Graph-Based Entity-Oriented Search: A Unified Framework in Information Retrieval

José Devezas[0000−0003−2780−2719]

INESC TEC and Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n, 4200-465, Porto, Portugal
jld@fe.up.pt

Abstract. Modern search engines have evolved beyond document retrieval. Nowadays, the information needs of the users can be directly satisfied through entity-oriented search, by taking into account the entities that better relate to the query, as opposed to relying exclusively on the best matching terms. Evolving from keyword-based to entity-oriented search poses several challenges, not only regarding the understanding of natural language queries, which are more familiar to the end-user, but also regarding the integration of unstructured documents and structured information sources such as knowledge bases. One opportunity that remains open is the research of unified frameworks for the representation and retrieval of heterogeneous information sources. The doctoral work we present here proposes graph-based models to promote the cooperation between different units of information, in order to maximize the amount of available leads that help the user satisfy an information need.

Keywords: Entity-oriented search · Graph-based models · Representation models · Retrieval models

1 Research Statement

We propose a graph-based unified framework to cover two fronts in entity-oriented search: representation and retrieval. On one side, we propose the joint representation of terms, entities and their relations, as a collection-based model of corpora and knowledge bases. Our idea is to provide a novel way to index documents and entities such that, from design, the aim is to seamlessly integrate units of information of different types, providing a higher-level of flexibility and expressiveness than the inverted index or the triplestore alone. On the other side, we propose a universal ranking function that should be issued over the representation model that we design, in order to rank the different units of information, based on other input units of information. The goal is to obtain a unique ranking function that, depending on the query and the target result, will be able to solve several tasks from entity-oriented search, be it document and entity retrieval or the recommendation-alike tasks of related entity finding and entity list completion.

2 Motivation

Classical information retrieval has focused on the representation of documents and their features, as well as the retrieval and ranking of those documents based on a representation model — usually the inverted index. Entity-oriented search, however, relies simultaneously on corpora and knowledge bases, which have mismatching representation models — usually the inverted index and the quad indexes in a triplestore. Moreover, if we look at the definitions of 'information retrieval' and 'entity-oriented search', which are ten years apart, we will find a collision between two paradigms that are seemingly different.

> *"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."*
>
> – Manning et al. [8, Ch.1], 2008

> *"Entity-oriented search is the search paradigm of organizing and accessing information centered around entities, and their attributes and relationships."*
>
> – Balog [2, §1.3, Def.1.5], 2018

While the first is focused on unstructured data (documents), the second largely relies on structured data (entities and relations). Moreover, entity-oriented search also incorporates ad hoc document retrieval, as long as it leverages entities [2, Ch.8]. This collision of definitions, along with the consideration for both documents and entities, means that an integration of representation and/or retrieval models must happen at one of the stages of indexing and/or search. One way to tackle the problem is through separated subsystems based on an inverted index and a triplestore, which either independently respond to the query or contribute with different signals to the final scores. Another way to approach the problem is by considering data as a joint representation of corpora and knowledge bases — also known as combined data, according to Bast el al. [4, Def.2.3].

Being able to jointly index and search over the documents and entities in such a collection of combined data, through their relations, should provide a way to harness all available information, both from unstructured and structured data, as well as from the cross-referencing of both. Bast and Buchhold [3] have justified the need for such a unified framework based on the example of a friendship relation that could only be found in the text, but should influence retrieved triples by establishing new connections. One way to tackle this problem and to build a seamless model of text and knowledge is to use graphs, since they have already been used for the retrieval of documents [5,11] and entities [1,14]. While models like the graph-of-word [11] were designed to answer keyword-based queries with a list of ranked documents, entities in a knowledge graph like DBpedia [1] are still more frequently retrieved through a structured query language like SPARQL [9]. Our goal is to propose a unified model that is able to support the ranking of documents and entities, for a given keyword query, entity, or set of entities, without the need for a structured query language. At the same time, the model should retain the complex relations between entities, exploiting them for all tasks.

3 Description

This doctoral work focuses on improving entity-oriented search performance by researching and proposing novel graph-based approaches to better harness the power of all available information sources. The goal is to satisfy the user's need to answer increasingly complex verbose queries by automatically cross-referencing and weighting related information from unstructured and structured data, at a common stage. It is particularly relevant to research and develop representation models that are able to support the integration of existing tasks through a general ranking function. This unified model should be able to provide a clear framework for entity-oriented search, similar to what the inverted index and TF-IDF are for ad hoc document retrieval, unlocking a larger universe of possibilities for the combination of heterogeneous and multimodal information sources.

The steps required to build such a model are the following:

1. Proposing a graph-based joint representation model of terms, entities and their relations, for indexing corpora and knowledge bases.
2. Proposing a universal ranking function, to be issued over the graph, that can support:
   (a) Ad hoc document retrieval (leveraging entities);
   (b) Ad hoc entity retrieval (leveraging documents);
   (c) At least one of the following recommendation-alike tasks:
      i. Related entity finding (leveraging documents and entities);
      ii. Entity list completion (leveraging documents and entities).
3. Validating the viability of the unified model based on:
   (a) The effectiveness of individual tasks;
   (b) Different combined data test collections.

The main research questions are linked to each of the previous three steps:

1. Which nodes, edges and respective weighting functions (if any) should we use to jointly represent documents and entities for the best overall performance of the considered retrieval tasks?
   – Is it constructive to consider synonyms?
   – And contextual similarity?
   – What about syntactic dependencies?
   – Which entity relations should we consider?
   – And which term-entity relations?
2. How can we model a ranking function, over the proposed representation model, such that the same ranking function can be applied for ranking different units of information, either regarding relevance to a keyword query, or relatedness with one or multiple input entities?
3. How can we evaluate the graph-based unified model?
   (a) Are there any combined data test collections that cover multiple tasks?
   (b) Which variations of the representation model should we consider?
   (c) Which parameter values of the ranking function should we assess?

4 Research Methodology

The research methodology is based on the well-established approach for information retrieval experimentation [10], supported on the empirical cycle and the test collections provided by evaluation initiatives [13]. An example of a combined data test collection is the INEX 2009 Wikipedia collection [12], which consists of a Wikipedia snapshot, formatted as XML, containing inter-document links, as well as entity and link annotations using XML tags based on the WordNet thesaurus. While this collection is from 2009, its great advantage is that there are topics and relevance judgments that can be used to evaluate the tasks of ad hoc document retrieval, ad hoc entity retrieval and entity list completion.

One of the main experiments we already carried was based on the graph-of-entity [6], a model that we proposed, inspired by the graph-of-word [11]. We also experimented in-depth with hypergraph-of-entity [7]. This model was built as an attempt to tackle the complexity and scalability issues of the graph-of-entity, as well as to improve the expressiveness and flexibility of the model and thus its generality. We evaluated both models based on subsets of the INEX 2009 Wikipedia collection, measuring mean average precision, precision at a cutoff of 10 and normalized discounted cumulative gain for the top-p results. While we were able to propose a hypergraph-based model and a universal ranking function (the random walk score), we have only tested ad hoc document retrieval and ad hoc entity retrieval. We are working on an experiment based on entity list completion, which we hope will further support our idea of a unified framework for entity-oriented search. All three experiments were designed around the topics and relevance judgments from the INEX 2010 Ad Hoc track and the INEX 2009 XML Entity Ranking track. Evaluation depends on the existence of test collections based on combined data [4, Def. 2.3], with multiple associated topics and relevance judgments corresponding to the different tasks to be supported and generalized by the model. So far, we have identified the INEX 2009 Wikipedia collection, as well as the TREC Washington Post Corpus as two potential collections to be used in this context.

5 Research Issues

- Can we jointly represent unstructured and structured data as a graph?
- Will this unlock novel ranking strategies?
- Is it possible for these ranking strategies to support the generalization of entity-oriented search tasks?
- Will the incorporation of explicit and implicit information derived from the relations between text, found in corpora, and entities, found in knowledge bases, improve overall retrieval effectiveness? Or is there, instead, a trade-off between generalization and effectiveness?

Acknowledgements

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program Human Capital (POCH), supported by the European Social Fund and by national funds from MCTES.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. pp. 722–735 (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Balog, K.: Entity-Oriented Search. Springer (2018)
3. Bast, H., Buchhold, B.: An Index for Efficient Semantic Full-text Search. In: Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management. pp. 369–378 (2013). https://doi.org/10.1145/2505515.2505689
4. Bast, H., Buchhold, B., Haussmann, E., et al.: Semantic search on text and knowledge bases. Foundations and Trends® in Information Retrieval **10**(2-3), 119–271 (2016)
5. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Information Retrieval **15**(1), 54–92 (2012). https://doi.org/10.1007/s10791-011-9172-x
6. Devezas, J., Lopes, C., Nunes, S.: Graph-of-Entity: A model for combined data representation and retrieval. In: 8th Symposium on Languages, Applications and Technologies, SLATE 2019, June 27-28, 2019, Coimbra, Portugal (2019). https://doi.org/10.4230/OASIcs.SLATE.2019.1
7. Devezas, J., Nunes, S.: Hypergraph-of-entity: A unified representation model for the retrieval of text and knowledge. Open Computer Science Journal **10**(1) (Jun 2019)
8. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
9. Pérez, J., Arenas, M., Gutiérrez, C.: Semantics and complexity of SPARQL. ACM Trans. Database Syst. **34**(3), 16:1–16:45 (2009). https://doi.org/10.1145/1567274.1567278
10. Robertson, S.E.: The methodology of information retrieval experiment. In: Information Retrieval Experiment, pp. 9–31. Butterworth-Heinemann (1981)
11. Rousseau, F., Vazirgiannis, M.: Graph-of-word and TW-IDF: New approach to ad hoc IR. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. pp. 59–68. ACM (2013)
12. Schenkel, R., Suchanek, F.M., Kasneci, G.: YAWN: A semantically annotated wikipedia XML corpus. In: Datenbanksysteme in Business, Technologie und Web (BTW 2007), 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Proceedings, 7.-9. März 2007, Aachen, Germany. pp. 277–291 (2007), http://subs.emis.de/LNI/Proceedings/Proceedings103/article1404.html

13. Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers. pp. 355–370 (2001). https://doi.org/10.1007/3-540-45691-0_34
14. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledge base. Communications of the ACM **57**, 78–85 (2014), http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext