Bringing Structure to Text: Mining Phrases, Entities, Topics, and Hierarchies

Jiawei Han, Chi Wang, Ahmed El-Kishky
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
hanj@illinois.edu, chiwang1@illinois.edu, elkishk2@illinois.edu

ABSTRACT

Mining phrases, entity concepts, topics, and hierarchies from massive text corpus is an essential problem in the age of big data. Text data in electronic forms are ubiquitous, ranging from scientific articles to social networks, enterprise logs, news articles, social media and general web pages. It is highly desirable but challenging to bring structure to unstructured text data, uncover underlying hierarchies, relationships, patterns and trends, and gain knowledge from such data.

In this tutorial, we provide a comprehensive survey on the state-of-the art of data-driven methods that automatically mine phrases, extract and infer latent structures from text corpus, and construct multi-granularity topical groupings and hierarchies of the underlying themes. We study their principles, methodologies, algorithms and applications using several real datasets including research papers and news articles and demonstrate how these methods work and how the uncovered latent entity structures may help text understanding, knowledge discovery and management.

Why Bring Structure to Text?

In the big data age, vast amount of data generated in the world is unstructured or loosely structured, in the form of text, ranging from news to social media, business, government, and scientific documents, web pages, social networks, and enterprise logs. It is highly desirable to mine such huge amount of text data to discover its underlying thematic structures, hierarchies, and relationships.

Transforming vanilla unigrams into a richer information-rich phrases, uncovering real-world entities such as people, locations, and organizations, creation of heterogenous information networks, and constructing semantically rich conceptual and topical groupings of data can take intractable quantities of unstructured data and provide a rigid organization that can facilitate human exploration and understanding. This induced order provides application to information retrieval, information summarization, knowledge-base construction, and a wide-spectrum of new applications.

KDD'14, August 24–27, 2014, New York, NY, USA.
ACM 978-1-4503-2956-9/14/08.
http://dx.doi.org/10.1145/2623330.2630804.

Tutorial Outline

Phrase Mining and Phrase Topical Modeling
We discuss the limitations of unigrams and the 'bag-of-words' assumption to effective big-data analysis of text. We will provide examples from within the community to move forward from unigrams to multi-grams and demonstrate the value added from n-gram and 'bag-of-phrases' formulation of text data. We focus specifically on the problem of topic modeling and demonstrate the benefits of various phrase, n-gram, and bi-gram statistical topic models over unigram varieties.

Mining Topical Hierarchies: Topics at Multiple Granularities
Often times, flat topical structures fail to capture the subtleties in topics. We demonstrates efforts to more naturally model text corpora as topical hierarchies with a topic/subtopics schema. We argue that such a natural structure can allow for a more systematic drill-down approach to exploring corpora based on topical criterion.

Mining Latent Structures for Heterogenous Information Network Construction
We outline the benefits creating heterogenous information networks that embody the interrelation between the multi-typed entities found in text. We outline how mining latent structures from text can assist in constructing quality heterogenous information networks (HIN) and how these HIN allow for a principled and user-guided organization, exploration, and understanding large quantities of text.

Acknowledgments

The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193, U.S. National Science Foundation grants IIS-1017362, IIS-1320617, IIS-1354329, DTRA, MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. Ahmed El-Kishky is supported by a National Science Foundation Graduate Research Fellowship grant number DGE-1144245.