On the Power of Massive Text Data

Jiawei Han
University of Illinois at Urbana-Champaign
hanj@illinois.edu

ABSTRACT

The real-world big data is largely unstructured, dynamic, and interconnected, in the form of natural language text. It is highly desirable to transform such massive unstructured data into structured knowledge. Many researchers and practitioners rely on labor-intensive labeling and curation to extract knowledge from unstructured text data. However, such approaches may not be scalable to web-scale or adaptable to new domains, especially considering that a lot of text corpora are highly dynamic and domain-specific. We argue that massive text data itself contains a large body of hidden patterns, structures, and knowledge. Equipped with domain-independent and domain-specific knowledge-bases, a promising direction is to develop more systematic data mining methods to turn massive unstructured text data into structured knowledge.

We introduce a set of methods developed recently in our own group on exploration of the power of big text data, including mining quality phrases using unsupervised, weakly supervised and distantly supervised approaches, recognition and typing of entities and relations by distant supervision, meta-pattern-based entity-attribute-value extraction, set expansion and local embedding-based multi-faceted taxonomy discovery, allocation of text documents into multi-dimensional text cubes, construction of heterogeneous information networks from text cube, and eventually mining multi-dimensional structured knowledge from massive text data. We show that massive text data itself can be powerful at disclosing patterns, structures and hidden knowledge, and it is promising to explore the power of massive, interrelated text data for transforming such unstructured data into structured knowledge.

KEYWORDS

Data mining, text mining, multi-dimensional text cube, heterogeneous information networks, data to knowledge

ACM Reference format:
Jiawei Han. 2018. On the Power of Massive Text Data. In *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining , Marina Del Rey, CA, USA, February 5–9, 2018 (WSDM 2018),* 1 pages.
https://doi.org/10.1145/3159652.3160604

Jiawei Han is the Abel Bliss Professor of Engineering in the Department of Computer Science, the University of Illinois at Urbana-Champaign. He has been researching into data mining, information network analysis, text mining, and database systems, with over 900 publications.

He served as the founding Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data (TKDD). He received ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), and IEEE Computer Society W. Wallace McDowell Award (2009). He is a Fellow of ACM and a Fellow of IEEE. His co-authored textbook "Data Mining: Concepts and Techniques" (Morgan Kaufmann) has been adopted worldwide. Professor Han is currently the co-Director of KnowEnG, a Center of Excellence in Big Data Computing, funded by NIH Big Data to Knowledge (BD2K) Initiative. He also served in 2009-2016 as the Director of Information Network Academic Research Center (INARC) supported by the Network Science-Collaborative Technology Alliance (NS-CTA) program of U.S. Army Research Lab.

ACKNOWLEDGMENTS

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The views and conclusions contained in this talk are those of the author and should not be interpreted as representing the policies of any government agencies.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02.
https://doi.org/10.1145/3159652.3160604