2019 IEEE 35th International Conference on Data Engineering (ICDE)

Building a Broad Knowledge Graph for Products

Xin Luna Dong
Amazon Inc
lunadong@amazon.com

I. Abstract

A knowledge graph (KG) describes entities and relations between them; for example, between entities *Amazon* and *Seattle*, there can be a *headquarter_located_at* relation. Recent years have witnessed broad applications of KG in search (*e.g.*, by *Google* and *Bing*) and question answering (*e.g.*, by *Amazon Alexa* or *Google Home*). In this talk, we ask the question: *Can one build a knowledge graph (KG) for all products in the world?*

The techniques for curating knowledge for popular domains such as *Music*, *Movies*, or *Sport* are fairly mature in industry [1], [2]. An *ontology* (schema) is defined to describe types and relations in a domain as the backbone for the KG. According to the ontology, knowledge is ingested for each domain from a few manually selected data sources, where Wikipedia and WikiData are often among the major ones, and then integrated using *entity linkage* [3], [4] techniques. Oftentimes heavy manual curation is required to maintain high quality of the knowledge [1], [5].

Various KGs have been proposed both in industry (*e.g.*, *Google Knowledge Graph*, *Bing Satori Graph* [2]), and in research literature (*e.g.*, Yago [6], NELL [7], Diadem [8], Knowledge Vault [9]). However, we are not aware of any KG that contains rich product information. Constructing a KG about products poses the following three unique challenges.

First, the number of product categories is in hundreds, making it hard to build the ontology. Product properties vastly differ between categories (*e.g.*, compare TVs and dog food categories), and also evolve over time (*e.g.*, older TVs did not have WiFi connectivity), making it hard to design comprehensive ontology and keep it up-to-date.

Second, the number of products is in billions and the number of product properties is in thousands, making it hard to collect knowledge for all products on their applicable properties. It is even more challenging because there is no major source to curate product knowledge from. The public data sources such as Wikipedia contains very limited information for products.

Third, a large number of new products emerge on a daily basis, making it hard to refresh the knowledge. Any manual process involved in building the graph would prevent the graph from staying up-to-date, and full automatic and efficient knowledge update mechanisms are required.

We start tackling the problem of building a product KG by building a prototype for a large product category. Instead of taking a traditional approach of curating a *rich* knowledge graph, we automatically build a *broad* graph. The broad graph is a bipartite graph, where one side contains nodes representing products, and the other side contains nodes representing product properties, such as brand, size, and color. We then enrich and clean the graph by applying machine learning (ML) techniques: we apply knowledge extraction to extract structured properties from product names and descriptions; we extract knowledge from external web sources and link to Catalog products; we group and normalize property values such as brands and flavors; and we detect invalid values and inconsistency in the data to clean the knowledge. The enrichment and cleaning are conducted in a *pay-as-you-go* fashion, triggered by detected customer needs, available data sources, or problems in data.

In particular, this talk present the following three contributions we made in this prototype.

1) We propose a new way for KG construction: instead of starting with a fully-fledged ontology, we start with a core ontology and gradually enrich the ontology and the data in a pay-as-you-go fashion.
2) We invented and applied a suite of state-of-the-art ML techniques that together lead us to full automation for knowledge enrichment and cleaning, paving the way to extend our prototype to other categories.
3) We present a prototype KG for a large category of products. To the best of our knowledge, this is the first KG that covers a large product category.

References

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 2008, pp. 1247–1250.
[2] Y. Gao, J. Liang, B. Han, M. Yakout, and A. Mohamed, "Building a large-scale, accurate and fresh knowledge graph," in *SigKDD*, 2018.
[3] X. L. Dong and D. Srivastava, "Big data integration," *PVLDB*, 2013.
[4] L. Getoor and A. Machanavajjhala, "Entity resolution: Theory, practice, & open challenges." *PVLDB*, vol. 5, no. 12, pp. 2018–2019, 2012.
[5] X. L. Dong, "Leave no valuable data behind: The crazy ideas and the business," *PVLDB*, 2016.
[6] M. S. Fabian, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge unifying wordnet and wikipedia," in *16th International World Wide Web Conference, WWW*, 2007, pp. 697–706.
[7] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, "Toward an architecture for never-ending language learning." in *AAAI*, vol. 5. Atlanta, 2010, p. 3.
[8] T. Furche, G. Gottlob, G. Grasso, O. Gunes, X. Guo, A. Kravchenko, G. Orsi, C. Schallhart, A. Sellers, and C. Wang, "Diadem: domain-centric, intelligent, automated data extraction methodology," in *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012, pp. 267–270.
[9] X. 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 *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 601–610.

2375-026X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDE.2019.00010