Knowledge Curation and Knowledge Fusion: Challenges, Models, and Applications

Xin Luna Dong
Google Inc.
Mountain View, CA
lunadong@google.com

Divesh Srivastava
AT&T Labs-Research
Bedminster, NJ
divesh@research.att.com

ABSTRACT

Large-scale knowledge repositories are becoming increasingly important as a foundation for enabling a wide variety of complex applications. In turn, building high-quality knowledge repositories critically depends on the technologies of knowledge curation and knowledge fusion, which share many similar goals with data integration, while facing even more challenges in extracting knowledge from both structured and unstructured data, across a large variety of domains, and in multiple languages.

Our tutorial highlights the similarities and differences between knowledge management and data integration, and has two goals. First, we introduce the Database community to the techniques proposed for the problems of entity linkage and relation extraction by the Knowledge Management, Natural Language Processing, and Machine Learning communities. Second, we give a detailed survey of the work done by these communities in knowledge fusion, which is critical to discover and clean errors present in sources and the many mistakes made in the process of knowledge extraction from sources. Our tutorial is example driven and hopes to build bridges between the Database community and other disciplines to advance research in this important area.

1. INTRODUCTION

Large-scale knowledge repositories have immense value for both humans and computers, and have been shown to be effective in facilitating web search and other complex tasks. Over the years we have seen many techniques proposed for (semi-) automatically building progressively larger repositories of knowledge, from Freebase [4] and YAGO [23, 16], to NELL [6], DeepDive [21], and Knowledge Vault [9] (see Table 1 for a comparison). Major companies including Google, Microsoft, Facebook, and Walmart have also launched their own efforts to organize knowledge [8].

Knowledge curation shares many similar goals with *data integration*: to integrate data of large diversity in representations of entities and of relations, to provide a unified interface to access and query the data, and to leverage the collective wisdom from a multitude of data sources. In turn, it faces many challenges that the data integration community has been facing for decades: identifying different mentions of the same real-world entity, matching different ways of representing the same attribute of an entity or the same relation between entities, discovering erroneous and out-of-date data, and so on. Web-scale knowledge discovery faces even more challenges, as the desire is to extract knowledge from both structured data and unstructured data on the web, across multiple domains and across many languages.

The goal of this tutorial is two-fold. First, we aim to introduce to the Database community the techniques proposed by other communities such as Knowledge Management, Natural Language Processing, and Machine Learning, in resolving the heterogeneity inherent in web-scale data, towards building a single coherent repository of the knowledge in the world. We describe critical techniques in this process, focusing on entity linkage and relation extraction. We point out the similarities and differences of the techniques with the data extraction, schema alignment, and entity resolution techniques that have been well studied in the Database community.

Second, we give a detailed survey of *knowledge fusion*, an important tool to discover and clean both errors present in data sources, and the many mistakes that can be made in the process of knowledge extraction from sources. Comparing with *data fusion*, which aims at resolving conflicts from sources [3, 12], knowledge fusion considers an additional dimension of errors–the errors made by knowledge extractors. We present how existing data fusion techniques can be adapted to solve the knowledge fusion problem; in addition, we present the knowledge fusion techniques from the Machine Learning community based on AdaBoost learning [9], random walk inference [18], and deep learning [7].

Comparing with previous tutorials that have covered data fusion [12, 13], this tutorial has only a very small portion introducing data fusion techniques, mainly for the purpose of comparison to highlight the new challenges faced by knowledge fusion. Comparing with previous tutorials on knowledge management [2, 24], this tutorial focuses more on the post-processing step of knowledge fusion, and on drawing parallels between the research that has been done by the Database community on data integration, data cleaning, and data management, and the research outside our community on knowledge extraction and curation. We hope our tutorial can build bridges across these communities and inspire more

SIGMOD'15, May 31–June 4, 2015, Melbourne, Victoria, Australia.

http://dx.doi.org/10.1145/2723372.2731083.

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
Knowledge Vault (KV)	1100	45M	4469	271M
DeepDive [21]	4	2.7M	34	7M[a]
NELL [6]	271	5.19M	306	0.435M[b]
PROSPERA [20]	11	N/A	14	0.1M
YAGO2 [16]	350,000	9.8M	100	4M[c]
Freebase [4]	1,500	40M	35,000	637M[d]
Knowledge Graph (KG)	1,500	570M	35,000	18,000M[e]

Table 1: Comparison of knowledge bases [9]. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Confident facts means with a probability of being true at or above 0.9.

[a]Ce Zhang (U Wisconsin), private communication.
[b]Bryan Kiesel (CMU), private communication.
[c]Core facts, `http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html`
[d]This is the number of non-redundant base triples, excluding reverse predicates and "lazy" triples derived from flattening CVTs (complex value types).
[e]`http://insidesearch.blogspot.com/2012/12/get-smarter-answers-from-knowledge_4.html`

inter-disciplinary research to leverage our expertise on data management for improving knowledge management.

2. TARGET AUDIENCE

The target audience for this tutorial is anyone with an interest in understanding knowledge management. The assumed level of mathematical sophistication is that of the typical conference attendees. Apart from a basic understanding of database technology, there is no prerequisite for this tutorial.

3. TUTORIAL OUTLINE

Knowledge curation and fusion are broad topics. Our tutorial is example driven and organized as follows.

3.1 Motivation

Our tutorial starts with a variety of examples of existing knowledge bases built by the academia and by industry (see Table 1). We motivate the importance of knowledge management by a few real-world examples such as web search.

3.2 Knowledge Extraction

We then briefly introduce the key techniques in knowledge extraction, namely, entity linkage and relation extraction.

Entity linkage is analogous to entity resolution (or record linkage) for structured data (surveyed in [15]). Whereas entity resolution focuses on structured records that refer to the real world, entity linkage considers entity mentions in text and semi-structured data (*e.g.*, DOM trees). There have been over 100 papers on this topic [17]; we focus on the techniques presented in the seminal paper [22].

Relation extraction is analogous to data extraction and schema alignment in data integration. We briefly review techniques for extracting relations from free text, DOM trees, web tables, and web annotations. In particular, we focus on two techniques: *distance supervision* based on existing schemas and training data [19], and *open IE* without requiring any human input [1, 14].

We end this session with a brief analysis of the knowledge extracted by 16 extractors from the web in the Knowledge Vault project [9].

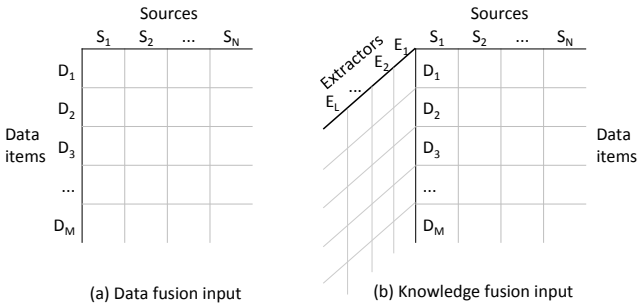

Figure 1: Input for data fusion is two-dimensional whereas input for knowledge fusion is three-dimensional [10].

3.3 Knowledge Fusion Models

The background of knowledge extraction serves as a good motivation for knowledge fusion, where we need to address both erroneous data present in web sources and mistakes in knowledge extraction. In characterizing knowledge fusion, we highlight the key difference from data fusion depicted in Figure 1; that is, data fusion assumes we know exactly the data provided by each source so the input is two-dimensional, whereas knowledge fusion reasons about the data extracted from the sources by multiple extractors so the input is three-dimensional.

We then present three types of knowledge fusion techniques. The first type takes signals such as the number of extractors that extract a triple and the extraction confidence, and learns a binary classifier for the correctness of an extracted triple. We present the techniques and the training data that are used in such supervised learning [9].

The second type learns graph-based priors to verify extracted knowledge. As an example, knowing that X and Y are both the parents of Z would increase the confidence that X and Y have a spouse relationship. This problem can be considered as link prediction–predicting edges from existing edges in a graph. We present two techniques: path ranking algorithm (PRA) [18] and neural network model (MLP) [7].

The third type extends data fusion techniques, which leverage agreements between sources and give higher trust to

high-quality sources. We show how basic data fusion techniques have been extended to take care of the mistakes made during knowledge extraction [10].

For each type of techniques, we briefly show experimental results on Knowledge Vault data for validation.

3.4 Knowledge Fusion Applications

We then briefly discuss potential applications for knowledge fusion. In particular, we highlight two applications. The first one generates a new signal for evaluating the quality of web sources: *knowledge-based trust* [11]. We discuss the techniques that are critical to generate such measures and our observations on how it complements existing measures such as *PageRank* [5]. The second one, called *Data X-Ray*, is a diagnosis framework that analyzes the common features among the identified wrong knowledge triples to provide insights on possible underlying systematic errors made in the process of knowledge curation [25].

3.5 Open Problems

We end our tutorial with a discussion of open problems in knowledge curation and knowledge fusion. It includes improving entity linkage on DOM-tree data by exploring the structure of the data, combining corpus-based matching and distance supervision for better relation extraction, advancing knowledge fusion for openIE, and so on. In particular, we issue a call to arms–*no valuable data left behind*, where we hope to significantly enrich knowledge bases by exploiting the large volume of "tail" data, including data about less popular entities, in less popular verticals, about non-current (historical) facts, from smaller sources, in languages other than English, and so on. How to effectively combine the techniques designed for structured data and those oriented towards unstructured data is the key for enabling greater success in this area.

4. CONCLUSIONS

This tutorial surveys the state-of-the-art techniques for knowledge curation and knowledge fusion. Our tutorial explores the similarities and differences between the techniques proposed for data integration and for knowledge management, aiming to build bridges between the research in the Database community and in other disciplines to develop more comprehensive techniques in this active research area.

5. BIOGRAPHIES

Xin Luna Dong is a Senior Research Scientist at Google Inc. She is one of the major contributors for the Knowledge Vault project and has led the Solomon data fusion project. She is a co-chair for WAIM'15 and has served as an area chair for Sigmod'15, ICDE'13, and CIKM'11.

Divesh Srivastava is the head of the Database Research Department at AT&T Labs-Research. He is an ACM fellow, on the board of trustees of the VLDB Endowment, the managing editor of PVLDB, and an associate editor of the ACM Transactions on Database Systems. He has served as PC co-chair of many conferences, including VLDB'07.

6. REFERENCES

[1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI*, 2007.

[2] D. Barbosa, H. Wang, and C. Yu. Shallow information extraction for the knowledge web. In *ICDE*, 2013.

[3] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1–41, 2008.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250, 2008.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. H. Jr., and T. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[7] J. Dean. Large scale deep learning. In *CIKM*, 2014.

[8] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. Building, maintaining, and using knowledge bases: A report from the trenches. In *Sigmod*, 2013.

[9] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.

[10] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. *PVLDB*, 2014.

[11] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. http://arxiv.org/abs/1502.03519, 2015.

[12] X. L. Dong and F. Naumann. Data fusion–resolving data conflicts for integration. *PVLDB*, 2009.

[13] X. L. Dong and D. Srivastava. Big data integration. *PVLDB*, 2013.

[14] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. Open information extraction: the second generation. In *IJCAI*, 2011.

[15] L. Getoor and A. Machanavajjhala. Entity resolution: Theory, practice, & open challenges. *PVLDB*, 5(12):2018–2019, 2012.

[16] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 2012.

[17] H. Ji. Entity linking and wikification reading list. http://nlp.cs.rpi.edu/kbp/2014/elreading.html, 2014.

[18] N. Lao, T. Mitchell, and W. Cohen. Random walk inference and learning in a large scale knowledge base. In *EMNLP*, 2011.

[19] M. Mintz, S. Bills, R. Snow, and D. Jurafksy. Distant supervision for relation extraction without labeled data. In *Prof. Conf. Recent Advances in NLP*, 2009.

[20] N. Nakashole, M. Theobald, and G. Weikum. Scalable knowledge harvesting with high precision and high recall. In *WSDM*, pages 227–236, 2011.

[21] F. Niu, C. Zhang, and C. Re. Elementary: Large-scale Knowledge-base Construction via Machine Learning and Statistical Inference. *Intl. J. on Semantic Web and Information Systems*, 2012.

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

[23] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO - A Core of Semantic Knowledge. In *WWW*, 2007.

[24] F. M. Suchanek and G. Weikum. Knowledge bases in the age of big data analytics. In *VLDB*, 2014.

[25] X. Wang, X. L. Dong, and A. Meliou. Data X-Ray: A diagnostic tool for data errors. In *Sigmod*, 2015.