MANIPAL UNIVERSITY JAIPUR

IEEE

Technically Co-Sponsored by IEEE computer society

Supported TCPP

Microsoft

MANIPAL UNIVERSITY JAIPUR

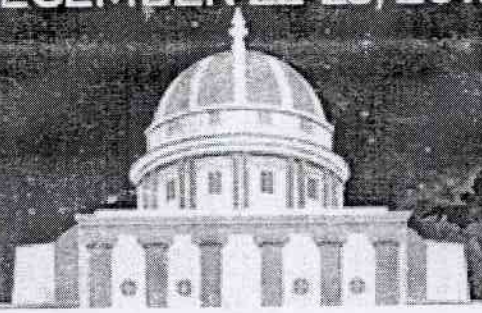
WELCOMES YOU

TO

INTERNATIONAL CONFERENCE ON INTELLIGENT COMMUNICATION AND COMPUTATIONAL TECHNIQUES.

ICCT 17

DECEMBER 22-23, 2017



Author Index

2017 International Conference on Intelligent Communication and Computational Techniques (ICCT)
Manipal University Jaipur, Dec 22-23, 2017

Efficient Entity Resolution using Multiple Blocking Keys for Bibliographic Dataset

Dolly Mittal*†, Emmanuel S Pilli*, Madhuri Gupta‡
*†Department of Computer Science and Engineering
*Malaviya National Institute of Technology, Jaipur
†Swami Keshvanand Institute of Technology, Jaipur
Rajasthan 302017
‡Formcept Technologies and Solutions, Bangalore
Karnataka 560076
Email: 2014pcp5269@mnit.ac.in, espilli.cse@mnit.ac.in, madhuri.gupta@formcept.com

***Abstract*—Entity Resolution(ER) is defined as identifying entities referring to the same real world object. The standard entity resolution process compare each entity with all other entities, which is inefficient for large datasets. A significant challenge in ER is to reduce the search space and execution time. The aim of this paper is to provide efficient entity resolution implementation for massive dataset by combining the use of multiple blocking key and parallel and distributed computing. In multiple blocking key concept, a record can belong to multiple blocks and it is possible that a record pair is generated multiple times for matching task. A solution to eliminate these duplicate pair is proposed, in addition to this character based similarity measure on sorted tokens is used for computing similarity between two record in the matching task. Efficient partitioning technique is used to remove the limitations of skewed dataset and matching task are evenly distributed among all the reducer. We used a bibliographic dataset in our experiment to show that our approach is less time consuming and scalable.**

Keywords*—MapReduce, Entity Resolution, Blocking, Hybrid similarity

I. Introduction

Entity Resolution is the task of identifying entities referring to the same real world object. For example to find duplicate entries of customer in an enterprise database or to match product offers for price comparison portals. Massive amount of data is being generated by many real world domains like social networking , pharmaceutical and healthcare, telecommunications, E-Commerce websites. For Efficient processing, management and analysis of large data collections various novel techniques are required [1]. That is why fields like data warehousing and data mining have gained popularity in both academia and industry.

Entity Resolution also known as entity/object matching, data deduplication, record linkage, merge/purge [2] is a complex problem and has a significant impact on data quality and data integration. Entities used in ER mainly refer to people, such as patients, customers, taxpayers, or travellers, but they are also applicable for consumer goods and products, publications or citations, or businesses. In the era of digitization research publications from academia and industry are available electronically and through online databases such as IEEE Xplore, Google Scholar, Scopus, CiteSeerX etc. This digitization of research sets a new direction for research scholars as they are able to access millions of publications. These online databases are also providing various services like analyses of impact of a particular publication , alert messages for new publications by an author, and notifications of new citations for given publications.

The biggest challenge in creating and maintaining these bibliographic databases is that it is very likely that several researchers may have same surname, same initials in a database, some even might be working in the same research domain. Sometimes it is difficult to decide whether two publications written by same individual or not when they are same authors, affiliated to same university even if full given names are provided. In addition to this, journal and conference names are also in the abbreviated form and do not follow a standardised format, and therefore different variations of the same publication reference can often be found.

Various approach, so far have been proposed for entity resolution. The standard ER process compare each entity with all other entities, resulting into a complexity of $O(n^2)$, which is inefficient for large datasets. So we need some mechanism which partitions the input data set into smaller blocks. The approach used to improve efficiency of this process is to reduce the search space by adopting blocking techniques based on blocking key. After blocking step partitions the input dataset, similarity computations are performed only among the entities within the same block. For example, first two character of the title of the article concatenated by first two later of author name or publisher of the article can be used as a blocking key for bibliographic databases.

But, the concept of single blocking key provides incomplete results which can be improved using the concept of multiple blocking keys, as the probability of identifying two similar entities increased if they share multiple blocking keys. In this approach, multiple blocking keys are used to partition the input dataset into multiple blocks, wherein an entity might belong to multiple blocks. So, if two similar entities do not match for the first blocking key, there is possibility that they may match for another blocking key. This approach decreases the chance of missing a similar entity pair but suffers from duplicated pair matching tasks, a same entity pair might belong to several blocks and comparison is performed for each block.

Entity resolution or data deduplication is data intensive task and very expensive process computationally that might

978-1-5386-3030-3/17/$31.00 ©2017 IEEE