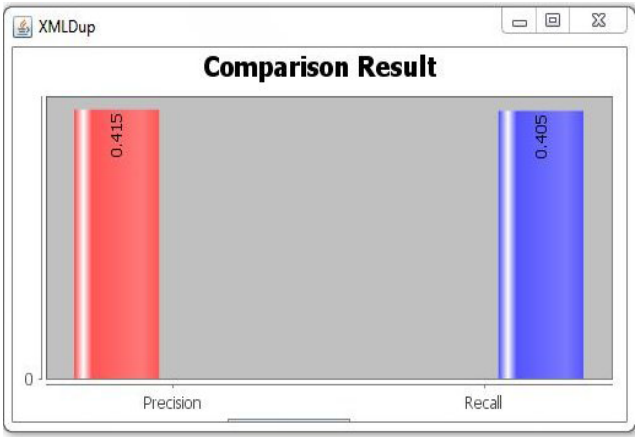

Figure 6: Duplicate Detection Using Normalized Edit Distance over Country Dataset

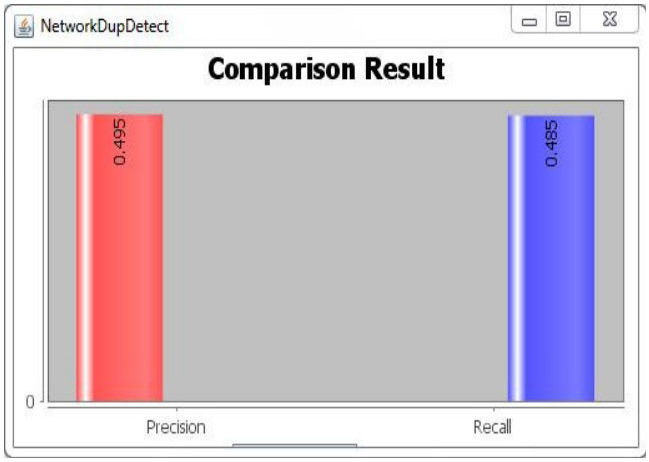

Figure 7: Duplicate Detection Using Levenshtein Distance Algorithm Over Country Dataset.

Why the Levenshtein distance algorithm is best for our System….

In approximate string matching, the main aim is to find matches for short length strings in many longer texts, where a small number of differences are to be expected. The short length strings that could come from a dictionary, for example here, one of the strings are short length, while the other is long string. This has a wide range of applications; for instance, spell checkers, correction systems for optical recognition of characters, and software to assist natural language translation based on translation memory.

The Levenshtein distance algorithm can also be computed between two longer length strings. The costs of computing distance, is proportional to the product of the two string lengths, which makes this impractical. Thus, when used to aid in fuzzy string searching in applications such as record linkage,to help improve speeds of comparisons the compared strings are usually short.

5. Conclusion

Levenshtein distance algorithm uses a Bayesian Network to determine the probability of two XML objects being Duplicates. The Bayesian Network model is composed from the structure of the objects being compared, thus probabilities of all objects are computed considering not only the information the objects contain, but also how such information is structured. Levenshtein distance algorithm is very flexible, which allows the use of different similarity measures and different ways of combining probabilities. Using Levenshtein Distance algorithm gives better result Than Normalized Edit Distance algorithm. Experiments performed on both artificial and real-world collected data showed that our algorithm is able to achieve high precision and recall scores in several contexts.

6. Acknowledgment

We thank the mysterious referees for their valuable suggestions to improvise the content and quality of this paper. The author is grateful to our principal for providing necessary facilities towards carrying out this work. We acknowledge the diligent efforts of our Head of the Department to guide us towards implementation of this review paper.

References

[1] Lu´s Leita˜ o, Pa´ vel Calado, Melanie Herschel “Efficient and Effective Duplicate Detection in Hierarchical Data”, Knowledge and Data Engineering, IEEE Transactions, Volume: 25, Issue: 5, ISSN: 1041-434,May 2013.
[2] Yuan Wang David J. DeWitt Jin-Yi Cai “Local X-Diff: An Effective Change Detection Algorithm for XML Documents” Data Engineering, 2003. Proceedings. 19th International Conference, ISBN: 0-7803-7665, March 2013.
[3] Diego Milano, Monica Scannapieco, Tiziana Catarci, “Structure-aware XML Object Identification”, in ’CleanDB’, 2006.
[4] Adrovane Marques Kade, Carlos Alberto Heuser “Matching XML Documents in Highly Dynamic Applications”, DocEng '08 Proceedings of the eighth ACM symposium on Document engineering’, ISBN: 978- 1-60558-081-4, 16 Sep 2008.
[5] Erhard Rahm, Hong Hai Do “Data Cleaning: Problems and Current Approaches”, IEEE Data Eng. Bull.23, no. 4 (2000): 3--13.
[6] Rohit Ananthakrishna, Surajit Chaudhuri, Venkatesh Ganti ba “Eliminating Fuzzy Duplicates in Data Warehouses”, VLDB '02Proceedings of the 28th international conference on Very Large Data Bases, August 2002.
[7] Melanie Weis, Felix Naumann, and Franziska Brosy, “A Duplicate Detection Benchmark for XML (and Relational) Data”, Proc. Of Workshop on Information Quality for Information Systems (IQIS),2006.
[8] zur Erlangung des akademischen “Duplicate Detection in XML Data .
[9] Le Chen, Lei Zhang, Feng Jing, Ke-Feng Deng and Wei-Yeing Ma,“Ranking Web Objects from Multiple Communities”, CIKM '06Proceedings of the 15th ACM international conference on Information and knowledge management, Pages 377-386, ISBN:1-59593-433-2, 2006.
[10] Melanie Weis and Felix Naumann, “DogmatiX Tracks down Duplicates in XML”, SIGMOD '05 Proceedings of the 2005 ACM SIGMOD Dinternational conference

on Management of data, ISBN: 1-59593-060-4.June 2005.

[11] Dmitri V. Kalashnikov and Sharad Mehrotra, "Domain-Independent Data Cleaning via Analysis of Entity-Relationship Graph", ACM TODS Journal, Vol. 31(2), June 2006.

[12] Zaiqing Nie, Yuanzhi Zhang, JiRon Wen, WeiYing Ma, "Object Level Ranking: Bringing Order to Web Objects", 05 Proceedings of the 14th international conference on World Wide Web, ISBN: 1-59593-046-9 May 2005.G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[13] L. Leita˜o, P. Calado, and M. Weis, "Structure-Based Inference of XML Similarity for Fuzzy Duplicate Detection," Proc. 16th ACM Int'l Conf. Information and Knowledge Management, pp. 293-302,2007.

[14] R.A. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison-Wesley Longman Publishing Co., Inc., 1999.