A semantic search approach for hyper relational knowledge graphs

Veronica dos Santos [1], **Sérgio Lifschitz (supervisor)**[1]

[1]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

{vdsantos, sergio}@inf.puc-rio.br

Level: Doctorate degree

Enrolment date: 2019/March

Due date: 2023/February

Completed activities: All mandatory course credits (2019-2020), Qualifying exam (2021), Bibliographic review (2020-2021), Problem Statement (2021.1)

Ongoing activities: Research proposal formulation (2021.1), proposal defense (2021.2), solution specification and implementation (2021-2022)

Future activities: thesis writing (2022.1-2023.1), thesis defense (2023.1), results publication (2021.2-2023.1)

Abstract. *Information Retrieval Systems usually employ syntactic search techniques to match a set of keywords with the indexed content to retrieve results. But pure keyword-based matching lacks on capturing user's search intention and context and suffers from natural language ambiguity and vocabulary mismatch. The hypothesis raised is that the use of embeddings in a semantic search approach will make the results more meaningful. Embeddings allow minimizing problems arising from terminology and context mismatch. This work proposes a semantic similarity function to support semantic search based on hyper relational knowledge graphs (KG). This function uses embeddings to find the most similar KG entities to satisfy a user query.*

Resumo. *Sistemas de Recuperação de Informações geralmente empregam técnicas de pesquisa sintática para combinar um conjunto de palavras-chave com o conteúdo indexado para recuperar os resultados. Mas correspondência baseada somente em palavras-chave não consegue capturar a intenção e o contexto de busca além de sofrer com a ambigüidade da linguagem natural e a incompatibilidade de vocabulário. Considerando esse cenário, a hipótese levantada é que o uso de embeddings em uma abordagem de busca semântica tornará os resultados mais significativos. Embeddings permitem minimizar problemas decorrentes de incompatibilidade de terminológica e de contexto. Este trabalho propõe uma função de similaridade semântica para apoiar a busca baseada em grafo de conhecimento (KG) hiper relacional. Esta função usa embeddings para encontrar as entidades do KG mais semelhantes que atendam uma consulta.*

1. Context and Motivation

Information Retrieval (IR) Systems usually employ search techniques to match a set of keywords from the user's query, that represents their information need, with the indexed content of extensive collections and scores these matches to ranking the result [Manning et al. 2008]. But pure keyword-based matching lacks on capturing user's search intention and context and suffers from natural language ambiguity and vocabulary mismatch.

At BioBD Research Lab from PUC-Rio, we developed two keyword search applications: Quem@PUC and Busc@NIMA. Both access a single NoSQL database (a TripleStore) that integrates academic and research information from different sources. Quem@PUC aims to answer questions like *Which researchers are related to subject X?* This relationship could be established through a research publication, a research project, or even an academic discipline from a graduate course. Busc@NIMA has the same objective in a narrower scope since the subject of interest X belongs to the Environment domain, which is interdisciplinary.

Keyword Search (KwS) is the ubiquitous approach to finding information in the IR field. So why do we employ KwS over databases? To provide database user interfaces as simple as IR systems, without requiring query language knowledge. Due to the increase in the number of knowledge bases (KB) published, new approaches for KwS over KBs have been developed. Different from other KwS approaches over RDF, as those analysed by [Dosso and Silvello 2020], here fixed SPARQL queries combined with text search operators, based on inverted indexes, retrieve all matched sub graphs from the database and build virtual documents at runtime. Currently, both applications perform mainly a syntactic search to match the keyword list with the indexed content.

However, the terms used to describe resources by researchers and academic staff may mismatch with the user keyword list. Variations of the same concept with different words (synonymy) and multiple meanings for the same word (ambiguity) require that the search engine considers meaning similarity. It is also essential to consider that when users provide keywords into the search engine query box, sometimes, they precisely know what they want to retrieve. In some other cases, users want to explore and extend their knowledge about related topics or people. Semantic search approaches can tackle these issues to improve the results.

From this context, the identified research problem is: Given an information source modeled as a hyper relation Knowledge Graph(KG) H and a keyword list Q as input, retrieve the top-K subgraphs h from H that are the most semantic similar to Q considering a context C. In order to convert a keyword list Q into a query over the KG H, first, we need to semantic parse the keywords to understanding their meaning and context. The hypothesis raised is that embedding in a semantic search approach will make search results more meaningful. Embeddings allow minimizing problems arising from terminology and context mismatch. The KG guides the knowledge acquisition user experience.

2. Background

For the scope of this research, we used the semantic search definition as stated bellow [Cudré-Mauroux 2019]: *"Semantic Search regroups a set of techniques designed to improve traditional document or knowledge base search. Semantic search aims at better*

grasping the context and the semantics of the user query or the indexed content by leveraging natural language processing, Semantic Web, and machine learning techniques to retrieve more relevant results from a search engine."

Semantic search techniques can be applied to the query, indexed content, or representation of the knowledge domain, a.k.a KB, to promote meaningful retrieval results. According to [Bast et al. 2016], semantic search solutions can be classified using two dimensions: data type (text, KB, and combination of both) and search approach (keyword list, structured query, natural language, and question and answering). Besides that, semantic search also employs specific ranking and indexing criteria and may also use ontologies, inference, and natural language processing (NLP).

Knowledge graphs

A simple graph G can be defined as a tuple (V, E), where V is its set of vertices, and $E \subseteq V \times V$ is its set of edges. A directed graph has an edges $e = u, v \in E$ as a tuple whose first element u is the source and the second v is the target vertex. A weighted graph can be modeled by a triple (V, E, w), where $w : E \mapsto R$ is function that gives the weight of a given edge. And a hypergraph H generalizes G enabling edges to have any number of vertices.

There are some semantic search systems target exclusively to KB, as can be found in [Bast et al. 2016]. KB can be represented by KG and ontologies. A KG can be defined as a multi-relational heterogeneous (more than one type of entity) graph composed of nodes representing entities with their attributes and edges representing relations between entities and connecting at least a pair of nodes. It is multi-relational because different kinds of relations can relate to the same pair of entities. When two (or more) entities are connected through a relation, we can call it a fact [Wang et al. 2017]. KG entities can have arbitrary string literals associated with attributes.

KG can be represented using Semantic Web standards such as RDF[1] and OWL. RDF is, at its core, a collection of triples. Due to its triple nature, RDF is not suitable to direct represent n-ary relations with n $>$ 2. When this is necessary, reification can be applied. Edges are described using three triples with the aid of blank nodes.

When a KG demands to represent tuples with more than three components, other graph data models are needed. If n-ary relations are considered, with n $>$ 2, the edges become hyper edges (connecting more than two nodes). For hyper relational graphs, the edges can be nodes of other edges. Hyper edges allow qualifiers on edges to represent provenance, spatial or temporal information. Graph data models effectively represent a KB, but their underlying symbolic nature usually makes KGs hard to manipulate.

Embeddings

Generally speaking, embeddings techniques convert any symbolic representation (Text, Image, Graph) into low dimensional vectors. For example, word embeddings is a real-valued vector, with dimension space much lower than the corpus vocabulary size, generated to represent each word from this corpus. Vector representations were developed to quantify word semantics, not the exact meaning of the word, but contextual. The primary

[1] `https://www.w3.org/TR/rdf11-concepts/`

intuitions are that the co-occurrence of words in similar contexts indicates that these words are semantically equal. The similarity between words, not in the sense of synonyms, but the proximity in the vector space can be calculated using similarity metrics such as cosine similarity. Semantic Search can use word embedding to capture keyword list semantic and also represent the semantic of the indexed documents [Cudré-Mauroux 2019].

Graph embeddings can be seen as a generalization of word embeddings. The purpose of knowledge graph embeddings is to represent KG components in continuous vector spaces without losing their inherent structure to facilitate computational tasks. Graph embeddings techniques can use only observed facts from the KG or even incorporate additional information. According to [Wang et al. 2017], fact-based methods can be grouped in two broader categories: translation-based, where relations are represented as a translation vector between connected two entities such as TransE [Bordes et al. 2013], and semantic matching models, that generates the embeddings using similarity-based scoring functions such as RESCAL [Nickel et al. 2011].

Such graph representations can simplify graph manipulation for in-KG tasks, such as completion and entity resolution, as well as out-of-KG applications, such as keyword search, question and answering (Q&A), and recommendation systems.

3. Related Work

Microsoft Academic Graph(MAG) [Shen et al. 2015] represents a heterogeneous KG about scholarly communications with six types of entities: scientific publication (Paper), their authors, institutions, journals, conferences, and Field of Study (FoS) Hierarchy. It also contains entity relationships and attributes. Microsoft Academic (MA) [Wang et al. 2019] is a semantic search interface target to query MAG. MA uses entities navigation and keyword search. The keywords use synonyms to refer to the same entity, for example, the acronym (acronym) and the conference name or the author's full name and citation names. The tool also allows term scope operators such as, for example, title: "graph," which will only match the title of publications. Search results ranking is based on salience measure, which is calculated using the eigencentrality concept of Graph Theory. According to [Bast et al. 2016] MA can be classified as a Semi structured Search on Combined Data.

Wikidata (WD) [Vrandecic 2012] is a community-built, multilingual, and general-purpose human and machine-readable KG. Each WD item has a unique identifier, properties (at least a label and description), and one or more aliases (alternative labels) associated with a language. Properties can be defined through statements composed of an item ID, a property ID, and a value. A statement can be annotated with property-value pairs, which can be qualifiers and references composed by a property ID and a value. Due to these characteristics, Wikidata is a multi-graph since the same statement can occur more than once with different qualifiers.

Scholia [Nielsen et al. 2017] is a semantic search interface target to query scholarly communications items from Wikidata. The purpose is to create academic profiles for researchers, organizations, journals, publishers, individual literary works, and research topics based on Wikidata. Scholia search interface by keyword shows matched items based on label, description, or aliases using Wikidata Query Service (WDQS)[2] Entity-

[2]Online available at `https://query.wikidata.org/`

Search, a full-text search API. According to [Bast et al. 2016] Scholia can be classified as Keyword Search on Knowledge Bases (here the KB is Wikidata KG).

A TSA+BM25 and the TSA+VDP keyword search systems based on the *virtual documents* approach and best-match assumption was proposed by [Dosso and Silvello 2020]. The computational complexity to build text documents based on the nodes (subjects and objects) and edges (predicates) from the RDF dataset is made off-line. Text retrieval techniques and data structures such as inverted indexes are used to speed up the search process for the on-line phase and to return a ranking of best answers. Two methods were applied on how the document is ranked and mapped back to answer subgraphs: BM25 and Virtual Document Pruning (VDP).

4. Expected Contributions

This research work proposes a semantic search based on hyper relational KGs. We will use embeddings in order to find the most similar subgraphs that satisfy a user query expressed as a keyword list. This approach aims to simplify graph manipulation as well as to promote meaningful retrieval of KG entity-centric results. A preliminary overview of the proposal, without context and ranking, is depicted in Figure 1.

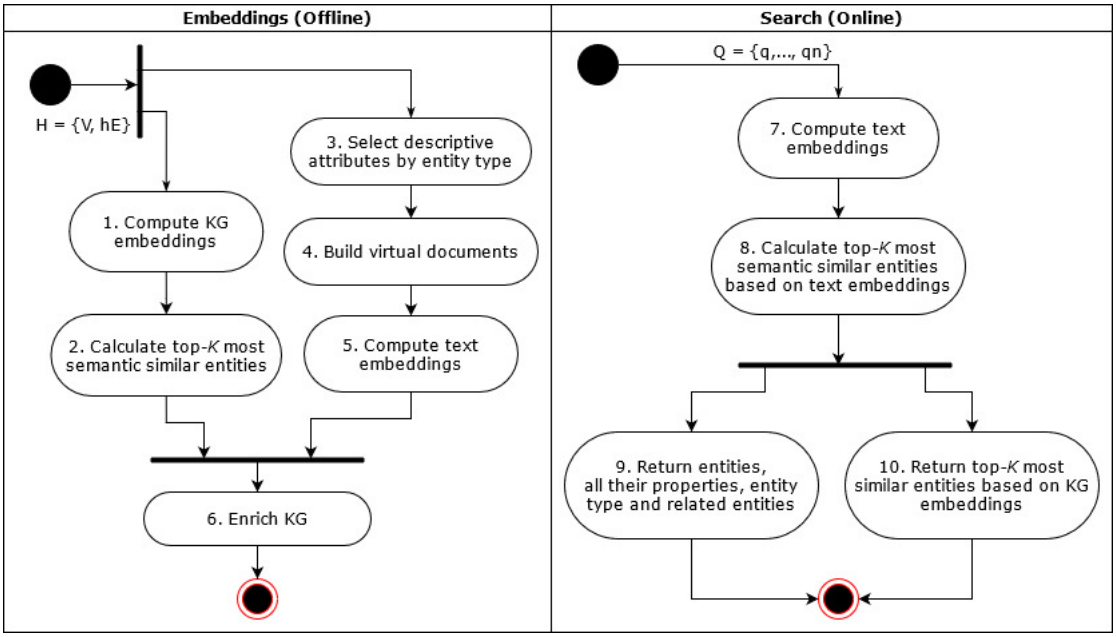

Figure 1. KG-based Semantic Search approach

In the offline phase, given a hyper relational graph $H = \{V, hE\}$ as input: (1) Compute KG embeddings for each entity node of H; (2) Calculate top-K most semantic similar entities nodes of the same entity type based on the KG embeddings; (3) Select descriptive attributes for each entity type that belongs to H; (4) Build virtual documents for each entity node from H based on the values of previously selected attributes; (5) Compute text embeddings of the virtual documents; and finally (6) Enrich the KG H with one simple edge connecting each entity node with its text embeddings plus one hiper edge connecting these node with the K most semantic similar ones.

In the online phase, given a keyword list $Q = \{q1, q2, ..., qn\}$ as input: (7) Compute text embeddings of the keyword list Q; (8) Calculate top-K most semantic similar entities nodes based on the text embeddings previously computed and the query embeddings representation; (9) Return entities nodes, all their properties (attributes and values), entity type and directed related entity nodes (neighbours) ordered by the similarity measure; and also (10) Return top-K most similar entities nodes based on KG embeddings, as calculated in step 2 and stored in the KG H in step 6.

Different from MA[Wang et al. 2019] this approach don't use KG structure to rank the subgraphs, it uses the similarity measure between Q and entity nodes based on their text embeddings. Different from Scholia [Nielsen et al. 2017], text embeddings are used instead of classical IR approaches to match keyword list and entity nodes literals. Similarly to TSA [Dosso and Silvello 2020] this proposal has two phases (offline to compute KG embeddings and online to match and ranking) and also build virtual documents to represent entity nodes but it is target to any hiper relational graphs instead of RDF model only.

5. Methodology

To achieve the previously mentioned contributions, we defined four research objectives:

a Define a similarity measure to answer questions such as *Which individuals of an entity type T are related to subject X (from context C)?*
b Design a solution that, given a keyword list Q, retrieves a set of subgraphs h from H corresponding (or closer) to that query limited by the searcher's context C and ordered by a weighting scheme of the subgraph's properties.
c Implement a flexible solution where the context C is optional and can be defined at search time.
d Evaluate the solution through experiments with hyper relational KGs.

6. Preliminary Results and Current Work Status

Currently, we are experimenting KGTK framework [Ilievski et al. 2020] for embedding computation and, also to convert public available RDF datasets into KGTK format. In parallel, we are building a hyper relational KG about scholarly communications based on CV Lattes of PUC-Rio researches extracted from the CNPq platform.

We have chosen KGTK because its graph data model can represent a hyper relational graph since edges can be nodes of other edges. KGTK model[3] represents an edge as a tuple $(Id, node1, label, node2)$. The Id is a unique identifier for an edge (every edge has a unique identifier). It enables edges to have other edges asserted about them without requiring adding extra triples to represent these edges. KGTK flexibility enables the representation of an arbitrary number of levels of edges on edges.

KGTK is also a complete framework written in Python to facilitate KG manipulation. RDF datasets can be converted into the KGTK model by the NTriples-import command and any extra nodes previously generated by reification can be removed using the unreify-RDF-statements command. KG embeddings can be computed based on the structure of nodes and their relations using ComplEx[Trouillon et al. 2017] (by default),

[3] `https://kgtk.readthedocs.io/en/latest/data_model/`

TransE, DistMult[Yang et al. 2015], or RESCAL. Text embeddings can be computed for nodes based on their properties and labels. A template is used to concatenate selected properties into sentences. These sentences are embedded using at least one of the 16 currently supported variants of three state-of-the-art pre-trained language models: BERT, DistilBERT, and Roberta.

References

Bast, H., Buchhold, B., and Haussmann, E. (2016). Semantic search on text and knowledge bases. *Foundations and Trends in Information Retrieval*, 10(2-3):119–271.

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *NIPS'13*, page 2787–2795. Curran Associates Inc.

Cudré-Mauroux, P. (2019). Semantic search. In Sakr, S. and Zomaya, A. Y., editors, *Encyclopedia of Big Data Technologies*. Springer.

Dosso, D. and Silvello, G. (2020). Search text to retrieve graphs: A scalable rdf keyword-based search system. *IEEE Access*, 8:14089–14111.

Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N. T., Yao, Y., Rogers, C., Li, R., Liu, J., Singh, A., Schwabe, D., and Szekely, P. (2020). Kgtk: A toolkit for large knowledge graph manipulation and analysis. In *ISWC 2020*, pages 278–293. Springer.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *ICML'11*, page 809–816, Madison, WI, USA. Omnipress.

Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2017). Scholia, scientometrics and wikidata. In *The Semantic Web: ESWC 2017 Satellite Events*, pages 237–259. Springer.

Shen, Z., Ma, H., and Wang, K. (2015). A Web-scale system for scientific knowledge exploration. *ACL 2018*, pages 87–92.

Trouillon, T., Dance, C. R., Gaussier, E., Welbl, J., Riedel, S., and Bouchard, G. (2017). Knowledge graph completion via complex tensor factorization. *J. Mach. Learn. Res.*, 18(1):4735–4772.

Vrandecic, D. (2012). Wikidata: a new platform for collaborative data collection. In Mille, A., Gandon, F., Misselis, J., Rabinovich, M., and Staab, S., editors, *WWW' 2012*, pages 1063–1064. ACM.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A., and Rogahn, R. (2019). A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2:45.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE TKDE*, 29(12):2724–2743.

Yang, B., Yih, W., He, X., Gao, J., and Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In Bengio, Y. and LeCun, Y., editors, *ICLR 2015*.