Different Blocking approach for end-to-end Entity Resolution

Viral A. Parekh
Department of Computer Engineering
C. U. Shah University, Gujarat, India

Dr. K. H. Wandra
Gujarat Maritime Board, Gujarat Technological University, India

Abstract: An end-to-end entity resolution aims to find out duplicates among clean-clean datasets as well as from dirty datasets. It consists of various phases like blocking, block processing, matching and clustering to find out duplicates. In this paper we have discussed a different approach for blocking and also discussed few results which supports that proposed blocking approach helps to improve recall in end-to-end entity resolution.

I. INTRODUCTION

Entity resolution is used to find out duplicates from the given input data collection. In end-to-end ER, Blocking helps to handle multiple characteristics of big data (i.e. variety and velocity, or volume and variety) [4]. Block processing further used to reduce redundant and unnecessary comparisons among input entities. Matching uses a match function to find out duplicates. Clustering groups all the matching entities into a single group and forms a cluster [4].

To improve precision, blocking methods can be modified. Here, main purpose of this research is to improve recall without much affecting precision. We have proposed a method, in which in blocking phase, if more than one blocking methods are combined, recall is improved in end-to-end entity resolution. Our result shows that for both, clean-clean ER and dirty ER, this blocking approach improves the results and for small datasets recall is achieved almost near to one. Proposed approach is implemented with the help of an open source toolkit for end-to-end ER [7].

II. RESULTS AND DISCUSSION

For experiments various real datasets are taken into consideration. Clean-Clean ER as well as Dirty ER results are taken for different datasets. The experiments were performed on a server with an Intel(R) Core(TM) i7-4770K CPU @ 3.50GHz and 8.00 GB RAM, running Windows 7. Here only result of Dirty-ER for dataset Cora is discussed. The Cora dataset consists of 2708 scientific publications classified into one of seven classes [3].

Here, for block building Extended Q-Grams Blocking is used individually and then it's combined with Extended Suffix Arrays Blocking. Here meta-blocking is also used. Then results are taken for both individual blocking and combined blocking for various clustering algorithms including Center Clustering, Connected Component Clustering, Cut Clustering, and Ricochet SR clustering. Various parameters are taken into consideration for taking all the results. Manual fine tuning of few parameters is also required.

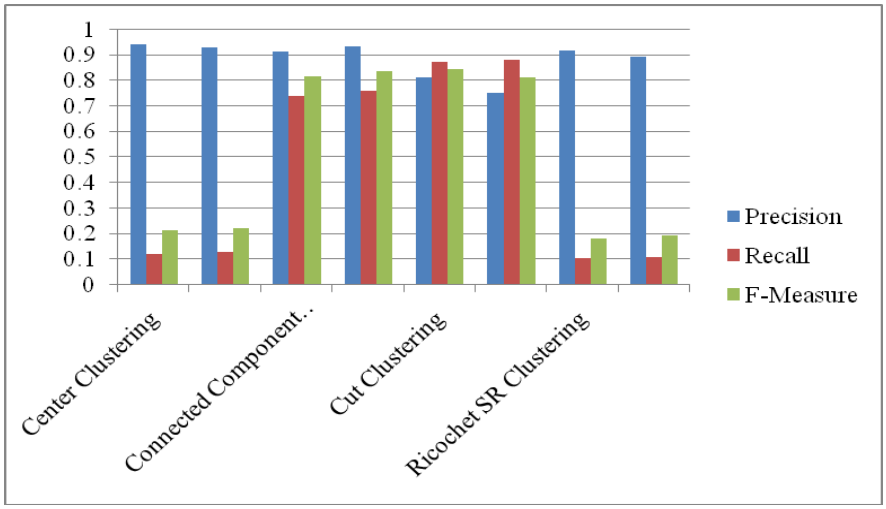

Figure 1: Comparison of precision, recall and F-Measure

Figure 1 shows the comparison chart of the results achieved after performing all experiments. Here, results show that for all clustering algorithms, recall is improved when more than one blocking is combined. Cut Clustering takes maximum time among all clustering algorithms for execution, but when individual blocking is used, Cut Clustering gives maximum F-measure, that shows that for Cora Profiles, Cut Clustering balances among precision and recall.

III. CONCLUSION AND FUTURE SCOPE

When more than one blocking methods are combined then there is an improvement can be seen in recall. For different datasets these results may vary. In future, this phenomenon can be implemented in parallel environment. Even this may also help to study the nature of different datasets.

REFERENCES

[1] Zhang, Fulin & Gao, Zhipeng & Niu, Kun. (2017). A pruning algorithm for Meta-blocking based on cumulative weight. Journal of Physics: Conference Series. 887. 012058. 10.1088/1742-6596/887/1/012058.

[2] Oktie Hassanzadeh, Fei Chiang, Hyun Chul Lee, and Renée J. Miller. 2009. Framework for evaluating clustering algorithms in duplicate detection. Proc. VLDB Endow. 2, 1 (August 2009), 1282–1293. DOI:https://doi.org/10.14778/1687627.1687771

[3] George Papadakis, George Alexiou, George Papastefanatos, and Georgia Koutrika. 2015. Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data. Proc. VLDB Endow. 9, 4 (December 2015), 312–323. DOI:https://doi.org/10.14778/2856318.2856326

[4] Christophides, Vassilis & Efthymiou, Vasilis & Palpanas, Themis & Papadakis, George & Stefanidis, Kostas. (2019). End-to-End Entity Resolution for Big Data: A Survey.

[5] Christen, Peter. (2011). A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. IEEE Transactions On Knowledge And Data Engineering. 24. 10.1109/TKDE.2011.127.

[6] Papadakis, George & Papastefanatos, George & Koutrika, Georgia. (2014). Supervised Meta-blocking. Proceedings of the VLDB Endowment. 7. 10.14778/2733085.2733098.

[7] George Papadakis, Leonidas Tsekouras, Emmanouil Thanos, George Giannakopoulos, Themis Palpanas, Manolis Koubarakis. The return of JedAI. PVLDB, 11 (5): 2150-8097, 2018. DOI: https://doi.org/TBD.

[8] Ruhaila Maskat, Norman W. Paton, and Suzanne M. Embury. 2016. “Pay-as-you-go Configuration of Entity Resolution”. In Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIX - Volume 10120. Springer-Verlag, Berlin, Heidelberg, 40–65. DOI:https://doi.org/10.1007/978-3-662-54037-4_2.

[9] George Papadakis, Jonathan Svirsky, Avigdor Gal, and Themis Palpanas. 2016.” Comparative analysis of approximate blocking techniques for entity resolution”. Proc. VLDB Endow. 9, 9 (May 2016), 684–695. DOI:https://doi.org/10.14778/2947618.2947624.

[10] George Papadakis, Leonidas Tsekouras, Emmanouil Thanos, George Giannakopoulos, Themis Palpanas, and Manolis Koubarakis. 2020. “Domain- and Structure-Agnostic End-to-End Entity Resolution with JedAI”. SIGMOD Rec. 48, 4 (December 2019), 30–36. DOI:https://doi.org/10.1145/3385658.3385664.