Construction of Structured Heterogeneous Networks from Massive Text Data

(Extended Abstract)

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

ABSTRACT

Network data analytics is important, powerful, and exciting. How big role may network data analytics play in the real world? Much real-world data is unstructured, in the form of natural language text. A grand challenges on big data research is to develop effective and scalable methods to turn such massive text data into actionable knowledge. In order to turn such massive unstructured, text-rich, but interconnected data into knowledge, we propose a data-to-network-to-knowledge (D2N2K) paradigm, that is, first transform data into relatively structured heterogeneous information networks, and then mine such text-rich and structure-rich heterogeneous networks to generate useful knowledge. We argue that such a paradigm represents a promising direction and network data analytics will play an essential role in transforming data to knowledge. However, a critical bottleneck in this game is mining structures from text data. We present our recent progress on developing effective methods for mining structures from massive text data and constructing structured heterogeneous information networks.

Keywords

Network mining; structure mining from massive text data

Introduction

In today's computerized and information-based society, we are soaked with vast amounts of text-based *unstructured* data, ranging from news articles, scientific publications, product reviews, to a wide range of textual information from social media. The success of network data analytics so far is largely attributed to the efficient and effective analysis of *structured* data/networks. Only if we can transform such unstructured but interconnected text data into structured heterogeneous networks, will network data analytics demonstrate its great power to impact the society.

NDA'17, May 19-19, 2017, Chicago, IL, USA

ACM ISBN 978-1-4503-4990-1/17/05.

DOI: http://dx.doi.org/10.1145/3068943.3068944

Fortunately, big data leads to big opportunities to uncovering structures of real-world entities (*e.g.*, `person`, `company`, `product`), attributes (*e.g.*, `age`, `weight`), relations (*e.g.*, `employee_of`, `manufacture`) from massive text corpora. By integrating the semantic-rich structures so generated with other inter-related structured data, powerful structured heterogeneous information networks can be constructed. This talk presents an organized picture of our recent research on information extraction and structure mining from massive text corpora, covering the following themes: (1) automated quality phrase mining: from `ToPMine` to `SegPhrase` and `AutoPhrase`; (2) entity/relation recognition and typing: from `ClusType` to `PLE`, `AFET` and `CoType`; (3) meta-pattern guided discovery of entities, attributes and their values: the `MetaPAD` approach; and (4) mining of text-generated structured heterogeneous networks. Research challenges are also identified, including integrated text mining and network mining for network construction.

About the speaker

Jiawei Han, Abel Bliss Professor, Department of Computer Science, Univ. of Illinois at Urbana-Champaign. His research areas encompass data mining, data warehousing, text mining, and information network analysis, with over 800 conference and journal publications. He is Fellow of ACM, Fellow of IEEE, and received 2004 ACM SIGKDD Innovations Award, 2005 IEEE Computer Society Technical Achievement Award, 2009 M. Wallace McDowell Award from IEEE Computer Society. His work has been suppoerted by U.S. National Science foundation, Network Science Collaborative Technology Alliance program of the U.S. Army Research Lab, and National Institute of Health.

Acknowledgement

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-1320617 and IIS 16-18481, 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 document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government.