INTERNATIONAL CONFERENCE

ON

COMMUNICATION, COMPUTING AND NETWORKING (ICCCN-2017)

CERTIFICATE OF PARTICIPATION

This is to certify that Dr./Mr./Ms. ***Dolly Mittal*** has participated in the International Conference organized by Department of Computer Science & Engineering held at National Institute of Technical Teachers Training & Research (NITTTR), Chandigarh, India during 23-24 March, 2017. He/She has participated/~~presented~~ the paper titled ***Entity Resolution using Multiple Blocking Keys and Hybrid Similarity***.

Dr. Rakesh Kumar
Coordinator

Dr. C. Rama Krishna
Conference Chair

Dr. P. K. Tulsi
Patron

INDEX

Volume I

Volume II

TRACK 5: Digital Signal and Image Processing **Invited Speaker & Session Chair**: Prof. Renu Vig, UIET, Panjab University, Chandigarh		
ICCCN-17-105	**A Review on GSA-FODPSO-SVM Based Feature Selection Algorithm** Purnima Gaba, Surjeet Singh, K.K. Paliwal	528-533
ICCCN-17-112	**Implementing Three Way Bit Wise Separation of the Speech Signal Based on Element Position in the Speech Signal** Divya Sharma	534-537
ICCCN-17-113	**Implementing Odd and Even Element Separation for Encryption of Speech Signal using MATLAB** Divya Sharma	538-542
ICCCN-17-131	**MIMO-OFDM PAPR Reduction using Fusion of Techniques** Raj Lakshmi Shukla, Monika Singh	543-548
ICCCN-17-146	**Advancement in Adaptive Algorithm for Non- Stationary Noise Cancellation** Shelly Garg, Heena gupta, Reecha sood	549-553
ICCCN-17-150	**Comparative Study of Efficient Face Recognition Methods: A Literature Survey** Shashi Kant Sharma, Maitreyee Dutta, Kota Solomon Raju	554-559
ICCCN-17-157	**Analysis of Gradient Based Edge Detection Operators** Jaspreet Kaur, Anand Sharma	560-563
ICCCN-17-185	**Implementing Software Defined Radio using USRP and GNU Radio** Nayan Basumatary	564-570
ICCCN-17-188	**A Historical Review of Image Forgeries** Savita Walia, Krishan Kumar and Deva Prasad	571-576
ICCCN-17-192	**Capacity Enhancement of MIMO System over Fading Channels: A Review** Shivani, Amita Soni	577-581
ICCCN-17-199	**A Hybrid CEC Approach For Estimation Of Color Light Source** Harpreet Kaur, Sandeep Sharma	582-588
ICCCN-17-203	**A Hybrid Approach For Stereo Matching Based On Joint Trilateral Filter & Improved Mean-Shift** Deepika Kumari, Kamaljit Kaur	589-594
ICCCN-17-212	**Determination of Light Direction using 3D Vector Analysis of Known Shapes in Images** Deva Prasad, Krishan Kumar, SavitaWalia	595-600
ICCCN-17-213	**A Survey on Nature Inspired Approaches for Brain Tumor Segmentation** Khyati, Amit Doegar, Mamta Juneja	601-607
ICCCN-17-232	**QoS Discrepancy Impact Aware Best Fit Channel Selection and Void Filling by Burst Segmenting and Scheduling** Venkata Rao Tavanam, B.N Bhandari, Karuna Sagar Dasari	608-616
ICCCN-17-244	**Performance Comparison of PAPR Reduction Techniques** Asra Mubeen, S. Shalini Priya, CH. Nipun, P.Sisira, B.N. Bhandari	617-623
ICCCN-17-253	**Image Processing Technique to Determine Quality Aspect of Microwave Processed Paneer Prepared by Automated Press** Chitranayak, M. Manjunatha, Mahesh Kumar G., Padma Priya, P.S. Minz, Amita Vairat	624-627
ICCCN-17-261	**A Survey of Gender Classification Techniques using Facial Characteristics**	628-635

Entity Resolution using Multiple Blocking keys and Hybrid Similarity

Dolly Mittal*†,Emmanuel S Pilli*, M.C. Govil*

*Department of Computer Science & Engg., Malaviya National Institute of Technology Jaipur

†Department of Computer Science Engg., Swami Keshvanand Institute of Technology Jaipur,

2014pcp5269@mnit.ac.in, espilli.cse@mnit.ac.in

***Abstract*—**Entity Resolution plays a significant role in data integration task and defined as identifying entities referring to the same real world object. 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. The approach used to improve efficiency of this process is to reduce the search space by adopting blocking techniques based on blocking key. 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. This approach decreases the chance of missing a similar entity pair but suffers from duplicated pair matching tasks.

Further, parallel and distributed execution framework MapRe- duce is used for efficient execution of this data intensive task of entity resolution. Efficient partitioning technique is used to remove the limitations of skewed dataset and matching task are evenly distributed among all the reducer.

So in this paper we propose an efficient MapReduce algorithm for entity resolution in bibliographic dataset using multiple blocking keys. Also, hybrid similarity measure is used to determine the similarity between two entities. We used a bibliographic dataset in our experiment to show that our approach is efficient as well as scalable.

Keywords—Map-Reduce,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, telecom- munications, 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 is a complex problem and has a significant impact on data quality and data integration. ER is an important step of data integration process. While entity match- ing is concerned with identifying and matching individual en- tities that refer to the same real world objects from several data sources, data integration is the overall process of integrating heterogeneous datasets or data repositories to provide a unified and clear view of the available data. 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. Take a bibliographic database for example, In the era of digitization research publications from academia and industry are available electronically and through online databases such as IEEE Xplore, Google Scholar etc, Knowledge.This digitization of research sets a new direction for research scholars as they are