Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies

— Proposal for a Tutorial at KDD 2014 Conference —

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

1 Introduction

Motivation: Why Bringing 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. If one can mine text data, to uncover latent structures of real-world entities, such as people, locations and organizations, and construct semantically rich structures that provide conceptual or topical grouping of them, one can bring order to big, unstructured text data, facilitate information retrieval, information summarization, knowledge-base construction, and a wide spectrum of new applications.

Real-world examples and typical data sets to be used in the tutorial:

It is always good to use many real-world good examples to illustrate concepts, principles and methods. To overview the principles and methods to be introduced in the tutorial, we will use multiple real-world text data sets, as illustrated below, as typical examples: (1) bibliographic data (essentially DBLP[1]), (2) news data (especially recent focused events in the news), and (3) some selected twitter datasets.

In bibliographic data, papers are explicitly linked with authors, venues and terms. However, the fundamental latent research topics of authors, venues, and papers are hidden, preventing insightful organization of the entities. When browsing scientific articles, individuals generally want to explore papers pertaining to a certain topic, yet the default organization by venue and author does not allow for this type of search. By organizing paper, authors, and terms into topics and subtopics, we can facilitate efficient search and exploration of relevant papers, yet the unavailability of topics for each of these entities makes it difficult to create this natural topic/subtopic hierarchy. We will show how one can benefit by bringing structures to the text-rich attributes, such as titles and abstracts.

For news data, we will show how meaningful phrases can be mined from such data, how categories/types can be extracted from news data, and how topical hierarchies can be generated from the analysis of such better structured news data.

For twitter data, we will show how phrase mining may help clustering and linking multiple lines of tweets

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

and facilitate multi-dimensional tweet summary and analysis.

What will be covered in this tutorial?

This tutorial presents a comprehensive overview of the techniques developed for topical structure discovery in recent years. We will discuss the following key issues.

- Phrase mining and phrase-based topic modeling
- Discovery of conceptual groupings of entities, that is, typing
- Discovery of hierarchical topics from text
- Case studies: Structure discovery on the collections of research publications, news articles, and social media.
- Research frontiers

Why a tutorial at KDD 2014?

In today's era of 'big data', people are exposed to an explosion of information in the form of text corpora and other document collections. Most of these collections are unstructured or loosely structured and can benefit from induced topical and conceptual organization. Automatically inferring and inducing conceptual groupings and mining high-quality topical structures at multiple granularity will facilitate many important tasks, such as search, information browsing, and knowledge discovery.

This tutorial will present an organized picture of recent research on phrase mining, entity-concept discovery, topical hierarchy construction and human-interpretable representations in text data. We will demonstrate how efficient and effective methods can be developed that may systematically bring structure to messy, unstructured data and show how exciting and interesting information can be discovered.

We believe such a comprehensive survey on new methods is in high demand since this is an emerging, exciting, and critically important theme. This tutorial will present a state-of-the-art review on phrase mining and entity-oriented structure discovery, and draw an organized picture in this important research frontier. It will also serve as a good learning experience for data mining and text mining researchers, information extraction researchers, industry developers, and graduate students.

Target audience and prerequisites:

Researchers and practitioners in the field of text mining, data mining, information extraction, information retrieval, web and information systems. While the audience with a good background in these areas would benefit most from this tutorial, we believe the material to be presented would give general audience and newcomers an introductory pointer to the current work and important research topics in this field, and inspire them to learn more. Only preliminary knowledge about text mining, information extraction, data mining, algorithms, and their applications are needed.

2 A preliminary outline of the tutorial

1. Mining structure from text: An introduction
   (a) Big text data: A major challenge in big data
   (b) Why bringing structure to text is critically important to text understanding, summary, and analysis?
   (c) Learning from the experience of database technology: Bringing structures will substantially enhance the power of search and mining
   (d) Our road map: from phrase mining and phrase-based topic modeling to entity concept (i.e., type) mining and hierarchical topical structure mining
2. Phrase Mining and Phrase-Based Topical Modeling
   (a) Unigram topic modeling
      i. Latent Semantic Analysis (LSA)
      ii. Probablistic Latent Semantic Analysis (PLSA)
      iii. Latent Dirichlet Allocation (LDA)
   (b) Topic visualization and limitation of uni-grams
      i. Visualizing topic models
      ii. How humans interpret topic models, word and topic intrusion
   (c) Phrase Mining Methods
      i. From bag-of-words to bag-of-phrases.
      ii. NLP and knowledge base tools for constructing bag-of-phrases representation
      iii. Mining "interesting" phrases from a corpus subset
      iv. Suffix tree phrase mining
      v. Significant N-grams based on mutual information
   (d) Phrase-based topic modeling
      i. TurboTopics: Combining topical uni-grams to phrases
      ii. KERT: A frequent pattern + ranking approach
      iii. Topical key phrase extraction and ranking from twitter
      iv. Probabilistic models for inferring phrase and topic: TNG and PD-LDA

v. Bag-of-phrases topic modeling: phrase mining and PhraseLDA

3. Discovery of Conceptual Grouping of Entities from Text Data

   (a) Large-scale taxonomies

   i. Built from human-curated knowledgebases, such as Wikipedia and Freebase: DBpedia, YAGO, WikiTaxonomy

   ii. Constructed from free text: KnowItAll, TextRunner, Probase

   iii. Constructed from web tables and lists: WebSets

   (b) Link entities in text to a taxonomy

   i. Wikifier: Link entity mentions in text to Wikipedia

   ii. AIDA: Collective mention-entity graph mapping

   iii. MENED: Mining evidence outside referent knowledgebases.

   iv. MentionRank: Find entities of a target concept

   (c) Link entities in web tables and lists to a taxonomy

4. Modeling topics and subtopics with topical hierarchies

   (a) Mining topical hierarchies

   i. Hierarchical topic models

   ii. CATHY: Topical hierarchy construction by clustering term co-occurences

   iii. Scalable Tensor Recursive Orthogonal Decomposition for Topical Hierarchy Construction

   (b) Mining topical hierarchies in heterogenous information networks

   i. RankClus and NetClus algorithms

   ii. CATHYHIN: Entity-enriched Topical Hierarchy Construction

5. Case studies: Research publications, news articles, and social media (Tweet analysis)

   (a) Phrase mining and phrase-based topical modeling in these datasets

   (b) Discovery of entity concept (type) structures in these datasets

   (c) Topical hierarchy construction in these datasets

6. Recent Research Progress and Research Frontiers

   (a) Mine entity relations and concepts from multiple sources

   (b) Combination of NLP, text modeling and data mining approaches

   (c) Deep learning: deep latent structures?

   (d) Construction of heterogeneous information networks from text data

   (e) Mining information networks with discovered structures

   (f) Information quality enhancement with structured text

   (g) Multi-dimensional summary and analysis over structured text

3 Tailoring the Tutorial to Different Durations

The duration of the tutorial is flexible: It is expected to be 3 hours, but it can be compressed into 1.5 hours, based on the need of the conference. The outline presented here is for the full length tutorial. For shorter duration of the tutorial, we plan to cut half of the text visualization section (section 3) and the entirety of conceptual grouping of entities (section 4).

4 About the Instructors

- **Jiawei Han**, Abel Bliss Professor, Department of Computer Science, Univ. of Illinois at Urbana-Champaign. His research areas encompass data mining, data ware-housing, information network analysis, and database systems, with over 600 conference and journal publications. He is Fellow of ACM and Fellow IEEE, received ACM SIGKDD Innovation Award (2004) and IEEE CS W. McDowell Award (2009). His co-authored textbook "Data Mining: Concepts and Techniques", 3rd ed., (Morgan Kaufmann, 2011) has been adopted popularly world-wide.

- **Chi Wang** is a Ph.D. candidate at Univ. of Illinois at Urbana-Champaign. He is the sole winner of Microsoft Research Graduate Research Fellowship in the history of CS, UIUC. He received KDDCUP'13 runner-up award for the entity-name disambiguation competition, and his work on topic hierarchy construction was nominated best paper candidates in CIKM'13 and ICDM'13.

- **Ahmed El-Kishky** is a Ph.D. candidate at Univ. of Illinois at Urbana-Champaign. He is a National Science Foundation Graduate Research Fellow and was twice selected to participate and conduct research at NSF REU sites at The University of Notre Dame and The University of Massachusetts Amherst.

5 A list of tutorials on the most related topics given by the same authors

The authors have presented over 20 tutorials in data mining and database conferences and workshops. The follow lists a set of related, recent tutorials given by the authors.

1. **Conference tutorial**: Jiawei Han and Chi Wang, "*Mining Latent Entity Structures from Massive Unstructured and Interconnected Data*", 2014 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'14), Snowbird, UT, June 2014.

2. **Conference tutorial:** Tim Weninger, and Jiawei Han, "Information Network Analysis and Extraction on the World Wide Web", 2013 Int. Conf. on the World Wide Web (WWW'13), Rio de Janeiro, Brazil, May 2013.

3. **Conference tutorial:** Tim Weninger and Jiawei Han, "Exploring Structure and Content on the Web: Extraction and Integration of the Semi-Structured Web", 2013 ACM Int. Conf. on Web Search and Data Mining (WSDM'13), Rome, Italy, Feb 2013.

4. **Conference tutorial:** Yizhou Sun, Jiawei Han, Xifeng Yan, and Philip S. Yu, "Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach", 2012 Int. Conf. on Very Large Data Bases (VLDB'12/PVLDB), Istanbul, Turkey, Aug 2012.

5. **Conference tutorial:** Jiawei Han, Yizhou Sun, Xifeng Yan, Philip S. Yu, "Mining Knowledge from Data: An Information Network Analysis Approach", 2012 IEEE Int. Conf. on Data Engineering (ICDE'12), Arlington, VA, Apr 2012.

This tutorial differentiates itself substantially from the above related tutorials by authors in technical contents. It contains no overlapped technical contents with our ICDE'12, VLDB'12, WSDM'13 and WWW'13 tutorials. It has some content overlap with our SIGMOD'14 tutorial. However, our SIGMOD'14 tutorial covers sections on entity-relationship discovery which will not be covered in this tutorial. This tutorial is focused on phrase mining and phrase topic modeling (which are not covered in SIGMOD'14), as well as share some technical contents with our SIGMOD'14 tutorial on entity concept mining and topic hierarchy mining.

References

[1] S. Bedathur, K. Berberich, J. Dittrich, N. Mamoulis, and G. Weikum. Interesting-phrase mining for ad-hoc text analytics. *Proceedings of the VLDB Endowment*, 3(1-2):1348–1357, 2010.

[2] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, volume 16, 2003.

[3] D. M. Blei and J. D. Lafferty. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*, 2009.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

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

[6] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

[7] A. J.-B. Chaney and D. M. Blei. Visualizing topic models. In *ICWSM*, 2012.

[8] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, volume 22, pages 288–296, 2009.

[9] B. B. Dalvi, W. W. Cohen, and J. Callan. Websets: Extracting sets of entities from the web using unsupervised information extraction. In *WSDM*, 2012.

[10] M. Danilevsky, C. Wang, N. Desai, J. Guo, and J. Han. Kert: Automatic extraction and ranking of topical keyphrases from content-representative document titles. *arXiv preprint arXiv:1306.0271*, 2013.

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

[12] 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.

[13] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

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

[15] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[16] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. 2006.

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

[18] 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.

[19] R. V. Lindsey, W. P. Headden III, and M. J. Stipicevic. A phrase-discovering topic model using hierarchical pitman-yor processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 214–222. Association for Computational Linguistics, 2012.

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

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

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

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

[24] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM, 2009.

[25] 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.

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

[27] 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.

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

[29] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 697–702. IEEE, 2007.

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

[31] M. Yetisgen-Yildiz and W. Pratt. The effect of feature representation on medline document classification. In *AMIA annual symposium proceedings*, volume 2005, page 849. American Medical Informatics Association, 2005.

[32] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 46–54. ACM, 1998.

[33] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 379–388. Association for Computational Linguistics, 2011.