3.1.1. Dataset Collection

We have collected data using Wget, a software program that retrieves content from webservers and have gathered our data sets. Here, we have collected by giving some URLs of the news websites which usually have replicas, so we can get more number of web documents with relevant information. Wget can optionally work like a web crawler by extracting resources linked from HTML pages and downloading them in sequence, repeating the process recursively until all the pages have been downloaded or a maximum recursion depth specified by the user has been reached. The downloaded pages are saved in a directory structure resembling that on the remote server.

3.1.2. Parsing of Web Pages

Once a page has been crawled, we need to parse its content to extract information that will feed and possibly guide the future path of the crawler. Parsing might also involve steps to convert the extracted URL to a canonical form, remove stop words from the page's content and stem the remaining words [33]. HTML Parsers are freely available for many different languages. They provide the functionality to easily identify HTML tags and associated attribute-value pairs in a given HTML document.

3.1.3. Stop-listing

When parsing a Web page to extract content information or in order to score new URLs suggested by the page, it is often helpful to remove commonly used words or stop words such as “it" and “can". This process of removing stop-words from text is called stop-listing [26].

3.1.4. Stemming Algorithm

Stemming algorithms, or stemmers, are used to group words based on semantic similarity. Stemming algorithms are used in many types of language processing and text analysis systems, and are also widely used in information retrieval and database search systems [25]. A stemming algorithm is an algorithm that converts a word to a related form. One of the simplest such transformations is conversion of plurals to singulars, another would be the derivation of a verb from the gerund form (the "-ing" word). A number of stemming or conflation algorithms have been developed for IR (Information Retrieval) in order to reduce morphological variants to their root form. A stemming algorithm would normally be used for document matching and classification by using it to convert all likely forms of a word in the input document to the form in a reference document [22]. Stemming is usually done by removing any attached suffixes, and prefixes from index terms before the assignment of the term. Since the stem of a term represents a broader concept than the original term, the stemming process eventually increases the number of retrieved documents [23].

3.1.5. Duplicate Detection (DD) Algorithm

Step 1: Consider the Stemmed keywords of the web page.

Step 2: Based on the starting character i.e. A-Z we here by assumed the hash values should start with1-26.

Step 3: Scan every word from the sample and compare with DB (data base) (initially DB Contains NO key values. Once the New keyword is found then generate respective hash value. Store that key value in temporary DB.

Step 4: Repeat the step 3 until all the keywords get completes.

Step 5: Store all Hash values for a given sample in local DB (i.e. here we used array list)

Step 6: Repeat step 1 to step 6 for N no. of samples.

Step 7: Once the selected samples were over then calculate similarity measure on the samples hash values which we stored in local DB with respective to webpages in repository.

Step 8: From similarity measure, we can generate a report on the samples in the score of % forms. Pages that are 80% similar are considered to be near duplicates.

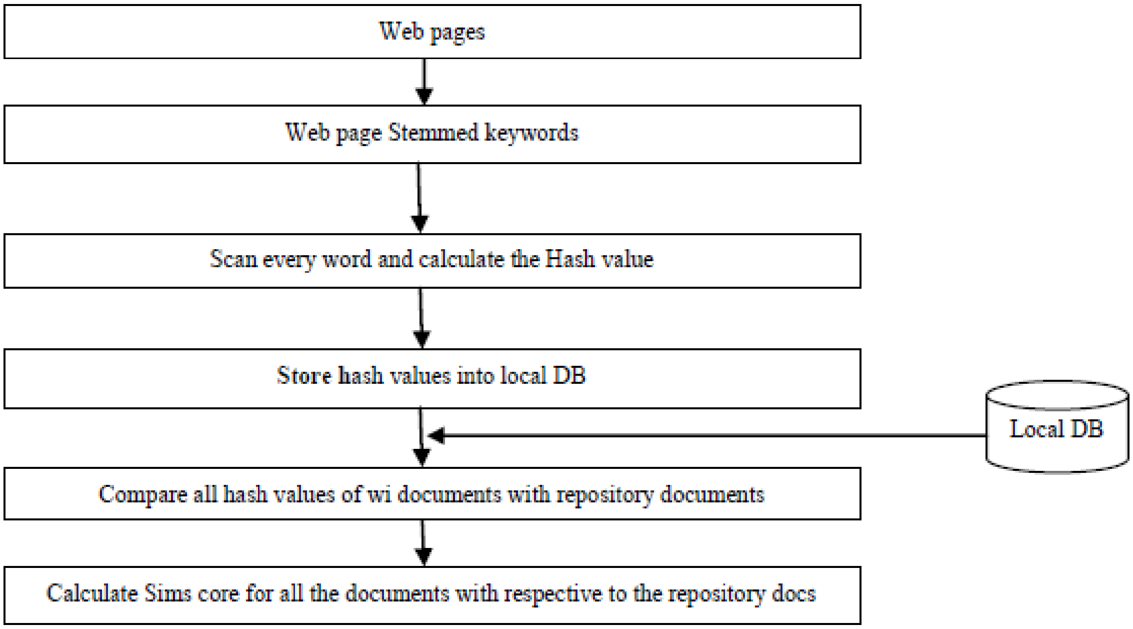

Web pages
Web page Stemmed keywords
Scan every word and calculate the Hash value
Store hash values into local DB
Local DB
Compare all hash values of wi documents with repository documents
Calculate Sims core for all the documents with respective to the repository docs

Figure 2. DD algorithm Work flow

This algorithm consists of three major steps. The first step is to calculate the hash vales of the each keyword in a web page w_1. Then in the second step they are stored in a temporary data base. This procedure is repeated for all the samples collected. The third step is to calculate the similarity of the pages on the hash values generated which are stored in a local DB (data base). If the threshold is above 80% then it is considered as a near duplicate web page. Whenever a new page w_n is coming and needs to be stored in the repository, its similarity with the web pages already existing in the repository needs to be compared. Similarity is calculated using the following formula.

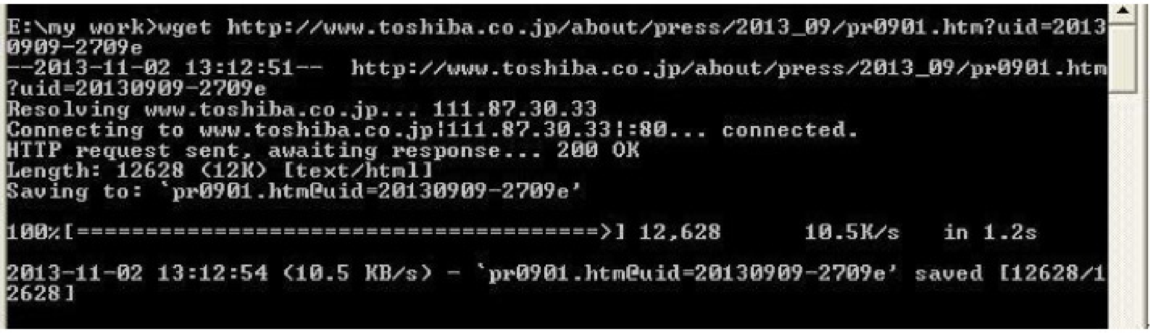
$$\text{SimScore} = \frac{\text{Number of hash values in } w_1, w_{2\ldots\ldots}\ w_n}{\text{Total } w_1, w_{2\ldots\ldots}\ w_n \text{ under consideration in Repository}} \text{ X } 100$$

4. Experimental Environment And Setup

The proposed near duplicate document detection system is programmed using Java (jdk 1.6) and the backend used is MS Access. The experimentation has been carried out on a 2.9 GHz, i5 PC machine with 4 GB main memory running a 32-bit version of Windows XP. The Web pages are collected using Wget computer program.

Table 1. Web pages dataset details

S. No	Documents data set	Number of Documents
1	Current Affairs	14
2	Technical articles	10
3	Shopping sites articles	15
4	Raw files	20
	Total	59

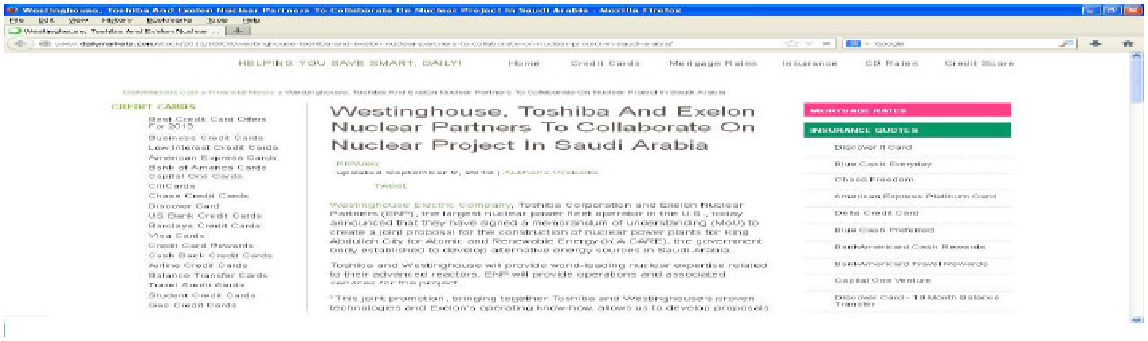


```
E:\my work>wget http://www.toshiba.co.jp/about/press/2013_09/pr0901.htm?uid=2013
0909-2709e
--2013-11-02 13:12:51--  http://www.toshiba.co.jp/about/press/2013_09/pr0901.htm
?uid=20130909-2709e
Resolving www.toshiba.co.jp... 111.87.30.33
Connecting to www.toshiba.co.jp|111.87.30.33|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 12628 (12K) [text/html]
Saving to: `pr0901.htm@uid=20130909-2709e'

100%[======================================>] 12,628      10.5K/s   in 1.2s

2013-11-02 13:12:54 (10.5 KB/s) - `pr0901.htm@uid=20130909-2709e' saved [12628/1
2628]
```

Figure 3. Data Collection using Wget tool.



Westinghouse, Toshiba And Exelon Nuclear Partners To Collaborate On Nuclear Project In Saudi Arabia - Mozilla Firefox

File Edit View History Bookmarks Tools Help
Westinghouse, Toshiba And Exelon Nuclear ...
www.dailymarkets.com
Google
HELPING YOU SAVE SMART, DAILY!
Home
Credit Cards
Mortgage Rates
Insurance
CD Rates
Credit Score

DailyMarkets.com » Financial News » Westinghouse, Toshiba And Exelon Nuclear Partners To Collaborate On Nuclear Project In Saudi Arabia

CREDIT CARDS

Best Credit Card Offers For 2013
Business Credit Cards
Low Interest Credit Cards
American Express Cards
Bank of America Cards
Capital One Cards
CitiCards
Chase Credit Cards
Discover Card
US Bank Credit Cards
Barclays Credit Cards
Visa Cards
Credit Card Rewards
Cash Back Credit Cards
Airline Credit Cards
Balance Transfer Cards
Travel Credit Cards
Student Credit Cards
Gas Credit Cards

Westinghouse, Toshiba And Exelon Nuclear Partners To Collaborate On Nuclear Project In Saudi Arabia

PRWeb
Updated September 9, 2013 | Author's Website
Tweet

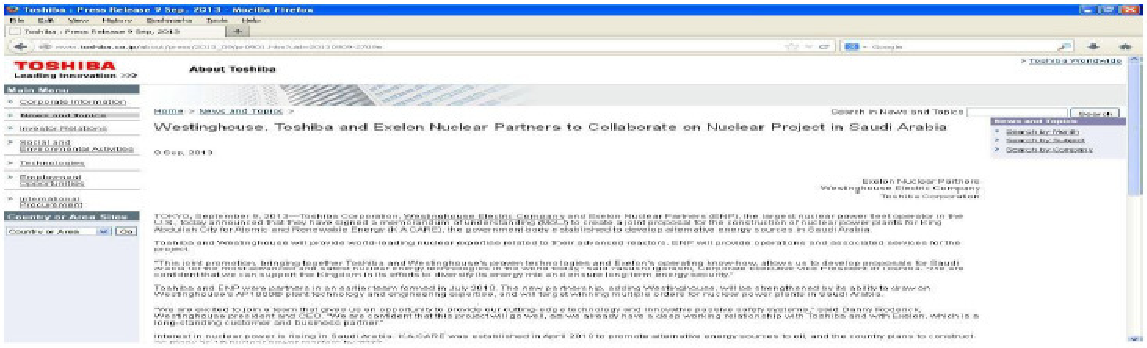
Westinghouse Electric Company, Toshiba Corporation and Exelon Nuclear Partners (ENP), the largest nuclear power fleet operator in the U.S., today announced that they have signed a memorandum of understanding (MoU) to create a joint proposal for the construction of nuclear power plants for King Abdullah City for Atomic and Renewable Energy (K.A.CARE), the government body established to develop alternative energy sources in Saudi Arabia.

Toshiba and Westinghouse will provide world-leading nuclear expertise related to their advanced reactors. ENP will provide operations and associated services for the project.

"This joint promotion, bringing together Toshiba and Westinghouse's proven technologies and Exelon's operating know-how, allows us to develop proposals

MORTGAGE RATES
INSURANCE QUOTES
Discover it Card
Blue Cash Everyday
Chase Freedom
American Express Platinum Card
Delta Credit Card
Blue Cash Preferred
BankAmericard Cash Rewards
BankAmericard Travel Rewards
Capital One Venture
Discover Card - 18 Month Balance Transfer

Figure 4. Data Collection Sample-1.



Toshiba : Press Release 9 Sep, 2013 - Mozilla Firefox

Home > News and Topics >

Westinghouse, Toshiba and Exelon Nuclear Partners to Collaborate on Nuclear Project in Saudi Arabia

9 Sep, 2013

Exelon Nuclear Partners
Westinghouse Electric Company
Toshiba Corporation

TOKYO, September 9, 2013—Toshiba Corporation, Westinghouse Electric Company and Exelon Nuclear Partners (ENP), the largest nuclear power fleet operator in the U.S., today announced that they have signed a memorandum of understanding (MoU) to create a joint proposal for the construction of nuclear power plants for King Abdullah City for Atomic and Renewable Energy (K.A.CARE), the government body established to develop alternative energy sources in Saudi Arabia.

Toshiba and Westinghouse will provide world-leading nuclear expertise related to their advanced reactors. ENP will provide operations and associated services for the project.

"This joint promotion, bringing together Toshiba and Westinghouse's proven technologies and Exelon's operating know-how, allows us to develop proposals for Saudi Arabia for the most advanced and safest nuclear energy technologies in the world today," said [illegible], Corporate Executive Vice President of Toshiba. "We are confident that we can support the Kingdom in its efforts to diversify its energy mix and ensure long-term energy security."

Toshiba and ENP were partners in an earlier team formed in July 2010. The new partnership, adding Westinghouse, will be strengthened by its ability to draw on Westinghouse's AP1000® plant technology and engineering expertise, and will target winning multiple orders for nuclear power plants in Saudi Arabia.

"We are excited to join a team that gives us an opportunity to provide our cutting-edge technology and innovative passive safety systems," said Danny Roderick, Westinghouse president and CEO. "We are confident that this project will go well, as we already have a deep working relationship with Toshiba and with Exelon, which is a long-standing customer and business partner."

Interest in nuclear power is rising in Saudi Arabia. K.A.CARE was established in April 2010 to promote alternative energy sources to oil, and the country plans to construct

Figure 5. Data Collection Sample-2.

5. RESULTS

Every dataset mentioned in the above section undergoes web page parsing, stop word removal and stemming process. It is observed that after applying the above mentioned steps the no. of keywords are reduced to an extent. The final output that is stemmed keywords are provided to the DD algorithm. The results are obtained by executing the DD program on the datasets which are shown in Table 2 and Table 3. These are representing the outcome obtained by checking every page of the dataset with all web pages from the database. As shown in Fig. 4 and Fig. 5, SimScore range is divided into 4 categories that <60%, 60% to 70%, 70% to 80% and >80%.

Table 2. Samples outcomes

Sample 1	Sample 2
Para 1: 23519 5125 31513 20158 31597 5245 14215 16119 12118 14215 16155 6125 15169 20154 1997 13513 21144 1315 3185 10155 16184 31514 14215 16155 16121 1195 12421 3120 12013 18524 5145 71522 2154 51912 45223 11220 5145 19153 19191 11812	**Para 1:** 20146 19572 20158 31597 23519 5125 31513 5245 14215 16119 12118 14215 16155 6125 15169 20154 1997 13513 21144 1315 3185 10155 16184 31514 14215 16155 16121 1195 12421 3120 12013 18524 5145 71522 2154 51912 45223 11220 5145 19153 19191 11812
Para 2 : 20158 23519 16189 23159 12514 14215 52416 18531 14221 18514 16189 15165 11915 19519 16171	**Para 2 :** 20158 23519 16189 23159 12514 14215 52416 18531 14221 18514 16189 15165 11915 19519 16171

Table 3: DD results for selected dataset

Dataset	Number of documents	< 60 %	60% to 70%	70% to 80%	>80%
Technical	w1	w10,w9,w8	w4,w6,w7	w3,w5	w2
	w2	w10,w8,w9	w4,w6,w7	w3,w5	w1
	w3	w4,w8,w9,w10	w6,w7	w1,w2,w5	--
	w4	w3,w5,w8,w9,w10	w1,w2,w6,w7	--	--
	w5	w4,w8,w9,w10	w6,w7	w1,w2,w3	--
	w6	w7,w8,w9,w10	w1,w2,w4,w3	w5	--
	w7	w6,w8,w9,w10	w1,w2,w3,w4,w5		--
	w8	w1,w2,w3,w4,w5,w6,w7,w9,w10			--
	w9	w1,w2,w3,w4,w5,w6,w7,w8,w10			--
	w10	w1,w2,w3,w4,w5,w6,w7,w8,w9			--

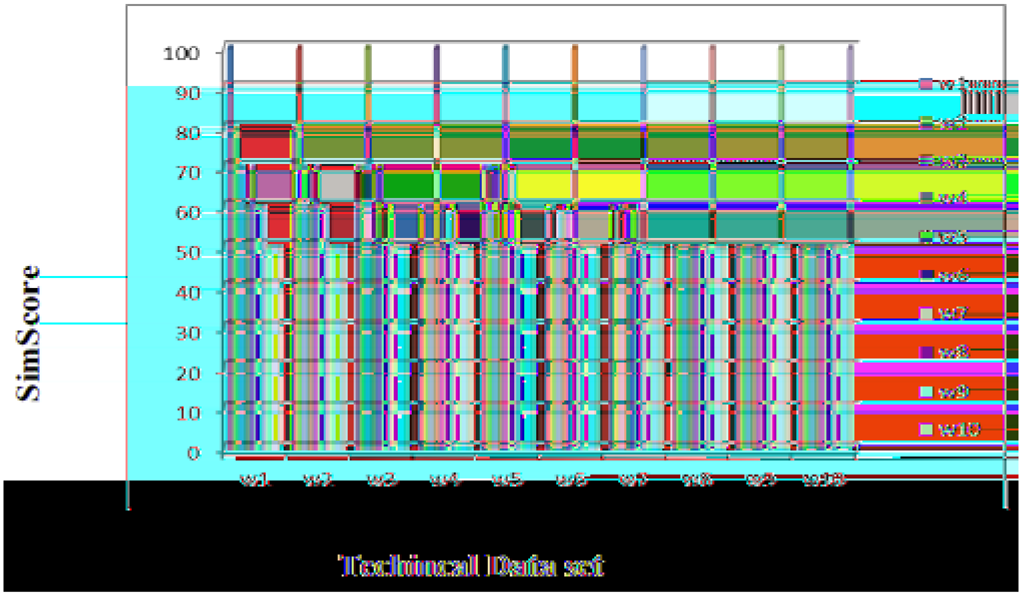


Figure 5. Simscore Illustrations for the datasets

6. Conclusions And Future Work

Web pages with similarity score of 0% represent totally unique documents. Web pages with similarity score of <60% are not similar and needs to be maintained in the database for future reference. Web pages with similarity score of 60% to 70% are suspicious near duplicates and needs to be maintained in the database for future reference. Web pages with similarity score of more than 70% are almost near duplicates.

Furthermore only top N pages are stored in the repository for each datasets instead of all documents which reduces memory space for the repository. Since web pages having SimScore greater than 60% are not stored in the repository, by detecting them as near duplicate web pages. The search efficiency and effectiveness will be achieved and document clustering can be facilitated.

Web pages which are near duplicates may appear closer to each other in search results, but provide very little benefit to the user. The future work will be research for more robust and accurate methods for near duplicate detection and elimination on basis of the detection. To increase the accuracy and effectiveness of the DD Algorithm a clustering mechanisms can also be applied to perform the effective document clustering.

REFERENCES

[1] P. Maes. Agents that reduce work and information overload. Communications of the ACM, 37(7):30–40, 1994.

[2] 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.

[3] Chuan Xiao, Wei Wang, Xuemin Lin, Efficient Similarity Joins for Near Duplicate Detection, Proceeding of the 17th international conference on World Wide Web, pp 131 – 140. April 2008. SimScore Techincal Data set

[4] 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, pp. 141 - 150, Banff, Alberta, Canada, 2007.

[5] Dennis Fetterly, Mark Manasse and Marc Najork, Detecting phrase-level duplication on the World Wide Web, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp.170 - 177, Salvador, Brazil, 2005

[6] A. Broder, on the resemblance and containment of documents, Proc. Compression and Complexity of Sequences (SEQS: Sequences97). pp. 21–29.

[7] A. Broder, S. Glassman, M. Manasse, G. Zweig, Syntactic clustering of the web, 6th International World Wide Web Conference, Apr. 1997, pp.393–404.

[8] A. Broder, M. Charikar, A.M. Frieze, M. Mitzenmacher, Min-wise independent permutations, Proc. STOC, 1998, pp. 327–336.)

[9] A. Broder, Identifying and filtering near-duplicate documents, Proc. Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science (eds. R. Giancarlo and D. Sankoff) 1848, 2000, pp.(1–10.)

[10] Broder, A. Z., Najork, M., and Wiener, J. L., 2003. “Efficient URL caching for World Wide Web crawling”, In International conference on World Wide Web.

[11] A. Chowdhury, O. Frieder, D.A. Grossman, M.C. McCabe, Collection statistics for fast duplicate document detection, ACM Transactions on Information Systems, 20, 2 (2002) 171–191.)

[12] B. Ganter, R. Wille, Formal concept analysis: mathematical foundations, Springer, Berlin, 1999.

[13] Lavanya Pamulaparty, C.V. Guru Rao, A Novel Approach to Perform Document Clustering Using Effectiveness and Efficiency of Simhash, ,International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013

[14] J. Cho, N. Shiva Kumar, H. Garcia-Molina, Finding replicated web collections, Proc. SIGMOD Conference, (2000) 355–366.) 216

[15] S. Brin, J. Davis, H. Garcia-Molina, Copy detection mechanisms for digital documents, 1995 ACM SIGMOD International Conference on Management of Data 1995, pp. 398–409.

[16] Jun Fan, Tiejun Huang “A fusion of algorithms in near duplicate document detection”.

[17] N. Heintze, Scalable document fingerprinting, Proc. of the 2nd USENIX Workshop on Electronic Commerce, 1996, pp. 191–200.) 216

[18] M. Henzinger, Finding near-duplicate web pages: a largeccale evaluation of algorithms, Annual ACM Conference on Research and Development in Information Retrieval, 2006, pp. 284–291.) 216

[19] T. C. Hoad, J. Zobel, Methods for identifying versioned and plagiarized documents, Journal of the American Society for Information Science and Technology, 54, 3 (2003) 203–215.) 216 232 D. Ignatov, K. J´anosi-Rancz, S. Kuznetsov

[20] Spertus, E., Sahami, M., and Buyukkokten, O., (2005) "Evaluating similarity measures: a large-scale study in the orkut social network", In KDD (2003), pp. 678-684.

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

[22] A. Kolcz, A. Chowdhury, J. Alspector, Improved robustness of signature-based near-replica detection via lexicon randomization, Proc. KDD'04(eds. W. Kim, R. Kohavi, J. Gehrke, W. DuMouchel) Seattle, 2004, pp.605–610.) 216, 217

[23] S.O. Kuznetzov, S.A. Obiedkov, Comparing performance of algorithms for generating concept lattices, Journal of Experimental and Theoretical Artificial Intelligence, 14 (2002) 189–216.) 221

[24] Hoad, T. C., and Zobel, J., (2003) "Methods for identifying versioned and plagiarized documents", JASIST, vol. 54, no. 3, pp.203-215

[25] U. Manber. Finding similar files in a large file system, Proc. of the USENIX Winter 1994 Technical Conference, 1994, pp. 1–10.) 216

[26] Gibson, D., Kumar, R., and Tomkins, A., (2005) "Discovering large dense sub graphs in massive graphs", Proceedings of the 31st international conference on Very large data bases, Trondheim, Norway, pp. 721-732.

[27] 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),ISSN: 2278-0181,Vol. 2 Issue 9, September – 2013

[28] Fetterly, D., Manasseh, M., and Najork, M., (2003) "On the evolution of clusters of near-duplicate web pages", Proceedings of the First Conference on Latin American Web Congress, pp. 37.

[29] Conrad, J. G., Guo, X. S., and Schreiber, C. P., (2003) "Online duplicate document detection: signature reliability in a dynamic retrieval environment", Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA, pp. 443-452.

[30] Bayardo, R. J., Ma, Y., and Srikant, R., (2007) "Scaling up all pairs similarity search", In Proceedings of the 16th International Conference on WorldWideWeb, pp.131-140.

[31] Cho, J., Shiva Kumar, N., and Garcia-Molina, H., (2000) "Finding replicated web collections", ACM SIGMOD Record, Vol. 29, no. 2,pp. 355-366.

[32] W. Pugh, M. Henzinger, Detecting duplicate and near-duplicate files, United States Patent 6658423 (December 2, 2003).) 216, 217

[33] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Efficient mining of association rules using closed item set lattices, Inform. Syst., 24, 1 (1999) 25–46) 220

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

[35] Y. Bernstein, J. Zobel ―Accurate discovery of co-derivative documents via duplicate text detection‖ Information Systems 31(2006) 595–609 Elsevier. doi:10.1016/j.is.2005.11.006.

AUTHORS

Lavanya Pamulaparty, Associate Professor and Head of the Dept. Dept of CSE, MCET, Osmania University, Hyderabad, obtained her Bachelor's degree in computer science from Nagpur University of KITS, Nagpur, India, and Masters Degree in Software Engineering from School of Informatics from JNT University Hyderabad, India, and Pursuing the PhD degree in computer science and engineering from JNT University. Her research interests include information storage and retrieval, Web Mining, Clustering technology and computing, performance evaluation and information security. She is a senior member of the ACM, IEEE and Computer Society of India.

Dr. Guru Rao C. V received his Bachelor's Degree in Electronics &Communications Engineering from VR Siddhartha Engineering College, Vijayawada, India. He is a double post graduate, with specializations in Electronic Instrumentation and Information Science & Engineering. He is a Doctorate holder in Computer Science & Engineering from Indian Institute of Technology, Kharagpur, India. With 24 years of teaching experience, currently he is the Professor & Principal, SR Engineering College Warangal, India. He has more than 25 publications to his credit. He is a life member of Indian Society for Technical Education, Instrumentation Society of India and member of Institution of Engineers, Institution of Electronics & Telecommunications Engineers and Institution of Electrical & Electronics Engineers (USA).

Dr. M Sreenivasa Rao, Professor, School of Information Technology, JNT University, Hyderabad, obtained his Graduation and Post-graduation in Engineering from JNT University, Hyderabad and Ph D from University of Hyderabad. Over 28 Years of IT Experience in the Academia& industry. As a Dean of the MS IT Program, in association with Carnegie Melon University, USA. De- signed and conducted post graduations level MSIT program. Guided more than 10 research students in JNTU, and continuing the research in IT.