ABSTRACT

The “big data” era is characterized by an explosion of information in the form of digital data collections, ranging from scientific knowledge, to social media, news, and everyone’s daily life. Examples of such collections include scientific publications, enterprise logs, news articles, social media and general Web pages. Valuable knowledge about multi-typed entities is often hidden in the unstructured or loosely structured but interconnected data. Mining latent structured information around entities uncovers semantic structures from massive unstructured data and hence enables many high-impact applications.

In this tutorial, we summarize the closely related literature in database systems, data mining, Web, information extraction, information retrieval, and natural language processing, overview a spectrum of data-driven methods that extract and infer such latent structures, from an interdisciplinary point of view, and demonstrate how these structures support entity discovery and management, data understanding, and some new database applications. We present three categories of studies: mining conceptual, topical and relational structures. Moreover, we present case studies on real datasets, including research papers, news articles and social networks, and show how interesting and organized knowledge can be discovered by mining latent entity structures from these datasets.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGMOD’14, June 22–27, 2014, Snowbird, UT, USA.
Copyright 2014 ACM 978-1-4503-2376-5/14/06 ..$15.00.
http://dx.doi.org/10.1145/2588535.2588890.

Categories and Subject Descriptors

H.4.0 [Information Systems]: General

Keywords

Latent structure; Entity Knowledge Engineering

1. INTRODUCTION

The success of database technology is largely attributed to the efficient and effective management of structured data. The construction of a well-structured database is often the premise of subsequent applications. However, the explosion of “big data” poses great challenges on this practice since the real world data are largely unstructured. Thus, it is crucial to uncover latent structures of real-world entities, such as people, locations and organizations, and bring massive unstructured data into better semantic structures. By mining massive unstructured or loosely structured data where the entities occur, one can construct semantically rich structures which reveal the relationships among entities and provide conceptual or topical grouping of them. The uncovered entity relationships and groupings facilitate browsing information and retrieving knowledge from data that are otherwise hard to handle due to the lack of structures.

Example: Latent entity structures in scientific literature and news media. In a bibliographic database like DBLP and PubMed, research papers are explicitly linked with authors, venues and terms. Many interesting semantic relationships such as advisor-advisee between authors are hidden in the publication records; moreover, the research topics of authors, venues and terms are also hidden or unorganized, preventing insightful organization of the entities. In news articles, multiple types of entities like people, locations and organizations are involved in many events of different topics. The topics, as well as the entity concepts and relations are buried in the text rather than in the form of relational tuples. In the web scale, it is crucial to mine the entity structures hidden in web pages and tables since they are extensive resources to enrich open-domain knowledge-bases; also, mining the entity structures hidden in social media will help reorganize scattered information and improve the service for individuals.

This tutorial presents a comprehensive overview of the principles and methods developed for latent entity structure discovery in recent years. The following key issues will be covered: (i) mining latent entity structures for construction of structured databases; (ii) typing entities; (iii) topic modeling with entities; (iv) discovery of hidden relationships among entities; (v) case studies: research publications, news articles and social networks; and (vi) research frontiers on mining latent entity structures.

2. A PRELIMINARY OUTLINE OF THE TUTORIAL

1. Preliminaries on mining latent entity structures

http://www.informatik.uni-trier.de/

3. REFERENCES

IIS-1017362, IIS-1320617, IIS-1354329.

and U.S. National Science Foundation grants CNS-0931975, Cooperative Agreement No. W911NF-13-1-0193, DTRA, DHS, 2-0053 (NS-CTA) and the U.S. Army Research Office under Laboratory under Cooperative Agreement No. W911NF-09-.

The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA) and the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193, DTRA, DHS, and U.S. National Science Foundation grants CNS-0931975, IIS-1017362, IIS-1320617, IIS-1354329.

3. REFERENCES


