Challenges and Innovations in Building a Product Knowledge Graph

Xin Luna Dong
Amazon.com
Seattle, USA
lunadong@amazon.com

ABSTRACT

Knowledge graphs have been used to support a wide range of applications and enhance search results for multiple major search engines, such as Google and Bing. At Amazon we are building a Product Graph, an authoritative knowledge graph for all products in the world. The thousands of product verticals we need to model, the vast number of data sources we need to extract knowledge from, the huge volume of new products we need to handle every day, and the various applications in Search, Discovery, Personalization, Voice, that we wish to support, all present big challenges in constructing such a graph.

In this talk we describe four scientific directions we are investigating in building and using such a knowledge graph. First, we have been developing advanced extraction technologies to harvest product knowledge from *semi-structured* sources on the web and from *text* product profiles. Our annotation-based extraction tool selects a few webpages (typically below 10 pages) from a website for annotations, and can derive XPaths to extract from the whole website with average precision and recall of 97% [1]. Our distantly supervised extraction tool, CERES, uses an existing knowledge graph to automatically generate (noisy) training labels, and can obtain a precision over 90% when extracting from long-tail websites in various languages [1]. Our OpenTag technique extends state-of-the-art techniques such as Recursive Neural Network (RNN) and Conditional Random Field with attention and active learning, to achieve over 90% precision and recall in extracting attribute values (including values unseen in training data) from product titles, descriptions, and bullets [3].

Second, we build tools that enable hands-off-the-wheel knowledge integration and cleaning. We train random forest models to link records that refer to the same real-world entity with high precision and recall. We design active learning strategies that can reduce the number of training labels by 2 orders of magnitude to reach 99% precision and recall. For cleaning, we apply state-of-the-art knowledge fusion techniques to clean errors resulted from extraction mistakes and source errors.

Third, we have been conducting graph mining to decide importance of entities and relationships in a graph, which can be taken as important signals for search ranking, and to generate embeddings for entities and relationships, which can be used as signals to train models for downstream tasks. As an example, our embedding techniques are able to identify false triples with linkage errors, mistaken relationships, and other random mistakes. We have also developed embeddings that provide evidence for linking entities from different knowledge graphs [2].

Finally, we have been developing techniques that enable us to apply knowledge learning with human in the loop. We are designing a system that best allocates resources including machines, labelers, and data scientists, to enable us to train high-quality models with the minimum resources.

This talk will present our progress to achieve near-term goals in each direction to make production impact, and show the many research opportunities towards our moon-shot goals.

KDD 2018, August 19-23, 2018, London, United Kingdom.

ACM ISBN 978-1-4503-5552-0/18/08.
DOI: https://doi.org/10.1145/3219819.3219938

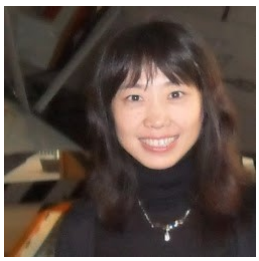
CCS Concepts/ACM Classifiers

• Information systems > Information integration

Author Keywords

Knowledge extraction, knowledge fusion, entity linkage, data cleaning, graph mining, human-in-the-loop.

BIOGRAPHY



Xin Luna Dong is a Principal Scientist at Amazon, leading the efforts of constructing Amazon Product Knowledge Graph. She was one of the major contributors to the Google Knowledge Vault project, and has led the Knowledge-based Trust project, which is called the "Google Truth Machine" by Washington's Post. She has co-authored book "Big Data Integration", published 70+ papers in top conferences and journals, and given 30+ keynotes/invited-talks/tutorials. She got the VLDB Early Career Research Contribution Award for "advancing the state of the art of knowledge fusion", and got the Best Demo award in Sigmod 2005. She serves in VLDB endowment and PVLDB advisory committee, and is the PC co-chair for Sigmod 2018 and WAIM 2015.

REFERENCES

[1] Colin Lockard, Xin Luna Dong, Arash Einolghozati, Prashant Shiralkar. Ceres: Distantly supervised relation extraction from the semi-structured web. In *VLDB*, 2018.

[2] Rakshit Trivedi, Bunyamin Sisman, Jun Ma, Christos Faloustos, Hongyuan Zha, Xin Luna Dong. LinkNBed: Multi-Graph representation learning with entity linkage. In *ACL*, 2018.

[3] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, Feifei Li. OpenTag: Open attribute value extraction from product profiles. In *SigKDD*, 2018.