

MANIPAL UNIVERSITY  
JAIPUR



IEEE

Technically Co-Sponsored by  
IEEE  
computer  
SOCIETY

Support  
TCP



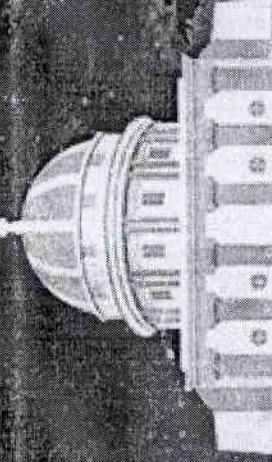
Microsoft



MANIPAL UNIVERSITY JAIPUR  
WELCOMES YOU  
TO  
INTERNATIONAL CONFERENCE ON INTELLIGENT  
COMMUNICATION AND COMPUTATIONAL TECHNIQUES.

**ICCT 17**

DECEMBER 22-23, 2017





## Author Index

A. K. Daniel	223	Kirti Kaur Sahota	228
Aditya Vats	200	Kuheli Pramanik	65
Afnan Salem Babrahem	182	Kuntam Babu Rao	17
Ahmed Qasim Mohammed	12	Lalit Kumar Awasthi	228
Aitha Nagaraju	102	Latika Singh	1
Ajay Gupta	262	M Abdul Rahman	123
Ajit Kumar	285	M. N. Giri Prasad	79,205
Amit Verma	86	M. Sridevi	200
Anamika Jain	138	Madhu Jain	138
Anil K Roy	90	Madhumita Das	217
Anil Kumar	189	Madhuri Gupta	108
Anil Swarnkar	252	Mamta Jain	189
Anita Shrotriya	256	Manju Mandot	36
Anju R	61	Manpreet Singh	239
Ankit Jyothish	200	Markus Hofmann	1
Anshul Garg	48	Milind Lalwani	200
Anshul Kumar	169	Mohammad Shahid	102
Anupam Ghosh	55	Mohit Mohta	32
Anurag Vidyarthi	169	Muhammad Mostafa Monowar	182
Anushree Pillai	26	N. M. Devashrayee	8
Aparajita Datta Sinha	65	Neelam Sharma	242
Ashikali M Hasan	90	Neeraj Kanwar	252
Asmita Rajawat	32	Neha Kashyap	36
Atul Kumar	96	Nikhil Gupta	252
Avinash Sharma	158	Nikita Bakshi	145
Avireni Srinivasulu	165	Oinam Robita Chanu	26
Bharti Nathani	119	P. Sudhakara Reddy	79
Bhupendra Singh	86	Pankaj Jain	242
Bindu Garg	48	Paramita Chowdhury	65
Brinda Chanv	151	Parminder Kaur	233
D. Satyanarayana	205	Partho Mallick	55
Devershi Pallavi Bhatt	71	Perla Anitha	79
Divyakant T. Meva	90	Piyanka Das	26
Dolly Mittal	108	Pooja Chaturvedi	223
Durjoy Majumder	194, 217	Pooja Jain	279
E Ilavarasan	42	Pooja V. Garach	272
Emmanuel S Pilli	108	Prachi Natu	173,177
Fatma Mubarak Said Al Siyabi	248	Prashant Vats	36
Gaurav Agrawal	279	Praveen Kumar Agrawal	138
Gurvir Kaur	285	Prem Sankar AU	211
Hardeep Singh	233	Priyanka Seth	55
Hargeet Kaur	96	Priyesh. P. Gandhi	8
Harpreet Singh Gill	132	Probir K. Dhar	194
Harsh Kumar Verma	114,228	Rajesh Bharati	12
Horesh Kumar	256	Rajni Goyal	23
Ikkurthy Kavya Sri	165	Rehna Kalam	123
Inderdeep Kaur	233	Rekha Vijayvergia	119
Jayanta Poray	55	Rhythm Bhatia	239
Jignesh Doshi	90	Richa Jain	242
K. R. Niazi	252	Rikin Thakkar	272
Karush Suri	32	Sagar Mal Nitharwal	114

Salina Adinarayana	42
Sandip Kumar Goyal	158
Sankaranarayanan N	200
Saurabh Ranjan Srivastava	132
Shachi Natu	173,177
Shubhi Shrivastava	61
Sindhu Hak Gupta	32
Smitha Sunil Kumaran Nair	248
Soman KP	211
Sophiya Sheikh	102
Spandan Sinha	26
Subarna Bhattacharya	65
Sujadevi VG	211
Sukhdev Singh	285
Suman Bhakar	71
Sunil Bakhru	151
T. Manasaveena	205
Tanuja Sarode	173, 177
Tapan Kumar	279
Tarun Jain	256
Tarun K. Naskar	194
Vandan Bhatia	239
Vibha Prabhu	145
Vijay Mehta	151
Vinay B Gavirangaswamy	262
Virendra Singh Kushwah	158
Vivek Kumar Verma	256
Yash Sharma	279
Zakiya Juma Khulaif Al Riyami	248



# Efficient Entity Resolution using Multiple Blocking Keys for Bibliographic Dataset

Dolly Mittal<sup>\*†</sup>, Emmanuel S Pilli<sup>\*</sup>, Madhuri Gupta<sup>‡</sup>

<sup>\*†</sup>Department of Computer Science and Engineering

<sup>\*</sup>Malaviya National Institute of Technology, Jaipur

<sup>†</sup>Swami Keshvanand Institute of Technology, Jaipur  
Rajasthan 302017

<sup>‡</sup>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