# Graph-Based Entity-Oriented Search

José Devezas
INESC TEC and Faculty of Engineering, University of Porto
*jld@fe.up.pt*

**Abstract**

Entity-oriented search has revolutionized search engines. In the era of Google Knowledge Graph and Microsoft Satori, users demand an effortless process of search. Whether they express an information need through a keyword query, expecting documents and entities, or through a clicked entity, expecting related entities, there is an inherent need for the combination of corpora and knowledge bases to obtain an answer. Such integration frequently relies on independent signals extracted from inverted indexes, and from quad indexes indirectly accessed through queries to a triplestore. However, relying on two separate representation models inhibits the effective cross-referencing of information, discarding otherwise available relations that could lead to a better ranking. Moreover, different retrieval tasks often demand separate implementations, although the problem is, at its core, the same. With the goal of harnessing all available information to optimize retrieval, we explore joint representation models of documents and entities, while taking a step towards the definition of a more general retrieval approach. Specifically, we propose that graphs should be used to incorporate explicit and implicit information derived from the relations between text found in corpora and entities found in knowledge bases. We also take advantage of this framework to elaborate a general model for entity-oriented search, proposing a universal ranking function for the tasks of ad hoc document retrieval (leveraging entities), ad hoc entity retrieval, and entity list completion.

At a conceptual stage, we begin by proposing the graph-of-entity, based on the relations between combinations of term and entity nodes. We introduce the entity weight as the corresponding ranking function, relying on the idea of seed nodes for representing the query, either directly through term nodes, or based on the expansion to adjacent entity nodes. The score is computed based on a series of geodesic distances to the remaining nodes, providing a ranking for the documents (or entities) in the graph. In order to improve on the low scalability of the graph-of-entity, we then redesigned this model in a way that reduced the number of edges in relation to the number of nodes, by relying on the hypergraph data structure. The resulting model, which we called hypergraph-of-entity, is the main contribution of this thesis. The obtained reduction was achieved by replacing binary edges with $n$-ary relations based on sets of nodes and entities (undirected document hyperedges), sets of entities (undirected hyperedges, either based on co-occurrence or a grouping by semantic subject), and pairs of a set of terms and a set of one entity (directed hyperedges, mapping text to an object).

We introduce the random walk score as the corresponding ranking function, relying on the same idea of seed nodes, similar to the entity weight in the graph-of-entity. Scoring based on

this function is highly reliant on the structure of the hypergraph, which we call representation-driven retrieval. As such, we explore several extensions of the hypergraph-of-entity, including relations of synonymy, or contextual similarity, as well as different weighting functions per node and hyperedge type. We also propose TF-bins as a discretization for representing term frequency in the hypergraph-of-entity. For the random walk score, we propose and explore several parameters, including length and repeats, with or without seed node expansion, direction, or weights, and with or without a certain degree of node and/or hyperedge fatigue, a concept that we also propose.

For evaluation, we took advantage of TREC 2017 OpenSearch track, which relied on an online evaluation process based on the Living Labs API, and we also participated in TREC 2018 Common Core track, which was based on the newly introduced TREC Washington Post Corpus. Our main experiments were supported on the INEX 2009 Wikipedia collection, which proved to be a fundamental test collection for assessing retrieval effectiveness across multiple tasks. At first, our experiments solely focused on ad hoc document retrieval, ensuring that the model performed adequately for a classical task. We then expanded the work to cover all three entity-oriented search tasks. Results supported the viability of a general retrieval model, opening novel challenges in information retrieval, and proposing a new path towards generality in this area.

**Awarded by:** Universities of Minho, Aveiro, and Porto, Portugal **on** 26 January 2021.
**Supervised by:** Sérgio Nunes.
**Available at:** https://hdl.handle.net/10216/133205.

## Selected Publications

José Devezas and Sérgio Nunes. Hypergraph-of-entity. *Open Computer Science*, 9(1):103–127, 2019. URL https://doi.org/10.1515/comp-2019-0006.

José Devezas and Sérgio Nunes. Army ANT: A workbench for innovation in entity-oriented search. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 449–453. Springer, 2020a. URL https://doi.org/10.1007/978-3-030-45442-5_56.

José Devezas and Sérgio Nunes. Characterizing the hypergraph-of-entity and the structural impact of its extensions. *Applied Network Science*, 5(1):79, 2020b. URL https://doi.org/10.1007/s41109-020-00320-z.