Mining Latent Entity Structures from Massive Unstructured and Interconnected Data

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

Chi Wang
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
chiwang1@illinois.edu

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 sematic 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.

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[1] and PubMed[2], 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 hihdden 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

[1] `http://www.informatik.uni-trier.de/`

[2] `http://www.ncbi.nlm.nih.gov/pubmed/`

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/10.1145/2588555.2588890.

   (a) Entity-related services as database applications

   (b) Entities in documentary data collections

   (c) Latent structures: concepts, topics and relations

2. Mining entity concepts for typing

   (a) Generation of taxonomies in large scale

      i. Taxonomies built from human-curated knowledge-bases: DBpedia, YAGO, WikiTaxonomy

      ii. Taxonomies constructed from free text: KnowItAll, TextRunner, Probase

   (b) Type entities in text: Wikifier, AIDA, MentionRank, MENED

   (c) Type entities in web tables and lists

3. Discovery of topical grouping of entities

   (a) PLSA and LDA: Bag-of-words topic modeling

   (b) CATHY and CATHYHIN: Phrase and entity based topical hierarchy construction

4. Discovery of hidden entity relations

   (a) Patterns, rules and constraints for reasoning

      i. Text and linkage patterns

      ii. Dependency rules and constraints

   (b) Representative mining paradigms

      i. Measuring similarities of relationships

      ii. Classifying or clustering entity relationships

      iii. Slot filling

   (c) Methodologies for dependency modeling

5. Case studies: Mining latent entity structures from research publications, news articles, web pages and online social networks

6. Recent research progress and open problems on mining latent entity structures

   (a) Mining relations and concepts from multiple sources

   (b) Integration of NLP and data mining approaches

Acknowledgments

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

[1] N. Bach and S. Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, 2007.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

[3] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Commun. ACM*, 51(12), Dec. 2008.

[4] J. Hoffart, M. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *EMNLP*, 2011.

[5] H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In *ACL*, 2011.

[6] H. Kim, X. Ren, Y. Sun, C. Wang, and J. Han. Semantic frame-based document representation for comparable corpora. In *ICDM*, 2013.

[7] R. Li, C. Wang, and K. Chang. User profiling in ego network: An attribute and relationship type co-profiling approach. In *WWW*, 2014.

[8] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. Mining evidences for named entity disambiguation. In *KDD*, 2013.

[9] G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and searching web tables using entities, types and relationships. *Proceedings of the VLDB Endowment*, 3(1-2):1338–1347, 2010.

[10] S. P. Ponzetto and M. Strube. Deriving a large scale taxonomy from wikipedia. In *AAAI*, 2007.

[11] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *ACL*, 2011.

[12] R. Socher, D. Chen, C. D. Manning, and A. Ng. Reasoning with neural tensor network s for knowledge base completion. In *NIPS*, 2013.

[13] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *WWW*, 2007.

[14] P. Venetis, A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering semantics of tables on the web. *Proceedings of the VLDB Endowment*, 4(9):528–538, 2011.

[15] C. Wang, K. Chakrabarti, T. Cheng, and S. Chaudhuri. Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *WWW*, 2012.

[16] C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. A phrase mining framework for recursive construction of a topical hierarchy. In *KDD*, 2013.

[17] C. Wang, M. Danilevsky, J. Liu, N. Desai, H. Ji, and J. Han. Constructing topical hierarchies in heterogeneous information networks. In *ICDM*, 2013.

[18] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *KDD*, 2010.

[19] C. Wang, J. Han, Q. Li, X. Li, W.-P. Lin, and H. Ji. Learning hierarchical relationships among partially ordered objects with heterogeneous attributes and links. In *SDM*, 2012.

[20] W. Wong, W. Liu, and M. Bennamoun. Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, 44(4):20, 2012.

[21] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *SIGMOD*, 2012.

[22] M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri. Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In *SIGMOD*, 2012.