Building Structured Databases of Factual Knowledge from Massive Text Corpora

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

ABSTRACT

In today's computerized and information-based society, people are inundated with vast amounts of text data, ranging from news articles, social media post, scientific publications, to a wide range of textual information from various domains (corporate reports, advertisements, legal acts, medical reports). To turn such massive unstructured text data into structured, actionable knowledge, one of the grand challenges is to gain an understanding of the *factual information* (*e.g.*, entities, attributes, relations) in the text.

In this tutorial, we introduce data-driven methods on mining structured facts (*i.e.*, entities and their relations/attributes for types of interest) from massive text corpora, to construct structured databases of factual knowledge (called StructDBs). State-of-the-art information extraction systems have strong reliance on large amounts of task/corpus-specific labeled data (usually created by domain experts). In practice, the scale and efficiency of such a manual annotation process are rather limited, especially when dealing with text corpora of various kinds (domains, languages, genres). We focus on methods that are minimally-supervised, domain-independent, and language-independent for timely StructDB construction across various application domains (news, social media, biomedical, business), and demonstrate on real datasets how these StructDBs aid in data exploration and knowledge discovery.

Keywords

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

1. INTRODUCTION

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 (unstructured or loosely-structured) 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`, `manufacture`) from massive text corpora. By integrating these semantic-rich structures with other inter-related structured data (*e.g.*, product specification, user transaction log), one can construct a powerful StructDB as a conceptual abstraction of the original text corpora. The uncovered StructDBs will facilitate browsing information and inferring knowledge that are otherwise locked in the text corpora. Computational machines can effectively perform algorithmic analysis at a large scale over these StructDBs, and apply the new insights and knowledge to improve human productivity in various downstream tasks. Our phrase mining tool, SegPhrase [28], won the grand prize of Yelp Dataset Challenge[1] and was used by TripAdvisor in productions[2]. Our entity recognition and typing system, ClusType [38], was shipped as part of the products in Microsoft Bing and U.S. Army Research Lab.

Example: StructDB for social media posts.
In a collection of tweets, entities of different types and the relations between these 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, the extreme language variability of text corpora from various domains poses significant new challenges to the existing systems:

(1) The lack of annotated domain data presents a major challenge for adopting supervised information extraction methods (*e.g.*, deep learning methods). Fortunately, a number of structured and semantically rich knowledge-bases are publicly available, and has "information overlap" with the text corpus at hand. This provides chances for *automated* extraction with labeled data heuristically generated using information in knowledge bases (*i.e.*, distant supervision).

(2) Many entity detection tools are trained on general-domain, grammatically clean text (*e.g.*, news articles written in English), but cannot work well on text corpora of other domains, genres or languages (*e.g.*, web

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

SIGMOD'17, May 14-19, 2017, Chicago, IL, USA

DOI: http://dx.doi.org/10.1145/3035918.3054781

forum posts written in Arabic). 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 the vast amounts of synonymous text patterns found in the massive corpora.

What will be covered in this tutorial? This tutorial presents a comprehensive overview of the information extraction techniques developed in recent years, for constructing structured databases of factual knowledge (see also Section 2 for a more detailed outline). We will discuss the following key issues: (1) data-driven approaches for mining quality phrases from massive, unstructured text corpora; (2) entity recognition and typing: preliminaries, challenges, and methodologies; (3) relation extraction: previous efforts, limitations, recent progress, and a joint entity and relation extraction method using distant supervision; (4) attribute name and value discovery: previous efforts, limitations, and a data-driven pattern mining approach; (5) exploration and mining of constructed StructDBs; (6) case studies on several application domains; and (7) research frontiers.

Why a tutorial at SIGMOD? In today's era of "big data", people are exposed to an explosion of information in the form of textual data collections, ranging from the scientific knowledge of all humanity, to the daily life of individuals. Most of these collections are unstructured or loosely structured. Effective extraction and typing of factual information is key to inducing structure and understanding from messy and scattered raw data. This tutorial will present an organized picture of recent research on information extraction for structuring massive, unstructured text corpora. We will show how exciting and surprising knowledge can be discovered from your own not so well-structured raw corpora.

Target audience and prerequisites. Researchers and practitioners in the field of database systems, information extraction, data mining, text mining, web mining, information retrieval, 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, data mining, algorithms and their applications are needed.

2. OUTLINE

This tutorial presents a comprehensive overview of the information extraction techniques developed in recent years, for automated extraction of factual information from text data (especially from a large, domain-specific text corpora). We will discuss the following key issues.

1. Motivation and Background of StructDB construction
   - (a) StructDB: Basic concepts and components
   - (b) StructDB vs. general knowledge bases
     - i. KnowledgeVault
   - (c) Applications of StructDB
   - (d) Related efforts: Stanford DeepDive system
   - (e) StructDB construction: An overview
     - i. From phrases, to entities, relations and attributes
2. Phrase Mining from Massive Text Corpora
   - (a) Preliminaries
     - i. Criteria of Quality Phrases
       - A. Popularity: frequent enough;
       - B. Concordance / Independence: not by chance or not dependent on other words;
       - C. Informativeness: indicative of a specific topic or concept;
       - D. Completeness: a complete semantic unit.
     - ii. The Origin of Phrase Mining
       - A. Automatic Term Recognition
       - B. Supervised Noun Phrase Chunking
       - C. Dependency Parser-based Methods
   - (b) Data-Driven Phrase Mining in A Large Text Corpus
     - i. Unsupervised Frequency-based Methods
       - A. Zipf's Law-based Heuristic
       - B. Ratio-based Heuristic
       - C. Z-Score-based Heuristic: ToPMine
     - ii. Weakly Supervised Method: SegPhrase
     - iii. Automated Quality Phrase Mining
       - A. No Extra Human Effort
       - B. Support Multiple Languages
       - C. High Performance
3. Automated Entity Recognition and Typing
   - (a) Preliminaries
     - i. Entities that are explicitly typed and linked externally with documents.
       - A. Wikilinks and ClueWeb corpora
       - B. Probase: A Probabilistic Taxonomy
       - C. MENED: Mining evidence outside referent knowledge bases
     - ii. Entities that can be extracted within text.
     - iii. Traditional named entity recognition (NER) systems
       - A. Entity extraction as a sequence labeling task
       - B. Classic coarse types and manually-annotated corpora
       - C. Sequence labeling models
   - (b) Entity Recognition and Typing in A Large, Domain-specific Corpus
     - i. Semi-supervised approaches
       - A. Combining local and global features
     - ii. Weakly-supervised approaches
       - A. Pattern-based bootstrapping methods
       - B. SEISA: A set expansion method
       - C. Extracting entities from web tables
     - iii. Distantly-supervised approaches
       - A. SemTagger: Seed-based contextual classifier for entity typing
       - B. ClusType: Effective entity recognition by relation phrase-based clustering
     - iv. Fine-grained entity typing approaches

A. FIGER: Multi-label classification with automatically annotated data
B. Embedding methods for entity typing: AFET and WSABIE

v. Label noise reduction in distant supervision
A. Noisy type issue in distant supervision
B. Simple pruning heuristics
C. Partial-label learning methods
D. Label noise reduction by heterogeneous partial-label embedding

4. Automated Extraction of Structured Entity Relationships
(a) Preliminaries of relation extraction (RE)
i. Basic concepts: relation instance, relation mention
ii. Explicit relation vs. implicit relation
iii. Downstream applications
A. Knowledge base completion
B. Question answering systems
(b) Traditional supervised RE systems
i. Supervised RE methods
A. Supervised models
B. Features for relation extraction
C. Training data
D. Evaluation of RE task
ii. Systems from Stanford and IBM
(c) Extracting typed relations from A Massive Corpus
i. Weak supervision methods
A. Pattern-based boostrapping methods
B. Seed examples selection
C. DIPRE system
D. KnowItAll system
E. Snowball system
ii. Distant supervision (DS) methods
A. Distant supervision for RE: A typical workflow
B. Challenges of DS: noisy candidate labels
C. Noise-robust DS models
iii. Joint extraction of entities and relations
A. Supervised methods: linear programming and sequence models
B. CoType: A distantly-supervised method

5. Mining Attribute Names and Values
(a) Attribute name discovery
i. Unsupervised approaches
ii. Supervised approaches
iii. Google's approaches using query streams:
A. Biperpedia system
B. ARI system
(b) Attribute tuple extraction
i. Slot-filling approaches
A. Unsupervised methods
B. Supervised methods with web tables
C. Supervised methods with annotated corpus
ii. Open information extraction (IE) approaches
A. Open IE using linguistic structures
B. Open IE with web data: TextRunner, ReVerb, and Ollie
C. Stanford open IE system
(c) Attribute discovery from Massive Text Corpora
i. Meta pattern-driven method

6. Exploration and Mining of StructDB
(a) Keyphrase Extraction
i. LAKI: Latent keyphrase representation
(b) Automatic Summarization
i. CaseOLAP: Multi-dimensional Summarization
ii. Comparative Document Analysis

7. Case studies: news articles, tweets, biomedical papers.
(a) Constructing StructDBs from these datasets
i. Phrase mining results in these datasets
ii. Extracting entities for types of interest in these datasets
iii. Attribute discovery in these datasets

8. Recent advances and research problems

3. ABOUT THE 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 2016 (as the sole winner in the category of Structured Data and Database Management in the world), KDD Rising Star (Rank No.3) by Microsoft Academic Search 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. Homepage: http://xren7.web.engr.illinois.edu/.

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. Homepage: http://www.meng-jiang.com/.

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. Homepage: http://shang7.web.engr.illinois.edu/.

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. Homepage: http://web.engr.illinois.edu/~hanj/.

4. RELATED TUTORIALS

A list of tutorials on the most related topics given by the same authors are shown as followed:

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). http://web.engr.illinois.edu/~elkishk2/www2016/.
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 and C. Wang "Mining Latent Entity Structures from Massive Unstructured and Interconnected Data" (SIGMOD'14). http://web.engr.illinois.edu/~hanj/pdf/sigmod14_jhan.pdf.
7. **Conference tutorial:** Y. Sun, J. Han, X. Yan, and P. S. Yu, "Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach"(VLDB'12).

The above are the related tutorials given by the authors. Several of our early tutorials were on mining heterogeneous information networks. However, the power of such mining comes from structures of such networks (typed entities as nodes, and typed relations as links). Recent advances on information extraction and entity extraction from massive unstructured text make it possible to garner the informative and illuminating structures lying beneath the raw text corpora. This tutorial presents this new line of research on starting with the shallow information extraction (quality phrases, entities and relations) and ending up with deep database exploration and mining, by mining factual structures from text and constructing a structured database of facts for knowledge discovery.

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.

5. 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] B. Ahmadi, M. Hadjieleftheriou, T. Seidl, D. Srivastava, and S. Venkatasubramanian. Type-based categorization of relational attributes. In *EDBT*, pages 84–95, 2009.
[3] N. Bach and S. Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*.
[4] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI*, 2007.
[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] S. Brin. Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, 1998.
[7] R. C. Bunescu and R. Mooney. Learning to extract relations from the web using minimal supervision. In *ACL*, 2007.
[8] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. *VLDB*, 1(1):538–549, 2008.
[9] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, and T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *WSDM*, 2010.
[10] P. Deane. A nonparametric method for extraction of candidate phrasal terms. In *ACL*, 2005.
[11] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *VLDB*, 2015.
[12] 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 *Proceedings of the 13th international conference on World Wide Web*, pages 100–110. ACM, 2004.
[13] V. Ganti, A. C. König, and R. Vernica. Entity categorization over large document collections. In *SIGKDD*, 2008.
[14] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano. Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, 8(1):41–48, 2006.
[15] R. Gupta, A. Halevy, X. Wang, S. E. Whang, and F. Wu. Biperpedia: An ontology for search applications. *PVLDB*, 7(7):505–516, 2014.
[16] S. Gupta and C. D. Manning. Improved pattern learning for bootstrapped entity extraction. In *CONLL*, 2014.
[17] A. Halevy, N. Noy, S. Sarawagi, S. E. Whang, and X. Yu. Discovering structure in the universe of attribute names. In *WWW*, pages 939–949, 2016.
[18] Y. He and D. Xin. Seisa: set expansion by iterative similarity aggregation. In *WWW*, 2011.
[19] 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.
[20] R. Huang and E. Riloff. Inducing domain-specific semantic class taggers from (almost) nothing. In *ACL*, 2010.
[21] B. Kimelfeld. Database principles in information extraction. In *PODS*, 2014.
[22] T. Koo, X. Carreras, and M. Collins. Simple semi-supervised dependency parsing. *ACL-HLT*, 2008.
[23] T. Lee, Z. Wang, H. Wang, and S.-w. Hwang. Attribute extraction and scoring: A probabilistic approach. In *ICDE*, 2013.
[24] Q. Li and H. Ji. Incremental joint extraction of entity mentions and relations. In *ACL*, 2014.

[25] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1070–1078. ACM, 2013.

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

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

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

[29] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *ACL*, 2014.

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

[31] P. McNamee and J. Mayfield. Entity extraction without language-specific resources. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. Association for Computational Linguistics, 2002.

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

[33] N. Nguyen and R. Caruana. Classification with partial labels. In *KDD*, 2008.

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

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

[36] D. Qiu, L. Barbosa, X. L. Dong, Y. Shen, and D. Srivastava. Dexter: large-scale discovery and extraction of product specifications on the web. *Proceedings of the VLDB Endowment*, 8(13):2194–2205, 2015.

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

[38] 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 *KDD*, 2015.

[39] 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.

[40] 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 *KDD*, 2016.

[41] 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*, 2017.

[42] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE*, (99):1–20, 2014.

[43] J. Shin, S. Wu, F. Wang, C. De Sa, C. Zhang, and C. Ré. Incremental knowledge base construction using deepdive. *VLDB*, 8(11):1310–1321, 2015.

[44] A. Silva, W. Meira Jr, and M. J. Zaki. Mining attribute-structure correlated patterns in large attributed graphs. *PVLDB*, 5(5):466–477, 2012.

[45] Y. Sun and J. Han. Mining heterogeneous information networks: a structural analysis approach. *SIGKDD Explorations*, 14(2):20–28, 2013.

[46] J. Tang, M. Qu, and Q. Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174. ACM, 2015.

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

[48] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492. ACM, 2012.

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

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

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

[52] D. Yogatama, D. Gillick, and N. Lazic. Embedding methods for fine grained entity type classification. In *ACL*, 2015.

[53] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. R. Voss, and M. Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *COLING*, 2014.

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

[55] C. Zhang, J. Shin, C. Ré, M. Cafarella, and F. Niu. Extracting databases from dark data with deepdive. In *SIGMOD*, 2016.

[56] M. Zhang, M. Hadjieleftheriou, B. C. Ooi, C. M. Procopiuc, and D. Srivastava. Automatic discovery of attributes in relational databases. In *SIGMOD*, pages 109–120, 2011.

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

[58] L. Zou, R. Huang, H. Wang, J. X. Yu, W. He, and D. Zhao. Natural language question answering over rdf: A graph data driven approach. In *SIGMOD*, pages 313–324, 2014.