“Mining Structures from Massive Text Data: A Cross Point of Data Mining, Machine Learning and Natural Language Processing”

Jiawei Han

The real-world big data are largely unstructured, interconnected, and dynamic, in the form of natural language text. It is highly desirable to transform such massive unstructured data into structured knowledge. Many researchers rely on labor-intensive labeling and curation to extract knowledge from such data. However, such approaches may not be scalable, especially considering that a lot of text corpora are highly dynamic and domain-specific. We argue that massive text data itself may disclose a large body of hidden patterns, structures, and knowledge. Equipped with domain-independent and domain-dependent knowledge-bases, we should explore the power of massive data itself for turning unstructured data into structured knowledge.

We introduce a set of methods developed recently in our own group on exploration the power of big text data, including mining quality phrases, recognition and typing of entities and relations by distant supervision, pattern-based information extraction, automated discovery of multi-faceted taxonomy, and construction of multidimensional text cubes. We show that it is critical to develop powerful and scalable methods by integrating the methodologies developed in Data Mining, Machine Learning and Natural Language Processing and demonstrate the promise of such a direction.