Constructing Structured Information Networks from Massive Text Corpora

Xiang Ren, Meng Jiang, Jingbo Shang, Jiawei Han
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA
{xren7, mjiang89, shang7, hanj}@illinois.edu

ABSTRACT

In today's computerized and information-based society, text data is rich but messy. People are soaked with vast amounts of natural-language text data, ranging from news articles, social media post, advertisements, to a wide range of textual information from various domains (medical records, corporate reports). To turn such massive unstructured text data into *actionable knowledge*, one of the grand challenges is to gain an understanding of the factual information (*e.g.*, entities, attributes, relations, events) in the text. In this tutorial, we introduce data-driven methods to construct *structured information networks* (where nodes are different types of entities attached with attributes, and edges are different relations between entities) for text corpora of different kinds (especially for massive, domain-specific text corpora) to represent their factual information. We focus on methods that are minimally-supervised, domain-independent, and language-independent for fast network construction across various application domains (news, web, biomedical, reviews). We demonstrate on real datasets including news articles, scientific publications, tweets and reviews how these constructed networks aid in text analytics and knowledge discovery at a large scale.

Keywords

Quality Phrase Mining; Entity Recognition and Typing; Attribute Discovery; Massive Text Corpora; Relation Extraction

Introduction

Motivation: Constructing structured information networks from massive text corpora The success of data mining technology is largely attributed to the efficient and effective analysis of structured data. The construction of a well-structured, machine-actionable database from raw data sources is often the premise of consequent applications. Although the majority of existing data generated in our society is unstructured, big data leads to big opportunities to uncover structures of real-world entities (*e.g.*, `person`, `company`, `product`), attributes (*e.g.*, `age`, `weight`), relations (*e.g.*, `employee_of`, `produce`) from massive text corpora. By integrating these semantic-rich structures with other interrelated structured data (*e.g.*, social networks, transaction logs), one can construct a powerful structured information network as a conceptual summarization of the original text corpora and structured information. The uncovered structure information networks will facilitate browsing information and inferring knowledge that are otherwise locked in the text corpora. Computers can effectively conduct algorithmic analysis over these networks, and apply the knowledge to improve human productivity in various downstream tasks. Our phrase mining tool, SegPhrase [23], won the grand prize of Yelp Dataset Challenge[1] and was used by TripAdvisorin in their products[2]. Our entity recognition and typing tool, ClusType [33], was shipped as part of the products in Microsoft Bing and U.S. Army Research Lab.

Example: StructNet for social media. In a collection of tweets, entities of different types and relations between entities are mentioned in text. For example, from the tweet "*Jean Joho, Chef of Eiffel Tower Restaurant, is on board to present at EC 2010.*", it is desirable to identify "*Jean Joho*" as `person`, "*The Eiffel Tower Restaurant*" as `restaurant`, and the relation `chef_of`(*Jean Joho*, *The Eiffel Tower Restaurant*). However, domain text corpora pose significant new challenges to the existing systems: **(1)** the lack of annotated domain data presents a major challenge for adopting traditional information extraction systems. Fortunately, a number of structured and semantically rich knowledge-bases are available, which provides chances for "*automatic*" extraction with *distant supervision*; **(2)** many entity detection tools are trained on general-domain, grammatically clean text (*e.g.*, news articles), but cannot work well on text of other domains, genres or languages (*e.g.*, tweets). A domain-agnostic phrase mining algorithm is required to efficiently generate entity mention candidates with minimal assumptions on the language formation; **(3)** Even though the surface content provide clues on the types of entities and relations, natural language has extreme variability in expressing the same meaning, causing data sparsity issues when discovering "common text patterns". A principled methodology is needed to resolve synonymous patterns.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051107

[1] `http://www.yelp.com/dataset_challenge`
[2] `http://engineering.tripadvisor.com/mining-text-review-snippets/`

What will be covered in this tutorial?

This tutorial presents a comprehensive overview of the techniques developed for constructing structured information networks in recent years (a more detailed outline will be presented later). We will discuss the following key issues: (1) mining quality phrases from massive, unstructured text corpora; (2) entity recognition and typing: paradigms and methodologies; (3) relation extraction: existing systems, limitations, general domain vs. specific domains, and a joint entity and relation extraction approach; (4) attribute discovery in a massive, domain-specific text corpus: previous efforts, limitations, and a data-driven pattern mining approach; and (5) research frontiers.

Target Audience and Prerequisites

Researchers and practitioners in the field of web search, information retrieval, data mining, text mining, database 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, data mining, algorithms and their applications are needed.

Tutorial Outline

Preliminaries

We introduce the audience to the broad subject of structured information network construction by providing motivation in the context of *information extraction for cold-start knowledge base construction* . Within this context, we introduce entities, entity types, relations, and knowledge bases. We then introduce several different tasks to information extraction. In particular, we will introduce identifying typed entities from text, extracting different types of relations between entity mentions and mining attribute name and values for entities.

Quality Phrase Mining

We will first introduce the criteria of quality phrases and formulate the problem.

1. **Supervised Methods:** We start from automatic term recognition, which is the origin of phrase mining in the NLP community. We will introduce supervised noun phrase chunking techniques built upon annotated documents and other methods relied on more sophisticated NLP features such as dependency parser.
2. **Unsupervised Methods:** We follow on by introducing a set of unsupervised data-driven approaches taking the advantages of massive text corpora together with their broad appcalitions. Frequency statistics in the corpus are utilized to address both candidate generation and quality estimation. We then focus on applications of ToPMine for topical phrase mining and noun-collocation mining.
3. **Distantly / Weakly Supervised Methods:** We focus on a variety of methods including incorporating outside information via dictionary. We mainly emphasize SegPhrase, a weakly supervised approach for extracting high-quality phrases and entity mentions with minimal supervision. We also outline a distantly supervised language-independent phrase mining method built upon SegPhrase.

Entity Recognition and Typing

We start by giving formal problem definition of entity recognition and typing.

1. **Individual General-Domain Document:** In the context of general text recognition, we discuss introduce many named entity recognition (NER) methods. We discuss entity recognition as sequence labeling as well as the coarse types and manually-annotated corpora these models leverage.
2. **Individual Domain-Specific Document:** In the context of domain-specific extraction, we discuss several approaches. We discuss twitter in the context of Tweet segmentation and chunking as well as LabeledLDA based on Freebase. In addition to twitter we discuss entity recognition in product reviews and biomedical text data.
3. **Domain-Specific Text Corpora:** We contrast single-document cases to the context of large single-domain corpora. Starting with semi-supervised approaches, we present sequence-labeling models. We transition to weakly supervised approaches and their merits - discussing pattern-based bootstrapping methods, set expansion methods, and a variety of probabilistic modeling methods as well as graph-based label propagation approaches. We then discuss several approaches for distantly supervised entity recognition. These methods include state of the art approaches such as FIGER which performs sequence labeling with automatically annotated data , SemTagger - a contextual classifier that uses seed data , APOLLO which performs label propagation on graphs, and ClusType which employs relation phrase-based clustering for effective entity recognition. We further demonstrate how to extend the distant supervision framework to fine-grained typing scenarios, discuss the new challenges bring by noisy knowledge base labels, and introduce noise-robust approaches PLE and AFET.

Relation Extraction

1. **General-Domain Document:** In the context of text documents from general domain (*e.g.*, news), we introduce supervised relation extraction systems, which are trained on large amount of human-annotated data and rely on output of entity mention detectors. We discuss kernel-based supervised approaches, examine different features used in the systems, and introduce recent popular neural network models . We further introduce several systems which adopt sequence models to jointly extract entities and the relations between them .
2. **Domain-Specific Text Corpora:** In context of domain-specific text corpora, we focus on domain-independent methods that rely on minimal human-annotated data and minimal linguistic assumptions: weak supervision , distant supervision (with the help of public knowledge bases). We also introduce open information extraction systems which do not rely on any pre-defined schema for entities and relations.

Attribute Discovery

1. **Attribute Name Extraction:** We discuss two lines of work: One is using heterogeneous data sources such as structured data (*e.g.*, web tables), semi-structured data (*e.g.*, query logs) and unstructured data (*e.g.*, web documents) to extract the type's attribute names. Massive query logs can be turned into an ontology of factual knowledge. The other is developing advanced weakly-supervised learning models to train annotated data.
2. **Attribute Tuple Extraction:** We discuss two lines of work: One is the open information extraction (open IE) systems. The open IE systems produce a set of short sentence fragments and generate relation tuple extractions based on linguistic assumptions. The other is the slot-filling task that were conducted by learning annotated corpus to complete the attribute values when given the entity and attribute name. We will introduce a novel domain-independent methodology that jointly extracts the attribute names and tuples. The large collection of the tuples can enrich the database of real-world facts to support information search and decision-making.

Conclusion and Research Frontiers

We conclude our tutorial by demonstrating the capabilities of many of the tools and methods mentioned on a variety of test cases and metrics. We then present a few case-studies on two real-world datasets consisting of news articles and tweets. Finally, we discuss the downstream applications, related tasks and future directions for the subject of structured information network construction.

Related Tutorials

1. **Conference tutorial:** X. Ren, A. El-Kishky, H. Ji and J. Han, "Automatic Entity Recognition and Typing in Massive Text Data" (SIGMOD'16). `http://xren7.web.engr.illinois.edu/sigmod2016tutorial.html`.
2. **Conference tutorial:** X. Ren, A. El-Kishky, C. Wang and J. Han, "Automatic Entity Recognition and Typing in Massive Text Corpora" (WWW'16).
3. **Conference tutorial:** M. Jiang and J. Han, "Data-Driven Behavioral Analytics: Observations, Representations and Models" (CIKM'16). `http://www.meng-jiang.com/tutorial-cikm16.html`
4. **Conference tutorial:** J. Han, H. Ji and Y. Sun, "Successful Data Mining Methods for NLP" (ACL'15). `http://acl2015.org/tutorials-t1.html`.
5. **Conference tutorial:** X. Ren, A. El-Kishky, C. Wang and J. Han, "Automatic Entity Recognition and Typing from Massive Text Corpora: A Phrase and Network Mining Approach" (SIGKDD'15). `http://research.microsoft.com/en-us/people/chiw/kdd15tutorial.aspx`.
6. **Conference tutorial:** J. Han, C. Wang and A. El-Kishky, "Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies" (SIGKDD'14).

Tutorial Material and Equipment

We will provide attendees a website and upload our tutorial materials (outline, slides, references, software) there. There is no copyright issue. Standard equipment will be enough for our tutorial.

Instructors

• **Xiang Ren**, Ph.D. candidate, Department of Computer Science, Univ. of Illinois at Urbana-Champaign. His research focuses on creating computational tools for better understanding and exploring massive text data. He has published over 25 papers in major conferences. He received Google Global PhD Fellowship in Structured Data and Database Management in 2016, KDD Rising Star by Microsoft Academic Search in 2016, C. W. Gear Outstanding Graduate Student Award in 2016, and Yahoo!-DAIS Research Excellence Award in 2015. Mr. Ren has rich experiences in delivering tutorials in major conferences, including SIGKDD 2015, SIGMOD 2016 and WWW 2016.

• **Meng Jiang**, Postdoctoral Research Associate, Department of Computer Science, Univ. of Illinois at Urbana-Champaign. His research focuses on behavioral modeling and social media analysis. He got his Ph.D. of Computer Science from Tsinghua University, Beijing in 2015. His Ph.D. thesis won the Dissertation Award at Tsinghua. His recent research won the SIGKDD 2014 Best Paper Finalist. His ICDM 2015 Tutorial won the honorarium.

• **Jingbo Shang**, Ph.D. candidate, Department of Computer Science, Univ. of Illinois at Urbana-Champaign. His research focuses on mining and constructing structured knowledge from massive text corpora. He is the recipient of Computer Science Excellence Scholarship and Grand Prize of Yelp Dataset Challenge in 2015.

• **Jiawei Han**, Abel Bliss Professor, Department of Computer Science, Univ. of Illinois at Urbana-Champaign. His research areas encompass data mining, data warehousing, information network analysis, etc., with over 600 conference and journal publications. He is Fellow of ACM, Fellow of IEEE, the Director of IPAN, supported by Network Science Collaborative Technology Alliance program of the U.S. Army Research Lab, and the Director of KnowEnG: a Knowledge Engine for Genomics, one of the NIH supported Big Data to Knowledge (BD2K) Centers.

Acknowledgement

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1320617 and IIS 16-18481, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In *ACM conference on Digital libraries*, pages 85–94, 2000.

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

[3] K. Bellare, P. P. Talukdar, G. Kumaran, F. Pereira, M. Liberman, A. McCallum, and M. Dredze. Lightly-supervised attribute extraction. In *NIPS*, 2007.

[4] P. Deane. A nonparametric method for extraction of candidate phrasal terms. In *ACL*, 2005.

[5] X. L. Dong, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.

[6] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *VLDB*, 2015.

[7] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in knowitall:(preliminary results). In *WWW*, 2004.

[8] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[9] M. R. Gormley, M. Yu, and M. Dredze. Improved relation extraction with feature-rich compositional embedding models. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2015.

[10] R. Gupta, A. Halevy, X. Wang, S. E. Whang, and F. Wu. Biperpedia: An ontology for search applications. *PVLDB*, 7(7):505–516, 2014.

[11] S. Gupta and C. D. Manning. Improved pattern learning for bootstrapped entity extraction. In *CONLL*, 2014.

[12] Y. He and D. Xin. Seisa: set expansion by iterative similarity aggregation. In *WWW*, 2011.

[13] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*, 2011.

[14] R. Huang and E. Riloff. Inducing domain-specific semantic class taggers from (almost) nothing. In *ACL*, 2010.

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

[16] T. Koo, X. Carreras, and M. Collins. Simple semi-supervised dependency parsing. *ACL-HLT*, 2008.

[17] Z. Kozareva, K. Voevodski, and S.-H. Teng. Class label enhancement via related instances. In *EMNLP*, 2011.

[18] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *SIGIR*, 2012.

[19] Q. Li and H. Ji. Incremental joint extraction of entity mentions and relations. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL*, 2014.

[20] D. Lin and X. Wu. Phrase clustering for discriminative learning. In *ACL*, 2009.

[21] W. Lin, R. Yangarber, and R. Grishman. Bootstrapped learning of semantic classes from positive and negative examples. In *ICML Workshop on The Continuum from Labeled to Unlabeled Data*, 2003.

[22] X. Ling and D. S. Weld. Fine-grained entity recognition. In *AAAI*, 2012.

[23] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *SIGMOD*, 2015.

[24] R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. Non-projective dependency parsing using spanning tree algorithms. In *EMNLP*, 2005.

[25] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL*, 2009.

[26] R. J. Mooney and R. C. Bunescu. Subsequence kernels for relation extraction. In *NIPS*, 2005.

[27] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[28] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *CIKM*, 2000.

[29] A. Parameswaran, H. Garcia-Molina, and A. Rajaraman. Towards the web of concepts: Extracting concepts from large datasets. *VLDB*, 3((1-2)), September 2010.

[30] M. Paşca and B. Van Durme. What you seek is what you get: Extraction of class attributes from query logs. In *IJCAI*, pages 2832–2837, 2007.

[31] V. Punyakanok and D. Roth. The use of classifiers in sequential inference. In *NIPS*, 2001.

[32] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *ACL*, 2009.

[33] X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, and J. Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *SIGKDD*, 2015.

[34] X. Ren, W. He, M. Qu, L. Huang, H. Ji, and J. Han. Afet: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *EMNLP*, 2016.

[35] X. Ren, W. He, M. Qu, C. R. Voss, H. Ji, and J. Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *SIGKDD*, 2016.

[36] X. Ren, Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han. CoType: Joint extraction of typed entities and relations with knowledge bases. In *arXiv:1610.08763*, under review.

[37] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *EMNLP*, 2011.

[38] M. Schmitz, R. Bart, S. Soderland, O. Etzioni, et al. Open language learning for information extraction. In *EMNLP*, 2012.

[39] S. Takamatsu, I. Sato, and H. Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In *ACL*, 2012.

[40] P. P. Talukdar and F. Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *ACL*, 2010.

[41] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *ACL*, 2010.

[42] E. Xun, C. Huang, and M. Zhou. A unified statistical model for the identification of english basenp. In *ACL*, 2000.

[43] M. Yahya, S. Whang, R. Gupta, and A. Y. Halevy. Renoun: Fact extraction for nominal attributes. In *EMNLP*, pages 325–335, 2014.

[44] D. Yu and H. Ji. Unsupervised person slot filling based on graph mining. In *ACL*, 2016.

[45] G. Zhou, J. Su, J. Zhang, and M. Zhang. Exploring various knowledge in relation extraction. In *ACL*, 2005.