



International Conference on Advanced Informatics for Computing Research

ICAICR 2021: **Advanced Informatics for Computing Research** pp 192–200

Efficient Entity Resolution for Bibliographic Data Using MapReduce

[Dolly Mittal](#) , [Anjana Sangwan](#) & [Veena Yadav](#)

Conference paper | [First Online: 25 June 2022](#)

122 Accesses

Part of the [Communications in Computer and Information Science](#) book series (CCIS, volume 1575)

Abstract

Entity resolution is a process to identify all objects/entities referring to the same real-world object. The standard process to extract matches is to compare every entity with all the available entities using some similarity computation technique. It results into inefficient quadratic time complexity for huge datasets. To reduce the number of comparisons for entity pairs, blocking techniques thus become necessary. The objective of this research is to provide an approach for entity resolution to overcome the challenges like scalability of big data, reduce the time for matching process and making it cost effective. An efficient implementation for entity resolution can be achieved by combining the use of blocking and distributed computing framework for massive datasets. The proposed hybrid blocking approach uses the MapReduce model because of it's particular partitioning

technique along with the solution to the data skew problem associated with traditional blocking approach.

Keywords

Entity resolution Blocking MapReduce

Partitioning

This is a preview of subscription content, [access via your institution](#).

▼ Chapter

EUR 29.95

Price includes VAT (India)

- DOI: 10.1007/978-3-031-09469-9_17
- Chapter length: 9 pages
- Instant PDF download
- Readable on all devices
- Own it forever
- Exclusive offer for individuals only
- Tax calculation will be finalised during checkout

Buy Chapter

▼ eBook

EUR 58.84

Price includes VAT (India)

- ISBN: 978-3-031-09469-9
- Instant PDF download
- Readable on all devices
- Own it forever
- Exclusive offer for individuals only
- Tax calculation will be finalised during checkout

Buy eBook

▼ Softcover Book

EUR 69.99

Price excludes VAT (India)

- ISBN: 978-3-031-09468-2
- Dispatched in 3 to 5 business days
- Exclusive offer for individuals only
- Free shipping worldwide
- [Shipping restrictions may apply, check to see if you are impacted.](#)
- Tax calculation will be finalised during checkout

References

1. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science & Business Media, Berlin (2012)
2. Ghemawat, S., Gobioff, H., Leung, S.T.: The google file system. In: Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, pp. 29–43. SOSP 2003, ACM, New York (2003)
3. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10. MSST 2010, IEEE Computer Society, Washington, DC, USA (2010).
<https://doi.org/10.1109/MSST.2010.5496972>
4. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
5. Newcombe, H.B.: Record linking: the design of efficient systems for linking records into individual and family histories. Am. J. Hum. Genet. **19**(3 Pt 1), 335 (1967)
6. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: a survey. Knowl. Data Eng.

7. Hernández, M.A., Stolfo, S.J.: The merge/purge problem for large databases. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pp. 127–138. SIGMOD 1995, ACM, New York (1995). <https://doi.org/10.1145/223784.223807>

8. Monge, A., Elkan, C.: An efficient domain-independent algorithm for detecting approximately duplicate database records (1997)

9. Aizawa, A., Oyama, K.: A fast linkage detection scheme for multi-source information integration. In: International Workshop on Challenges in Web Information Retrieval and Integration, pp. 30–39. IEEE (2005)

10. De Vries, T., Ke, H., Chawla, S., Christen, P.: Robust record linkage blocking using suffix arrays. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 305–314. ACM (2009)

11. Baxter, R., Christen, P., Churches, T., et al.: A comparison of fast blocking methods for record linkage. In: ACM SIGKDD, vol. 3, pp. 25–27. Citeseer (2003)

12. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. *Knowl. Data Eng. IEEE Trans.* **24**(9), 1537–1555 (2012)

13. Kolb, L., Thor, A., Rahm, E.: Dedoop: efficient deduplication with hadoop. *Proc. VLDB Endowment* **5**(12), 1878–1881 (2012)

14. Kolb, L., Rahm, E.: Parallel entity resolution with dedoop. *Datenbank-Spektrum* **13**(1), 23–32 (2013)

15. Kolb, L., Thor, A., Rahm, E.: Parallel sorted neighborhood blocking with mapreduce. *arXiv preprint [arXiv:1010.3053](https://arxiv.org/abs/1010.3053)* (2010)

16. Kolb, L., Thor, A., Rahm, E.: Load balancing for mapreduce-based entity resolution. In: *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pp. 618–629. IEEE (2012)

17. Kolb, L., Thor, A., Rahm, E.: Multi-pass sorted neighborhood blocking with mapreduce. *Comput. Sci. Res. Dev.* **27**(1), 45–63 (2012)

18. Mestre, D.G., Pires, C.E.S.: Improving load balancing for mapreduce-based entity matching. In: *2013 IEEE Symposium on Computers and Communications (ISCC)*, pp. 000618–000624. IEEE (2013)

19. Hsueh, S.C., Lin, M.Y., Chiu, Y.C.: A load-balanced mapreduce algorithm for blocking-based entity-resolution with multiple keys. In: *Proceedings of the Twelfth Australasian Symposium on Parallel and*

20. Dataset. <https://aminer.org/citation>, (Online)

Author information

Authors and Affiliations

Department of IT, SKIT, Jaipur, India

Dolly Mittal

Department of CSE, SKIT, Jaipur, India

Anjana Sangwan

Department of CSE, PCE, Jaipur, India

Veena Yadav

Corresponding author

Correspondence to [Dolly Mittal](#).

Editor information

Editors and Affiliations

Papua New Guinea University of Technology, Lae,

Papua New Guinea

Ashish Kumar Luhach

Namibia University of Science and Technology,

Windhoek, Namibia

Prof. Dharm Singh Jat

Universiti Malaysia Pahang, Pekan, Malaysia

Prof. Kamarul Bin Ghazali Hawari

University of Eastern Finland, Kuopio, Finland

Prof. Dr. Xiao-Zhi Gao

Saint Mary's University, Halifax, NS, Canada

Rights and permissions

[Reprints and Permissions](#)

Copyright information

© 2022 Springer Nature Switzerland AG

About this paper

Cite this paper

Mittal, D., Sangwan, A., Yadav, V. (2022). Efficient Entity Resolution for Bibliographic Data Using MapReduce. In: Luhach, A.K., Jat, D.S., Hawari, K.B.G., Gao, XZ., Lingras, P. (eds) Advanced Informatics for Computing Research. ICAICR 2021. Communications in Computer and Information Science, vol 1575. Springer, Cham. https://doi.org/10.1007/978-3-031-09469-9_17

[.RIS](#)  [.ENW](#)  [.BIB](#) 

DOI

https://doi.org/10.1007/978-3-031-09469-9_17

Published	Publisher Name	Print ISBN
25 June 2022	Springer, Cham	978-3-031-09468-2

Online ISBN	eBook Packages
978-3-031-09469-9	Computer Science
	Computer Science
	(R0)