Entity Resolution for Big Data using Combination of Supervised Meta-Blocking and pay-as-you-go Configuration

Viral A. Parekh, K. H. Wandra

Abstract*—Entity resolution refers to the method of identifying the same real world object from multiple data sets. In Data cleaning and data integration application, entity resolution is an important process. When data is large the task of entity resolution becomes complex and time consuming. End-to-end entity resolution proposal involves stages like blocking (efficiently identifies duplicates), detailed comparison (refines blocking output) and clustering (identifies the set of records which may refer to the same entity). In this paper, an approach for feedback based optimization of complete entity resolution is proposed in which supervised meta-blocking is used for blocking stage. This paper proposes a technique for entity resolution which does optimization of each phase of entity resolution with benefits of supervised Meta-blocking to improve performance of entity resolution for big data.

I. INTRODUCTION

Entity resolution is first proposed by [8]. Entity resolution (ER) seeks to find out entity profiles which refer to the same real world object from different data collections. The demand for entity resolution and its optimization increases as the amount of data increases exponentially. Entity resolution is also an important and challenging AI problem for structured, semi-structured and unstructured data sets. With the increased use of internet and social media, everyday data more than in terabytes is generated. Such a big volume data sets need some extra and special efforts for analysis. To analyze and identify similar record sets from large data sets, optimization may be used to increase efficiency. Optimization of Entity Resolution is required for time efficient data analysis. For effective data analysis, integration or cleaning in many diverse applications, identification of records that refer to the same entity is very important. Three important phases of entity resolution are likely: (i) *blocking*, in this phase, with the use of low-cost, approximate comparison schemes, pairs of records which are candidate duplicates are identified which helps to improve the run-time of entity resolution; (ii) *detailed comparison*, in this phase properties ofpairs of records which are candidate duplicates that are identified from blocking are compared using distance function; and (iii) *clustering*, in this phase, on the basis of the results of the distance function, candidate duplicates that are identified through blocking are grouped into clusters [1].The quality of the entity resolution's result is significantly depends on configuration parameters used in all of these three phases and there may be subtle inter-dependencies between these parameters. Lots of work is done on each of these phases [3, 4].

The main goal of blocking is to remove all redundant comparisons and to avoid most superfluous comparisons while maintaining precision and recall. Though blocking improves execution time of entity resolution, still it involves some unnecessary comparisons due to which performance of entity resolution may be affected. Much work is done to improve functionality of blocking phase. This paper proposes a technique to optimize complete lifecycle of entity resolution. With optimization, for blocking phase supervised Meta-blocking [2] is proposed to enhance performance of entity resolution.

II. RELATED WORK

This section discusses mainly work done for optimization of various phases of entity resolution and supervised meta-blocking in entity resolution. Only little work is done to co-optimize whole life-cycle of entity resolution. Previously probably Corleone [5] searches for optimizing completeentity resolution process, that optimizes different facets of an entity resolution pipeline in sequence instead of together, emphasizing comparison functions instead of wider configuration parameters, and instead of clustering concentrating on pairwise comparison [1]. Also Ruhaila [1] describes pay-as-you-go approach for complete entity resolutionconfiguration. In this approach, users can provide as little or as much feedback as they like on candidate duplicates (either to accept or reject candidate duplicates), the results are reviewed in the light of the feedback provided to date and additional feedback is supplied until users are satisfied [1]. This maximizes the quality of the resulting clusters.

The term pay-as-you-go has different meanings in relation to entity resolution. In [6] the payment is in the form of computational resource usage, whereas in this paper, the payment takes the form of feedback on the result of entity resolution process [1].

Meta-blocking is a method which transforms a redundancy positive block collection which it takes as an input into a new block collection that generates fewer comparisons, but keeps most of the detected duplicates [7]. In meta-blocking a block collection is restructured in order

Revised Manuscript Received on April 12, 2019.
Viral A. Parekh,C. U. Shah University, Gujarat, India.(viral.ccet@gmail.com)
Dr. K. H. Wandra,Gujarat Maritime Board, Gujarat Technological University, India.(khwandra@rediffmail.com)

to prune unnecessary comparisons to improve run-time of entity resolution.In supervised Meta-blocking three aspects are identified and examined to determine the performance: (i) the set of features annotating the edges of the blocking graph, (ii) the training data, and (iii) the classification algorithm and its configuration [2]. George [2] has shown that supervised techniques for blocking are robust to different classifiers and their configurations. Supervised techniques show better time efficiency than the best alternatives, for achieving equivalent recall [2]. For some pruning methods, Supervised Meta-blocking scales better than the unsupervised one with respect to both effectiveness and efficiency [2]. It is also proved that Multi-core Meta-blocking and parallel meta-blocking also improves performance of an ER.

III. PROPOSED WORK& RESULTS

In big data the data is large in volume, variety and velocity. It is very important to identify same real world object from different data sets and merge similar object into single entity. Data may come from heterogeneous knowledge bases or data sets. Entity resolution is an expensive process for large data sets. Also in big data era, ER is known as uncertain ER. The main ER process consists of several phases like blocking, detailed comparison of record pairs, classification, clustering and merging of record pairs. To improve the performance of entity resolution, pay-as-you-go approach is used to optimize various phases of ER. Pay-as-you-go is an approach based on feedback that is received from crowds on candidate duplicates for matched and unmatched entities and based on feedback further improvement is done. Feedback may be implicit (by observing user's regular behavior and interaction with system) or explicit [1]. Feedback for matched entities as well as for plausible unmatched entities is important to take necessary decision [1]. This may require human intervention for entity resolution process, though ER may be an automatic process. Feedback is generated algorithmically from the ground truth [1].

For supervised meta-blocking, blocking graphs are taken into consideration to find matching pairs more efficiently. This paper proposes a method in which for each phase of ER, optimization is done as discussed in [1] except for blocking. In [2], it is proved that even with small training sets, supervised meta-blockingcan achieve high performance, and also verified that most configurations of establishedclassification algorithms have little impact on the overallperformance. So by supervised meta-blocking with pay-as-you-go configuration in big data, performance of entity resolution can be improved. There are many tools available for big data mining and entity resolution. The proposed work will be implemented and verified using some real and synthetic big data sets to prove that pay-as-you-go approach with supervised meta-blocking helps to improve performance of entity resolution for big data.

IV. CONCLUSION

For Big data analysis, entity resolution is one of the most difficult tasks. Entity resolution is also time consuming process for big data. Performance of entity resolution for big data can be improved using pay-as-you-go approach and considering feedback as payment with supervised meta-blocking. In future this algorithm will be implemented to test various quality parameters like time, precision and recall to measure efficiency of proposed work.

REFERENCES

1. RuhailaMaskat, Norman W. Paton, Suzanne M. Embury,"Pay-as-you-go Configuration of Entity Resolution", Springer Trans. Large-Scale Data- and Knowledge-Centered Systems XXIX, 2016, VII, 135p, 40-65.
2. George Papadakis, George Papastefanatos, Georgia Koutrika, "Supervised Meta-Blocking", Proc. VLDB Endowment, Vol.7, Issue 14, 1929-1940, 2014
3. P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication". IEEE Trans. Knowl. Data Eng. 24(9), 1537–1555 (2012).
4. [4]. O. Hassanzadeh, F. Chiang, H. C. Lee, R. J. Miller, "Framework for evaluating clustering algorithms in duplicate detection", Proc. VLDB Endow. 2(1), 1282–1293 (2009).
5. C. Gokhale, S. Das, A. Doan, J. F. Naughton,N. Rampalli, J. W. Shavlik, X. Zhu, "Corleone: hands-off crowdsourcing for entity matching", In: SIGMOD Conference, pp. 601–612 (2014).
6. S. E. Whang, D. Marmaros, H. Garcia-Molina, "Pay-as-you-go entity resolution", IEEE Trans. Knowl. Data Eng. 25(5), 1111–1124 (2013).
7. G. Papadakis, G. Koutrika, T. Palpanas, W. Nejdl, "Meta-blocking: Taking entity resolute onto the next level", IEEE Trans. Knowl. Data Eng., 26(8):1946–1960, 2014.
8. H. B. Newcombe,J. M. Kennedy, S. J. Axford, A. P. James, "Automatic Linkage of Vital Records", 1959 Science 130.3381:954.