Knowledge Graphs: In Theory and Practice

Nitish Aggarwal
IBM Research
1 PUBLICATION 0 CITATIONS
SEE PROFILE

Saeedeh Shekarpour
University of Bonn
42 PUBLICATIONS 328 CITATIONS
SEE PROFILE

Amit Sheth
Wright State University
963 PUBLICATIONS 26,412 CITATIONS
SEE PROFILE

Knowledge Graphs: In Theory and Practice

Nitish Aggarwal
IBM Watson
San Jose, USA
nitish.aggarwal@ibm.com

Saeedeh Shekarpour
Knoesis Research Centre
Ohio, USA
saeedeh@knoesis.org

Sumit Bhatia
IBM Research
New Delhi, India
sumitbhatia@in.ibm.com

Amit Sheth
Knoesis Research Centre
Ohio, USA
amith@knoesis.org

ABSTRACT

Through the proposed tutorial, we aim to cover the state-of-the-art approaches in Knowledge Graph Construction from various types of data (i.e. unstructured, semi structured and structured data) and using both manual as well as automated methods. We also wish to review applications from various disciplines that benefit from the structure and semantics offered by knowledge graphs. Lastly, we will present case studies describing our experiences in construction of IBM Watson's Knowledge Graph and its applications in life sciences and intelligence domains.

KEYWORDS

Enterprise Knowledge Management, Knowledge Graph, IBM Watsonn

ACM Reference format:
Nitish Aggarwal, Saeedeh Shekarpour, Sumit Bhatia, and Amit Sheth. 2017. Knowledge Graphs: In Theory and Practice. In *Proceedings of CIKM 2017, Singapore, October 2017 (CIKM'17),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

1.1 Why This Tutorial?

We are transitioning from the era of Big Data to Big Knowledge, and semantic knowledge bases such as knowledge graphs play an important role in this transition. This is evident from the increased investments in Knowledge Graph research and development by major industrial players resulting in widely used systems such as IBM's Watson, Google's entity search, Apple's Siri, and Amazon's product graph [9, 18].

Knowledge Graphs can be constructed either manually (facts authored by humans) or automatically (facts extracted from text using Machine Learning tools). Manually curated knowledge graphs such as DBpedia[11][1], YAGO[2], have little or no noisy facts as they are carefully authored, but they require very large human efforts. This problem is further exacerbated in enterprise domains and custom domains such as life sciences, finance, intelligence, where domain expertise is also crucial to add good quality facts in the graph. As a result, efforts have been made for development of systems for automatic construction of semantic knowledge bases for domain specific corpora and systems that use such domain specific knowledge bases [13] are gaining prominence.

Through this tutorial, we propose to review the state-of-the-art in Knowledge Graph construction and curation, as well as share the problems encountered in building a Knowledge Graph from scratch in enterprise settings. We will share our learning and experiences through a *case study describing the practical problems encountered* while working on IBM Watson's Knowledge Graph and solutions developed using the Graph for clients from intelligence and life sciences domain. Further, as the relevant literature is widely dispersed across different communities like Web Mining (WSDM, WWW, and KDD), Artificial Intelligence (IJCAI and AAAI), Natural Language Processing (ACL and EMNLP), Semantic Web (ISWC), and Data Management (SIGMOD, VLDB, and ICDE), the tutorial also serves as a guided tour on the latest research in these venues and aims to offer a unifying big picture.

Specifically, our objectives for this tutorial are as follows.

(1) To review different approaches commonly used for constructing knowledge graphs. These approaches range from curated methods performed mainly for (semi) structured data [8] to the automated techniques that rely on substantial machine learning methods, and are often used for unstructured data [12].
(2) To provide an overview of applications and advantages of using knowledge graphs in various disciplines. We will discuss how the structure and semantics offered by Knowledge Graphs can be exploited to better solve a variety of problems in various disciplines including question answering, natural language processing, information retrieval, machine learning and information extraction.
(3) To share our experiences while constructing the Watson Knowledge Graph for building industrial applications, and provide an overview of practical problems encountered building such applications and possible solutions to these problems.

[1] http://wiki.dbpedia.org/
[2] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

CIKM'17, Singapore

DOI: 10.1145/nnnnnnn.nnnnnnn

1.2 Related Events

- In AAAI 2017 which is a top-tier conference on artificial intelligence, Jay Pujara, Sameer Singh, and Bhavana Dalvi presented a tutorial on fiKnowledge Graph Construction from Textfi[3]. They presented an overview of the existing approaches for extracting candidate facts from text and incorporating these into a well-formed knowledge graph. This tutorial was structured in the following direction (i) knowledge extraction, with a focus on the underlying NLP tasks and successful approaches for converting text into candidate facts. (ii) machine learning approaches such as tensor factorization, deep learning, probabilistic graphical models, and random walk strategies for integrating candidate facts into a complete and coherent knowledge graph.
- In WSDM 2017 which is also a top-tier conference on Web search and data mining, Laura Dietz, Alexander Kotov, Edgar Meij organized a tutorial about fiUtilizing Knowledge Graphs in Text-centric Information Retrievalfi[4]. This tutorial presents the progress in this emerging area to industry practitioners and researchers with the focus on entity linking and entity retrieval methods.

While these tutorials provided an in-depth review of the state-of-the-art in Knowledge Graph construction and applications in information retrieval, they lacked the coverage of practical problems encountered while deploying Knowledge Graph based systems in real life settings. Through the proposed tutorial, we wish to to share our experiences of working on Watson Knowledge Graph and solutions developed to solve such problems.

2 DETAILED DESCRIPTION

2.1 Content Overview

Our tutorial will consist of three sessions:

1. **Introduction to knowledge graph**
   (a) Introduction which provides the required background and motivation.
   (b) Knowledge graph construction approaches:
      - Curated approaches e.g, DBpedia (Human in loop)
      - Automatic approaches mainly relying on fact extraction methods (e.g., OpenIE, Relation extraction). These approaches might employ human for validating extracted fact (e.g., Google KG) or exclusively rely on automatic validation methods (e.g., Watson KG)
2. **Knowledge graph for text analytics.** Majority of AI-based applications dealing with text require to recognize entities with respect to the underlying background knowledge. The recognized entities provide more abstract semantics rather than purely statistical analysis of text. In this part of tutorial, we discuss the traditional approaches versus the knowledge graph-empowered approaches in various tasks and disciplines:
   (a) Traditional problems
      (i) Entity Recognition (NER) and Disambiguation [5, 10]
      (ii) Entity Search
      (iii) Entity Recommendation [1, 5, 6]
   (b) Emerging problems
      (i) NLP tasks: Entity discovery and consolidation. Relation discovery and ranking [3].
      (ii) Question answering tasks: Federated approaches on distributed knowledge graphs, semantic parsing, path discovery and ranking [4, 7, 14, 17].
      (iii) Machine learning: features extracted from semantics as well as structure of knowledge graphs [2, 15, 16].
3. **Case Study:** Watson Knowledge Graph for Life Sciences and Intelligence

2.2 Aims and Learning Objectives

Our learning objectives are the following:

- Provide an overview of the state-of-the-art of existing knowledge graphs.
- Discuss challenges occurring in the entire life-cycle of knowledge graphs i.e. creation, maintenance, and updating.
- Elaborate on differences, between building applications using curated knowledge like DBpedia and automatically constructed knowledge like Waston Knowledge Graph.
- Walk through advantages of employing semantics as well as structure of knowledge graphs for enhancing cognitive capabilities of various applications and discipline.
- Guided tour over research challenges and open problems in building industrial applications using knowledge graph.

2.3 Material, Presentation Style and Format

Our presentation relies on slides prepared by embedding high quality description of the subjects, challenges and examples tuned for the running subject. We are strongly welcome to audience questions or challenges during our presentation ensuring the audience walk through the learning process. We will organize three sessions lasting 45 minutes each.

2.4 Required Prior Knowledge

There are no specific prerequisites for attending the tutorial.

3 AUDIENCE

We believe that the tutorial will be of interest to researchers from both industry and academia with diverse backgrounds spanning semantic web, information retrieval, text mining, machine learning. The tutorial will specifically be beneficial for people that are interested in or are currently involved in building systems and applications based on domain specific knowledge graphs.

4 LENGTH

This will be a half day tutorial.

[3] http://www.aaai.org/Conferences/AAAI/2017/aaai17tutorials.php#SUP2
[4] http://www.wsdm-conference.org/2017/tutorials/

5 TECHNICAL REQUIREMENTS

We will only need standard projection equipment.

6 PRESENTERS

Dr. Nitish Aggarwal
IBM Watson, USA,
nitish.aggarwal@ibm.com
Nitish Aggarwal is a Research Scientist in Watson Knowledge Graph Department at IBM Watson, Almaden Research Centre, USA, where he is leading the research effort in building intelligent industrial applications using knowledge graph. He received his PhD from Insight Centre for Data Analytics, National University of Ireland. He works on the intersection of natural language processing, Information Retrieval and semantic web technologies. Nitish has contributed to several European funded projects in the area of knowledge graphs construction, mining and retrieval. He was the organizing chair of Proactive Information Retrieval workshop, collocated with ECIR 2016, and has served in program committee of multiple conferences and journals including ISWC, ESWC, AAAI, ACL, WWW, JASIST, IP&M, SWJ and JWS.

Dr. Sumit Bhatia
IBM Research India
Sumit Bhatia is a Research Scientist in Knowledge Engineering Department at IBM India Research Laboratory where he is working on developing a shared knowledge infrastructure for different client engagements. Previously, as a Researcher in IBM Watson, he led the development of cognitive analytic algorithms build on top of Watson's Knowledge Graph. He was a Post-doctoral Researcher at Xerox PARC and as a part of CiteSeerX project at Penn State, Sumit developed a search engine that searches for algorithms and pseudo-codes in academic documents. Sumit's primary research interests are in the fields of Knowledge Management, Information Retrieval and Text Analytics, and he has published 25+ papers in top journals and conferences. He was the organizing chair of Proactive Information Retrieval workshop, collocated with ECIR 2016 and Social Multimedia Data Mining Workshop, collocated with ICDM 2014. He has served as a reviewer for multiple conferences and journals including WWW, CIKM, ACL, TKDE, TOIS, WebDB, JASIST, IJCAI, and AAAI.

Dr. Saeedeh Shekarpour
Kno.e.sis Research Center, USA,
saeedeh@knoesis.org
Homepage: http://knoesis.org/researchers/saeedeh/
Saeedeh Shekarpour accomplished her PhD research in Germany at the University of Bonn. She spent one year as a postdoctoral researcher in the EIS research group at the Bonn University and 1+ year as a postdoctoral researcher at Knoesis research center. Her research interests are question answering, Semantic Web, NLP, statistical classifiers and social network mining. She successfully published her research results in the top-tier and prestigious conferences and journals of her field including WWW, AAAI, Web Intelligence conference, IEEE Confs, Journal of Web Semantics, Semantic Web Journal.

Prof. Dr. Amit Sheth
Kno.e.sis Research Center, USA,
amit@knoesis.org
Homepage: http://knoesis.wright.edu/amit/research/
Prof. Dr. Amit Sheth is a computer scientist at Wright State University in Dayton, Ohio. His work has been cited by 36,355+ publications with an h-index of 94 which puts him among the top 100 computer scientists with the highest h-index.

REFERENCES

[1] Nitish Aggarwal, Kartik Asooja, Housam Ziad, and Paul Buitelaar. 2015. Who are the american vegans related to brad pitt?: Exploring related entities. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 151–154.

[2] Nitish Aggarwal and Ken Barker. 2015. Medical Concept Resolution.. In *Proceedings of the International Semantic Web Conference*.

[3] Nitish Aggarwal, Sumit Bhatia, and Vinith Misra. 2016. Connecting the Dots: Explaining Relationships Between Unconnected Entities in a Knowledge Graph. In *The Semantic Web - ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*. 35–39. https://doi.org/10.1007/978-3-319-47602-5_8

[4] Nitish Aggarwal and Paul Buitelaar. 2012. A system description of natural language query over dbpedia. In *Proceedings of Interacting with Linked Data (ILD 2012)*.

[5] Nitish Aggarwal and Paul Buitelaar. 2014. Wikipedia-based Distributional Semantics for Entity Relatedness. In *2014 AAAI Fall Symposium Series*.

[6] Nitish Aggarwal, Peter Mika, Roi Blanco, and Paul Buitelaar. 2015. Insights into Entity Recommendation in Web Search.. In *Proceedings of IESD@ISWC*.

[7] Nitish Aggarwal, Tamara Polajnar, and Paul Buitelaar. 2013. Cross-lingual natural language querying over the web of data. In *International Conference on Application of Natural Language to Information Systems*. Springer, 152–163.

[8] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[9] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 601–610.

[10] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 782–792.

[11] Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics* 7, 3 (2009), 154–165.

[12] Tom Mitchell. 2010. *Never-ending learning*. Technical Report. DTIC Document.

[13] Meenakshi Nagarajan, Angela D. Wilkins, Benjamin J. Bachman, Ilya B. Novikov, Shenghua Bao, Peter J. Haas, María E. Terrón-Díaz, Sumit Bhatia, Anbu K. Adikesavan, Jacques J. Labrie, Sam Regenbogen, Christie M. Buchovecky, Curtis R. Pickering, Linda Kato, Andreas M. Lisewski, Ana Lelescu, Houyin Zhang, Stephen Boyer, Griff Weber, Ying Chen, Lawrence Donehower, Scott Spangler, and Olivier Lichtarge. 2015. Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 2019–2028. https://doi.org/10.1145/2783258.2788609

[14] Saeedeh Shekarpour, Sören Auer, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Sebastian Hellmann, and Claus Stadler. 2011. Keyword-Driven SPARQL Query Generation Leveraging Background Knowledge. In *Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011, Campus Scientifique de la Doua, Lyon, France, August 22-27, 2011*. 203–210.

[15] Saeedeh Shekarpour, Konrad Höffner, Jens Lehmann, and Sören Auer. 2013. Keyword Query Expansion on Linked Data Using Linguistic and Semantic Features. In *2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013*. 191–197.

[16] Saeedeh Shekarpour, Edgards Marx, Sören Auer, and Amit Sheth. 2017. RQUERY: Rewriting Natural Language Queries on Knowledge Graphs to Alleviate the Vocabulary Mismatch Problem. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, USA*.

[17] Saeedeh Shekarpour, Axel-Cyrille Ngonga Ngomo, and Sören Auer. 2013. Question Answering on Interlinked Data. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*. 1145–1156.

[18] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge Base Completion via Search-based Question Answering. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, New York, NY, USA, 515–526. https://doi.org/10.1145/2566486.2568032