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