Utilizing Knowledge Graphs in Text-centric Information Retrieval

Laura Dietz
University of New Hampshire
Durham NH, USA
dietz@cs.unh.edu

Alexander Kotov
Wayne State University
Detroit MI, USA
kotov@wayne.edu

Edgar Meij
Bloomberg L.P.
London, United Kingdom
edgar.meij@acm.org

Abstract

The past decade has witnessed the emergence of several publicly available and proprietary knowledge graphs (KGs). The increasing depth and breadth of content in KGs makes them not only rich sources of structured knowledge by themselves but also valuable resources for search systems. A surge of recent developments in entity linking and retrieval methods gave rise to a new line of research that aims at utilizing KGs for text-centric retrieval applications, making this an ideal time to pause and report current findings to the community, summarizing successful approaches, and soliciting new ideas. This tutorial is the first to disseminate the progress in this emerging field to researchers and practitioners.[1]

1. MOTIVATION

General-purpose knowledge graphs (KGs) such as Wikipedia, DBpedia, Freebase, and WikiData but also proprietary KGs such as Google's Knowledge Graph and Microsoft's Satori are growing in breadth and depth every day, rendering them a valuable resource for search systems, both on the Web and elsewhere. The research advances in entity linking and retrieval—both critical tasks in the contexts of KGs—that started with INEX and TREC initiatives, led to a surge in recent interest around utilizing KGS for text retrieval.

In this tutorial we define an *entity* as an atomic, identifiable object which can have a distinct and independent existence. Moreover, entities are interconnected using heterogeneous relationships and can be described using names, descriptions, and types. We therefore distinguish an entity from a *mention*, i.e., a text segment which refers to an entity. Most (if not all) KGs share this notion of entities and, historically, information retrieval methods have used such information as a source for expansion terms [1]. This tutorial places these established methods in context and explains how KGs can be used in even more effective ways for text retrieval. Given text, entity linking methods mark-up entity mentions and provide unambiguous pointers to entities in the KG. In particular, with effective entity representation schemes [6], fast and accurate entity retrieval [11, 18] and linking [2, 10] methods as well as collections [5], bidirectional connections between text and KGs are readily available. These developments give rise to novel research angles on how to effectively utilize KGs in the context of textual information retrieval tasks, ranging from ad hoc document retrieval to digital assistants.

Research on text search systems typically addresses three core angles: (1) keyword matching and indexing, (2) query expansion models using (pseudo-)relevance feedback and query logs, and (3) minimizing redundancy and result diversification. Most work on these fronts operate at the level of terms and phrases. Entity linking and retrieval make it feasible to efficiently tap into the rich information provided by KGs.

In particular, a new line of research is emerging on how to effectively use entity-centric knowledge repositories to understand textual data and estimate relevance for information needs. This research direction encompasses a range of tasks assessing document-centric entity prominence [4], discovering emerging entities [7], as well as extracting and mining entity aspects [13] from search logs. This kind of entity-centric structured information can be effectively exploited to estimate the relevance of text with respect to an information need—especially when combined with positional information in documents through entity links and proximity with query terms. Such approaches match information in KGs with information in text and obtain state-of-the-art performance for ad hoc document retrieval [3, 9, 12, 17]. This tutorial explains how to use existing technology to obtain similar performance improvements for related tasks and applications.

2. TOPICS

In this tutorial, we provide an overview of state-of-the-art methods and outline open research problems in order to encourage new contributions in this area. A cross-cutting issue addressed in this tutorial is the heterogeneity of different kinds of data and perspectives offered by KGs. For example, each entity is represented by different names, textual attributes, relations, and memberships in taxonomic types/category hierarchies. Throughout the tutorial we discuss how each of these different types of information can be used to: (1) retrieve a set of entities for a textual information need, or more broadly, how to assess the relevance of KB elements for the topic, (2) how to recognize mentions of entities from a KG in a textual fragment and (3) how to use these mentions to determine relevance of documents and text fragments.

[1] All tutorial resources are available online at http://github.com/laura-dietz/tutorial-utilizing-kg

WSDM 2017 February 06-10, 2017, Cambridge, United Kingdom

ACM ISBN 978-1-4503-4675-7/17/02.

DOI: http://dx.doi.org/10.1145/3018661.3022756

We start the tutorial with a brief overview of different types of KGs, their structure, and information contained in popular general-purpose and domain-specific KGs. Next, we provide a recap on entity linking and retrieval on knowledge graphs. We discuss different entity representation methods [6], which is followed by presentation of recent advances on the design of retrieval models for ad hoc entity retrieval [11, 18] and ranking [14]. This is essential technology which the remainder of the tutorial builds on.

We then present the details of previously proposed systems that successfully leverage KGs to improve ad hoc document retrieval. These systems combine the notion of entity retrieval and semantic search on one hand, with text retrieval models and entity linking on the other. Examples of systems to be discussed include EsdRank [17], Entity Query Feature Expansion [3], as well as Latent Entity Space [9]. We stress findings on how information from semantic networks, latently relevant entities, entity types, and relations integrate with textual retrieval models. We cover query expansion approaches as well as adaptations of the learning-to-rank paradigm for this task [3, 8, 17]. We also highlight the results of a recent study comparing statistical term association graphs with knowledge bases for query expansion [1]. Many entities have different aspects [9, 13], of which only one needs to be relevant in order to render the entity relevant for the query. In turn, to rank text, even with perfect entity linking accuracy, it is not sufficient to match relevant entities. We discuss approaches to extract relevant aspects of entities and how they are expressed in text in order to assess relevance for the information.

Most KGs contain both hyperlinks as well as typed relational facts between entities, the former appearing in abundance and the latter being often sparse and biased to entities of particular types. This graph structure can help understand the context as long as concept drift can be avoided [8]. We discuss relation extraction systems that extracted such information from text and, with the advent of schema-less, so called "open information extraction" methods, through which even more links with term-associations become available. Schuhmacher et al. [15] found that schema-based relation extraction can be used to find relevant relations for a query. In contrast, Voskarides et al. [16] focuses on the inverse problem of retrieving support passages for given relations.

We also include a detailed, worked-out example of an end-to-end application of utilizing knowledge graphs for IR with the aim of bringing together the tutorial topics in a single, unified example to illustrate how to utilize KGs with existing retrieval models and tools for enriching text. Specifically, for every element in a KG, we answer the following two questions: (1) "How to determine if an element is relevant to an information need?" and (2) "Assuming that an element is relevant, how can it inform us which documents are relevant?". We conclude the tutorial with a brief discussion of open research challenges in this emerging field.

Acknowledgements

This work was in part funded by the Elitepostdoc program of the BW-Stiftung and the University of New Hampshire.

Presenters

Prof. Dr. Laura Dietz is an Assistant Professor at University of New Hampshire. Her teaching and research topics connect information retrieval, machine learning, and knowledge graphs. Before that she was working at Mannheim University, and University of Massachusetts and obtained her Ph.D. form. the Max Planck Institute for Informatics.

Prof. Dr. Alexander Kotov is an Assistant Professor in the Department of Computer Science at Wayne State University. His general research interests lie at the intersection of information retrieval, textual data mining, and health informatics. Previously he was a post-doctoral fellow at Emory University and obtained his Ph.D. at University of Illinois at Urbana-Champaign.

Dr. Edgar Meij is a Senior Scientist at Bloomberg. His research focuses on all applications and aspects of knowledge graphs, entity linking, and semantic search. Previously he was a research scientist at Yahoo Labs and a postdoc at the University of Amsterdam, where he also obtained his Ph.D. He regularly teaches university courses and conference tutorials, e.g., at EACL, SIGIR, WWW, WSDM, and ICTIR.

References

[1] S. Balaneshinkordan and A. Kotov. An empirical comparison of term association and knowledge graphs for query expansion. In *ECIR*, 2016.

[2] R. Blanco, G. Ottaviano, and E. Meij. Fast and space-efficient entity linking for queries. In *WSDM*, 2015.

[3] J. Dalton, L. Dietz, and J. Allan. Entity query feature expansion using knowledge base links. In *SIGIR*, 2014.

[4] J. Dunietz and D. Gillick. A new entity salience task with millions of training examples. In *EACL*, 2014.

[5] E. Gabrilovich, M. Ringgaard, and A. Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1, 2013.

[6] D. Graus, M. Tsagkias, W. Weerkamp, E. Meij, and M. de Rijke. Dynamic collective entity representations for entity ranking. In *WSDM*, 2016.

[7] J. Hoffart, D. Milchevski, and G. Weikum. STICS: Searching with Strings, Things, and Cats. In *SIGIR*, 2014.

[8] A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In *WSDM*, 2012.

[9] X. Liu and H. Fang. Latent entity space: a novel retrieval approach for entity-bearing queries. *Information Retrieval Journal*, 18(6):473–503, 2015.

[10] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM*, 2012.

[11] F. Nikolaev, A. Kotov, and N. Zhiltsov. Parameterized fielded term dependence models for ad-hoc entity retrieval from knowledge graph. In *SIGIR*, 2016.

[12] H. Raviv, O. Kurland, and D. Carmel. Document retrieval using entity-based language models. In *SIGIR*, 2016.

[13] R. Reinanda, E. Meij, and M. de Rijke. Mining, ranking and recommending entity aspects. In *SIGIR*, 2015.

[14] M. Schuhmacher, L. Dietz, and S. Paolo Ponzetto. Ranking Entities for Web Queries through Text and Knowledge. In *CIKM*, 2015.

[15] M. Schuhmacher, B. Roth, S. P. Ponzetto, and L. Dietz. Finding relevant relations in relevant documents. In *ECIR*, 2016.

[16] N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke, and W. Weerkamp. Learning to explain entity relationships in knowledge graphs. In *ACL-IJCNLP*, 2015.

[17] C. Xiong and J. Callan. Esdrank: Connecting query and documents through external semi-structured data. In *CIKM*, 2015.

[18] N. Zhiltsov, A. Kotov, and F. Nikolaev. Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In *SIGIR*, 2015.