# 6 Leveraging Information Retrieval over Linked Data

Edgard Marx[1,2]

## Abstract

The Semantic Web vision has been realized, and a plethora of data from a multitude of domains is currently openly available on the Web. The semantic technologies have achieved reasonable maturity and spread across different industry segments. To date, over ten thousand knowledge graphs (KGs) have been published under the linked open data (LOD) cloud. In many cases, those KGs are very large, making their exploration and browsing time-consuming while maintaining their access a very resource-demanding task. Studies revealed that only one-third of public linked data (LD) access endpoints have a high availability rate, making them an unreliable option.

In addition, the Semantic Web architectural choices assemble the Web itself, making the LOD cloud a gigantic distributed and decentralized database composed of domain-specific as well as open-domain interlinked KGs. This very nature causes information to be duplicated and dispersed across various KGs. For example, an agent might be interested in verifying or conciliating the political boundary of a place using LinkedGeoData or its social indicators using DBpedia. However, a substantial effort might be necessary to check thousands of potential data sources. As if that were not enough, the information available in the LOD requires users to be familiar with formal query languages and data structures, which input a substantial obstacle to data consumption as well as content access. Simplifying LD search and discovery is important to enabling users to access and explore large amounts of information from multiple and distributed KGs. Within this chapter, we investigate the issues related to information retrieval over LD. We provide methods and evaluations of conceptual approaches that facilitate information access through formal and natural language queries.

The first challenge addressed in this chapter is the lack of studies in selecting relevant fragments of RDF data from distributed KGs. Thus, we present RDFSlice, a novel approach that enables the selection of well-defined slices of datasets fulfilling typical information needs. We show that the proposed approach is much faster and resource-efficient than the conventional methods of loading and retrieving the slices of the whole dataset from a triple store. RDFSlice is flexible and supports the sub-graph selection through basic graph patterns (BGP) for which each connected triple pattern shares

[1] ORCID: https://orcid.org/0000-0002-3111-9405
[2] Hochschule für Technik, Wirtschaft und Kultur, Leipzig

https://doi.org/10.33968/9783966270502-06

a maximum of one variable with one another. It also contains a query execution planner that runs the most optimized routine based on the query.

The second problem addressed in this chapter is the lack of efficient architectures for publishing and consuming RDF data. To date, RDF KGs are mainly consumed over large data files and SPARQL endpoints. However, SPARQL endpoints often have accessibility issues, while large data files are cumbersome. This thesis proposes a new distributed and decentralized publishing architecture that simplifies data sharing and querying by transparently shifting query execution on KGs to the network's edge. The evaluation over traditional publishing methods shows that the proposed architecture is more reliable and efficient regarding query runtime with a cost of data replication across the different network peers. Finally, we give an outlook on the future of RDF sharing and querying.

The third topic addressed in this chapter is the lack of studies in ranking functions for RDF data. For a long time, ranking functions have been used to facilitate information access in a wide range of tasks such as Search, EL, QA, Link Discovery, and Machine Learning to name a few. They often explore the data structure and statistics to measure relevance and take into account the context of the data. Although many ranking functions have been proposed over the last decades, there needed to be more studies of their impact on the Semantic Web domain. To overcome this gap, we create a benchmark for evaluating ranking functions using 60 users from two countries. We evaluate over a dozen ranking functions for RDF data, applied to properties, classes, and entities, and propose two ranking functions, DBtrends and MIXED-RANK.

The fourth subject addressed in this chapter is the need for effective methods for converting natural language utterances to a target KG. Over the last decades, many approaches have been proposed to enable RDF content access through natural language queries. Overall, those systems are commonly penalized concerning their precision due to their reliance on traditional IR bag of words methods. Overcoming this limitation is pivotal to enabling lay users to access information. In this thesis, we propose a scoring function based on Term Networks dubbed as *P (read star path) that allows factual query interpretation using the underlying graph structure of the RDF KGs. We compared *P with different state-of-the-art ER, QA, and EL over standard benchmarks and showed that it achieved better performance in factual based keyword queries. We further evaluate the use of the method in the Triple Scoring evaluation campaign achieving the general fourth place worldwide.

## 6.1 Extracting Relevant subsets of RDF data

In spite of the high availability of data, organizations still encounter an accessibility challenge while consuming Linked Open Data (LOD). RDF datasets are mostly accessible via either SPARQL endpoints or RDF data dumps. Part of these challenges lies on access points. In

an experimental study by Aranda et al.,[3] where 427 public endpoints were examined, the result revealed that around only one third of them have an availability rate of more than 99%; therefore, for accessing data, public endpoints are not a reliable option. Another option, i.e., using dumps of LOD datasets is also problematic. Since many of the LOD datasets are very large, both loading and querying them via a triple store is extremely time-consuming and resource-demanding. For example, DBpedia[4] and LinkedGeoData[5] encompass significantly more than 1 billion triples each. The loading time amounts to approximately 8 hours for DBpedia and 100 hours for LinkedGeoData on standard hardware.

It is possible that organizations, as well as ordinary users, may not be interested in the entire dataset; sometimes, a very specific fragment of these datasets suffices their need. For instance, for a consumer with an interest in entertainment topics, a fragment of DBpedia containing facts about, e.g., movies and actors is adequate. Another example is providing users with points-of-interest information from the LinkedGeoData dataset starting from the users' location. In both scenarios, only a tiny fraction of the underlying knowledge base is sufficient for a particular use case. In the above DBpedia example, all instances from classes `Actor` (2,431 instances) and `Film` (71,715 instances) are the required resources. In case of LinkedGeoData, we can omit all nodes and relations which do not have type `lgdo:PointOfInterest` or any of its sub-classes; thus, 98% of triples can be purged.

Slicing datasets is an emerging concept which enables users to section datasets in order to decrease the time and resources needed. Naturally, this significantly increases query performance since irrelevant but potentially very large parts of a dataset are discarded. As the extracted slices include only the required amount of information, closed-domain Semantic Web applications (i.e., applications with a specific topic) can perform more efficiently.

Figure 6.1 depicts a conceptual model from publishing to consumption of LOD datasets. The process is inspired by the classic ETL process known from Data Warehouse. However, other than ETL, the LOD consumption considers both new dataset versions being published *and* revisions being applied to internally used (parts of) data sets. The steps are described as follows:

1. *Publishing Data* is a prerequisite for the remaining consumption steps and comprises the publication of an RDF data set by a data publisher, mostly through a data set dump or a SPARQL endpoint.

[3] Carlos Buil Aranda et al. "SPARQL Web-Querying Infrastructure: Ready for Action?" In: *The Semantic Web – ISWC 2013 – 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21–25, 2013, Proceedings, Part II*. 2013, pp. 277–293.

[4] Jens Lehmann et al. "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia." In: *Semantic Web Journal* 6.2 (2015), pp. 167–195.

[5] Sören Auer, Jens Lehmann, and Sebastian Hellmann. "LinkedGeoData – Adding a Spatial Dimension to the Web of Data." In: *Proc. of 8th International Semantic Web Conference (ISWC)*. 2009. DOI: doi:10.1007/978-3-642-04930-9_46.

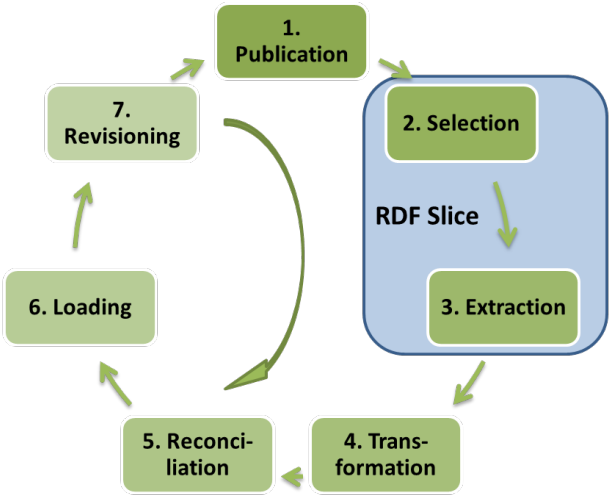


**Figure 6.1:** Linked Open Data consumption process.

2. *Slicing Data* includes the two steps described below:
   - *Selection* comprises the definition and specification of a relevant fragment of a dataset, which is envisioned to be used internally by a consuming organization.
   - *Extraction* processes the dataset dump and extracts the relevant fragment.
3. *Consuming Data* comprises all inter-organizational operations for using data, which briefly can be categorized as follows:
   - *Transformation* comprises mapping of the extracted data structure to match the organization's internal data structures.
   - *Reconciliation* applies revisions made by the organization to earlier versions of the dataset to the actual version.
   - *Loading* makes the dataset available for internal services and applications, for example, by means of a SPARQL endpoint.
   - *Revisioning* allows the organization to apply (manual) changes to the dataset, such as deleting instances or changing properties. Revisions applied to a certain version of the dataset should be persistent and be automatically reapplied (after an update of the dataset) at the respective reconciliation step.

RDFSlice[6] focuses particularly on both the selection and extraction steps. These steps are essential to reduce space and time complexity

[6] Edgard Marx et al. "Towards an Efficient RDF Dataset Slicing." In: *International Journal of Semantic Computing* 7 (2013), p. 455. DOI: 10.1142/S1793351X13400151 and Edgard Marx et al. "Torpedo: Improving the State-of-the-Art RDF Dataset Slicing." In: 2017. DOI: 10.1109/ICSC.2017.79.

in the whole process, since the retrieved fragment is a subset (i.e., a *slice*) of the original dataset. In this chapter, we devise a fragment of SPARQL dubbed SliceSPARQL, which enables the selection of well-defined slices of datasets fulfilling typical information needs. SliceSPARQL supports graph patterns for which each connected subgraph pattern involves a maximum of one variable or IRI in its join conditions. This restriction guarantees the efficient processing of the query against a sequential dataset dump stream. As a result, our evaluation shows that the proposed approach is much faster than using the conventional method of loading and retrieving the slices of the whole dataset from a triple store. Precisely, extracting the relevant fragment from large datasets (e.g. LOD Cloud) in-place is more efficient than downloading, indexing and extracting over triple stores. Although there are many existing approaches for LSD,[7] they are designed for continuous data streaming with high change rate, e.g. once per second. Differently from the SPARQL Streaming approaches, Slicing is not designed for continuous data streaming. Rather, we aim to extract relevant fragments from atomic data streaming, i.e., large files in the distributed static RDF-based LOD. However, the slicing engine can be extended to exploit the temporal order of data in the stream to improve the performance.

## 6.2 A Distributed and Decentralized RDF Publishing Architecture

Since the inception of the Web of Data, many open knowledge graphs were made available in RDF format. Examples of such knowledge graphs are DBpedia,[8] Freebase[9] and Wikidata.[10] Together, these knowledge sources alone encompass more than three billion facts covering a multitude of domains. Despite this data being freely available, lay users, as well as researchers and enterprises, still face difficulties in consuming RDF. The main obstacle is that using the data is still a very cumbersome and resource-demanding task. As a result, users often rely on publicly available SPARQL endpoints and RDF dump files.

On the one hand, SPARQL endpoints are not reliable, as it has been shown that the query evaluation problem for SPARQL

[7] Davide Francesco Barbieri et al. "An Execution Environment for C-SPARQL Queries." In: *Proceedings of the 13th International Conference on Extending Database Technology*. EDBT '10. Lausanne, Switzerland: ACM, 2010, pp. 441–452. DOI: 10.1145/1739041.1739095; Jean-Paul Calbimonte, Oscar Corcho, and Alasdair J. G. Gray. "Enabling ontology-based access to streaming data sources." In: *Proceedings of the 9th International Semantic Web Conference on The Semantic Web – Volume Part I*. ISWC'10. Shanghai, China: Springer-Verlag, 2010, pp. 96–111; Darko Anicic et al. "EP-SPARQL: a unified language for event processing and stream reasoning." In: *Proceedings of the 20th international conference on World wide web*. WWW '11. Hyderabad, India: ACM, 2011, pp. 635–644. DOI: 10.1145/1963405.1963495.

[8] Jens Lehmann et al. (2015).

[9] Kurt Bollacker et al. "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.

[10] Denny Vrandečić and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase." In: *Communications of the ACM* 57.10 (2014), pp. 78–85.

is "*PSPACE*-complete even without filter conditions".[11] Therefore, high-demand services are generally expensive to host, "which makes reliable public SPARQL endpoints an exceptionally difficult challenge".[12] For instance, a study monitoring 427 endpoints for 27 months shows that SPARQL endpoints have "an average fixed HTTP cost of $\sim$ 300 ms per query".[13] Moreover, the mean endpoint availability of the SPARQL endpoints decreased over time (i.e., from 83% in the beginning to 51% at the end of the experiment), while at least 24.3% of the SPARQL endpoints were always down.[14] To tackle the reliability problem of SPARQL endpoints, some proposed approaches reach from (i) improved indexing techniques[15] to (ii) novel architecture patterns such as LDF.[16] However, these methods often impose limitations as, for example, high network bandwidth consumption[17] as well as restrictions on SPARQL features – i.e., some indexes restrict the SPARQL query features to only basic graph patterns.[18]

On the other hand, consuming RDF dump files can be a very cumbersome, time-consuming, and resource-demanding task as there is a high effort necessary for: (1) identifying; (2) downloading and; (3) setting up the infrastructure for RDF data management including indexing the desired portion of the RDF graph. All these obstacles make it difficult for users to query Linked Data and build applications on top of it.

To overcome this challenges Marx et al.[19] proposes the Knowledge Box (KBox), an approach to transparently shift the query execution on knowledge graphs to the user or application (i.e., the edge of the network). KBox is based on a decentralized architecture for publishing and dereferencing RDF Knowledge Graphs that transfers the query execution from the server to the user or application.

## 6.3 Ranking Linked Data

Over the last decades, we have seen an emerging necessity in developing ranking functions in order to facilitate content access. This necessity became evident in the Semantic Web domain with the emerg-

[11] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. "Semantics and Complexity of SPARQL." in: *ACM Trans. Database Syst.* 34.3 (Sept. 2009), 16:1–16:45. DOI: 10.1145/1567274.1567278.

[12] Ruben Verborgh et al. "Querying Datasets on the Web with High Availability." In: *International Semantic Web Conference*. Springer. 2014, pp. 180–196.

[13] Carlos Buil-Aranda et al. "SPARQL Web-Querying Infrastructure: Ready for Action?" In: *International Semantic Web Conference*. Springer. 2013, pp. 277–293.

[14] Carlos Buil-Aranda et al. (2013).

[15] Pingpeng Yuan et al. "TripleBit: A Fast and Compact System for Large Scale RDF Data." In: *Proc. VLDB Endow.* 6.7 (May 2013), pp. 517–528. DOI: 10.14778/2536349.2536352; Javier D. Fernández et al. "Binary RDF Representation for Publication and Exchange (HDT)." in: *Web Semantics: Science, Services and Agents on the World Wide Web* 19 (2013), pp. 22–41. URL: http://www.websemanticsjournal.org/index.php/ps/article/view/328.

[16] Ruben Verborgh et al. (2014).

[17] Ruben Verborgh et al. (2014).

[18] Javier D. Fernández et al. (2013).

[19] Edgard Marx et al. "KBox: Transparently Shifting Query Execution on Knowledge Graphs to the Edge." In: *11th IEEE International Conference on Semantic Computing, 2017, San Diego, CA, USA*. 2017.

ing large structured datasets. Although many of these datasets are freely available, users can not easily consume them. During the last years, many ranking functions were designed to address a specific or broad range of purposes such as entity summarization[20] document retrieval[21] and entity linking[22] among others. This ranking functions usually explores statistics[23] or the structure of the data[24] to measure its relevance. A fundamental principle of the Semantic Web is that the resources represent concepts in the real world. Therefore, there are a huge amount of features and indicators that can be used to measure how important an information is. For example, to measure the relevance of a country to a person or a policy action, one can use the GDP or the HDI. Furthermore, the relevancy is highly tied to the context. For instance, a public policy coordinator can choose to use the HDI in an ascending order to decide welfare policies, while an emigrant can use the same index in descending order to decide where to move. Another important observation is that the relevance can change over time.

Presently, ranking algorithms have started to become more personalized. This means that instead of using only the data structure itself, approaches have begun to use third-party information, e.g., information that cannot be found in the data itself. For instance, one can use the location, language, or previously visited websites and their frequency. That information helps to enhance the rank of the query results.[25]

---

20 Gong Cheng, Thanh Tran, and Yuzhong Qu. "RELIN: Relatedness and Informativeness-based Centrality for Entity Summarization." In: *Proceedings of the 10th International Conference on The Semantic Web – Volume Part I*. ISWC'11. Bonn, Germany: Springer-Verlag, 2011, pp. 114–129.

21 Aidan Hogan, Andreas Harth, and Stefan Decker. "ReConRank: A Scalable Ranking Method for Semantic Web Data with Context." In: *In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*. 2006; Cristiano Rocha, Daniel Schwabe, and Marcus Poggi Aragao. "A Hybrid Approach for Searching in the Semantic Web." In: *Proceedings of the 13th International Conference on World Wide Web*. WWW '04. New York, NY, USA: ACM, 2004, pp. 374–383. DOI: 10.1145/988672.988723; Li Ding et al. "The Semantic Web – ISWC 2005: 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, 2005." In: ed. by Yolanda Gil et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. Chap. Finding and Ranking Knowledge on the Semantic Web, pp. 156–170. DOI: 10.1007/11574620_14.

22 Gong Cheng, Danyun Xu, and Yuzhong Qu. "Summarizing Entity Descriptions for Effective and Efficient Human-centered Entity Linking." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 184–194.

23 Li Ding et al. (2005).

24 Aidan Hogan et al. (2006); Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999–66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/.

25 Steve Lawrence. "Context in web search." In: *IEEE Data Eng. Bull.* 23.3 (2000), pp. 25–32; Nicolaas Matthijs and Filip Radlinski. "Personalizing Web Search Using Long Term Browsing History." In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: ACM, 2011, pp. 25–34. DOI: 10.1145/1935826.1935840.

According to Zhuang and Cucerzan,[26] a *good* method to measure the importance of information is its occurrence in real users query. Hence, query logs are highly useful for ranking information. The central idea of using query logs is that it allows to extraction of the users' interests across time. As users' interests tend to change over time, query logs provide a better idea about resource relevance when compared with other methods that use only graph-based metrics. Thereafter, query logs can also be used to generate a more personalized ranking, e.g., users from different countries may search for different things. Marx et al.[27] proposes an extension of Spearman's Footrule (C. Spearman. "The Proof and Measurement of Association Between Two Things." In: *American Journal of Psychology* 15 (1904), pp. 88–103) to deal with heterogeneous rankings and an extensive evaluation between main property, class, and entity ranking functions in a standard benchmark for measuring RDF ranking functions. He also proposes[28] two ranking functions for RDF data: DBtrends, a ranking function that uses external information to rank resources in the dataset, more precisely, the query logs, and; MIXEDRANK, a ranking function that uses a combination of DBtrends and the best-evaluated ranking function.

## 6.4 Information Retrieval through Factual keyword-queries

Although the use of triple stores leads to direct and efficient access to the data, lay users cannot be expected to make themselves familiar with the underlying formal languages and modeling structures. The use of ER and QA systems can enhance access to the data. However, they often rely on methods adapted from traditional IR, including approaches such as document retrieval and the exploration of triple stores. On one hand, a typical QA approach begins by converting the input query into a syntax tree. Then, it generates and ranks potential answer graphs by relying either on a triple store or document retrieval techniques.[29] On the other hand, a common approach for ER consists of adapting document retrieval engines and their score functions – e.g., term frequency–inverse document frequency (`TF-IDF`)[30] – to

[26] Ziming Zhuang and Silviu Cucerzan. "Re-ranking Search Results Using Query Logs." In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. CIKM '06. Arlington, Virginia, USA: ACM, 2006, pp. 860–861. DOI: 10.1145/1183614.1183767.

[27] Edgard Marx et al. "DBtrends : Publishing and Benchmarking RDF Ranking Functions." In: *2nd International Workshop on Summarizing and Presenting Entities and Ontologies, co-located with the 13th Extended Semantic Web Conference*. 2016.

[28] Edgard Marx et al. "DBtrends: Exploring Query Logs for Ranking RDF Data." In: *Proceedings of the 12th International Conference on Semantic Systems*. SEMANTiCS 2016. Leipzig, Germany: Association for Computing Machinery, 2016, pp. 9–16. DOI: 10.1145/2993318.2993322.

[29] Saeedeh Shekarpour et al. "SINA: Semantic Interpretation of User Queries for Question Answering On Interlinked Data." In: *Journal of Web Semantics* 30 (2015), pp. 39–51.

[30] Karen Spärck Jones. "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of Documentation* 28.1 (1972).

ER.[31] However, document retrieval engines rely on the assumption that the frequency of a term is related to the topic of the document.[32]

Overall, both categories of systems that rely on traditional IR methods are commonly penalized with respect to their precision. The research in the area of search over LD has thus shifted towards developing methods for efficient ER[33] or QA[34] that take the topology of RDF data into consideration. This is due to evidence that supports the idea that better results can be achieved by exploring the graph structure of the RDF knowledge bases. This assumption is derived from linguistics[35] and supported by results in ER[36] and QA.[37] However, these approaches face low accuracy, especially when dealing with a large volume of data. In this work, we address the following research question: *How to increase the accuracy of the current IR scoring functions on RDF knowledge graphs (KGs)?*

While ER engines seek to retrieve the *top-k* most relevant entities associated with the query intent, QA systems seek to retrieve answers from the knowledge graphs. In both cases, there is a need to correctly segment and ultimately annotate the query with the KG resources. Many QA[38] and ER[39] approaches perform this task using an EL. However, solemnly EL approaches do not suffice because to achieve

[31] Gong Cheng and Yuzhong Qu. "Searching Linked Objects with Falcons: Approach, Implementation and Evaluation." In: *Int. J. Semantic Web Inf. Syst.* 5.3 (2009), pp. 49–70. DOI: 10.4018/jswis.2009081903; Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. "Searching Web Data: an Entity Retrieval and High-Performance Indexing Model." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 10 (2012).

[32] Hans Peter Luhn. "A statistical approach to mechanized encoding and searching of literary information." In: *IBM Journal of research and development* 1.4 (1957), pp. 309–317.

[33] Roi Blanco, Peter Mika, and Sebastiano Vigna. "Effective and Efficient Entity Search in RDF Data." In: *Proceedings of the 10th International Conference on The Semantic Web – Volume Part I*. ISWC'11. Springer-Verlag, 2011, pp. 83–97; Roberto De Virgilio and Antonio Maccioni. "Distributed Keyword Search over RDF via MapReduce." In: *The Semantic Web: Trends and Challenges*. Berlin Heidelberg, Germany: Springer, 2014, pp. 208–223. DOI: 10.1007/978-3-319-07443-6_15.

[34] Lei Zhang et al. "Semplore: An IR Approach to Scalable Hybrid Query of Semantic Web Data." In: *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*. ISWC'07/ASWC'07. Busan, Korea: Springer-Verlag, 2007, pp. 652–665.

[35] Ferdinand de Saussure. *Course in General Linguistics*. Translated by Wade Baskin. New York: McGraw-Hill, 1959; Richard A. Hudson. *Language networks: The new word grammar*. Oxford linguistics. Oxford University Press, 2007.

[36] Roi Blanco et al. (2011); Roberto De Virgilio et al. (2014).

[37] Lei Zhang et al. (2007); Saeedeh Shekarpour et al. (2015).

[38] Ricardo Usbeck et al. "HAWK–hybrid question answering using Linked Data." In: *European Semantic Web Conference*. Springer. 2015, pp. 353–368; Mohnish Dubey et al. "AskNow: A Framework for Natural Language Query Formalization in SPARQL." in: *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, 2016*. Springer International Publishing, 2016, pp. 300–316. DOI: 10.1007/978-3-319-34129-3_19.

[39] Edgar Meij, Krisztian Balog, and Daan Odijk. "Entity Linking and Retrieval for Semantic Search." In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. WSDM '14. New York: ACM, 2014, pp. 683–684. DOI: 10.1145/2556195.2556201; Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. "Entity Linking in Queries: Tasks and Evaluation." In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ICTIR '15. New York: ACM, 2015, pp. 171–180. DOI: 10.1145/2808194.2809473.

the final goal of a QA or ER engines there is a need for a method to correctly identify the resources (Entities, Properties, and Objects). As EL approaches[40] rely primarily on document retrieval methods and frameworks. One hypothesis is that a single scoring function can be used to correctly annotate the resources and, consequently, the search results.

For many years, scientists from the diverse fields of cognitive science, such as psychology, neuroscience, philosophy, linguistics and artificial intelligence, have tried to explain and reproduce the human cognition system. While diverse theories have been developed, a commonly shared idea is that knowledge is organized as a network.[41] Hudson[42] goes further and claims that grammar is organized as a network as well. According to Hudson's work, the syntactic structure of a sentence consists of a network of dependencies between single terms. Thus, everything that needs to be said about the syntactic structure of a sentence can be represented in such a network. Hudson explores Saussure's[43] idea that "*language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others.*". He also argues about the psycholinguistic evidence for the use of *spreading activation* in supporting knowledge reasoning. However, according to,[44] the main challenge consists in finding how the activation occurs in mathematical terms.

> *"How exactly does spreading activation work? How does such a crude, unguided process help us to achieve our cognitive goals, rather than leave us drifting aimlessly around our mental networks? It is very unclear exactly how it works in mathematical terms, but the ... hypothesis is that a single formula controls activation throughout the network".*[45]

The intuition is that, since the KG contains a network of terms formed by the label of its resources, entities, properties, and literals can be used to query. Although there is no evidence that the previous works were influenced by Hudson's theory, some of the proposed models[46] follow this assumption.

However, one of the biggest challenges in IR for RDF data lies in evaluating the relatedness between an entity in a KG and the users's intent. Document retrieval engines rely on term frequency weighting functions based on the assumption that the more frequently a term occurs, the more related it is to the topic of the document.[47] While a

[40] Joachim Daiber et al. "Improving Efficiency and Accuracy in Multilingual Entity Extraction." In: *Proceedings of the 9th International Conference on Semantic Systems*. I-SEMANTICS '13. Graz, Austria: ACM, New York, NY, USA, 2013, pp. 121–124. DOI: 10.1145/2506182.2506198; Diego Moussallem et al. "MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach." In: *K-CAP 2017: Knowledge Capture Conference*. ACM. 2017, p. 8.

[41] Daniel Reisburg. *Cognition: Exploring the science of the mind*. New York: Norton, 1997.

[42] Richard A. Hudson (2007)

[43] Ferdinand de Saussure (1959).

[44] Richard A. Hudson (2007)

[45] Richard A. Hudson (2007).

[46] Lei Zhang et al. (2007);Saeedeh Shekarpour et al. (2015).

[47] Hans Peter Luhn (1957).

good retrieval method needs to take frequency into account, it suffers from frequent yet unspecific words such as "the", "a" or "in". Inverse document frequency corrects this by diminishing the weight of words that are frequently occurring in the corpus, leading to the combined term frequency–inverse document frequency[48] to score documents for a query. However, document retrieval approaches are not designed for RDF because the most important feature of RDF is not merely the term occurrence, but the relation of the concepts underlying its graph structure. Entity retrieval on KGs has been a long-studied research topic for many years. Early approaches rely on bag-of-words models[49] that suffers from *unrelatedness*[50] and *verbosity*.[51] They were built under the assumption that the distribution of keywords is proportional to its subject relatedness.[52] This idea contradicts the fact that people can describe things differently. Authors can be more descriptive or verbose than others. Particularly in the case of DBpedia, editors' experience or knowledge can unconsciously influence keyword frequency or even graph connectivity. To address the problem of verbosity, researchers proposed to score keywords normalized by the information (entity) length.[53] Other generation of ER approaches focused on the problem of unrelatedness by employing field retrieval models.[54] Late studies focused on evaluating how to weight fields differently to improve ER accuracy.[55] Nevertheless, field retrieval models are unable to relate query keywords with a specific predicate or object because they are treated as one, a bag-of-(field-words). Recent approaches introduced the use of two-stage techniques employing ER followed by an EL.[56]

---

48 Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of documentation* 28.1 (1972), pp. 11–21.

49 Haofen Wang et al. "Semplore: A Scalable IR Approach to Search the Web of Data." In: *Journal of Web Semantics* 7.3 (Sept. 2009). DOI: 10.1016/j.websem.2009.08.001; Gong Cheng et al. (2009); Renaud Delbru et al. (2012).

50 Nick Craswell, Hugo Zaragoza, and Stephen Robertson. "Microsoft Cambridge at TREC 14: Enterprise Track." In: *TREC*. ed. by Ellen M. Voorhees and Lori P. Buckland. Vol. Special Publication 500-266. National Institute of Standards and Technology (NIST), 2005. URL: http://dblp.uni-trier.de/db/conf/trec/trec2005.html#CraswellZR05.

51 Stephen E Robertson et al. "Okapi at TREC-3." In: *Nist Special Publication Sp* 109 (1995), p. 109.

52 Hans Peter Luhn (1957).

53 Stephen E Robertson et al. (1995).

54 Nick Craswell et al. (2005).

55 Roi Blanco et al. (2011); Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. "Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data." In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. Santiago, Chile: ACM, New York, NY, USA, 2015, pp. 253–262. DOI: 10.1145/2766462.2767756.

56 Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. "Exploiting entity linking in queries for entity retrieval." In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ACM. 2016, pp. 209–218.

To overcome those problems, Marx et al.,[57] propose $*P$, an ER approach to facilitate information access using keyword factual queries in RDF knowledge graphs. Factual queries are those whose intent can be formalized by triple graph patterns. $*P$ addresses the ER problem in the following manner. It relies on a Semantic Weight Model (SWM) that works in threefold. A query triggers an activation function that measures the relatedness of KG resources w.r.t. the query. The resource relatedness values are then spread to their connected entities using a conditionally backward propagation, and, in a latter process, a conditionally forward one. The individual resource relatedness measurement addresses the problem of finding the query's intent. The conditional propagation avoids the over- and the under-estimation of frequent and rare keywords. The next sections describe how the (1) Activation, (2) Conditional Backward Propagation and (3) Conditional Forward Propagation works.

## References


Anicic, Darko et al. "EP-SPARQL: a unified language for event processing and stream reasoning." In: *Proceedings of the 20th international conference on World wide web*. WWW '11. Hyderabad, India: ACM, 2011, pp. 635–644. DOI: 10.1145/1963405.1963495.

Aranda, Carlos Buil et al. "SPARQL Web-Querying Infrastructure: Ready for Action?" In: *The Semantic Web – ISWC 2013 – 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21–25, 2013, Proceedings, Part II*. 2013, pp. 277–293.

Auer, Sören, Jens Lehmann, and Sebastian Hellmann. "LinkedGeoData – Adding a Spatial Dimension to the Web of Data." In: *Proc. of 8th International Semantic Web Conference (ISWC)*. 2009. DOI: doi:10.1007/978-3-642-04930-9_46.

Barbieri, Davide Francesco et al. "An Execution Environment for C-SPARQL Queries." In: *Proceedings of the 13th International Conference on Extending Database Technology*. EDBT '10. Lausanne, Switzerland: ACM, 2010, pp. 441–452. DOI: 10.1145/1739041.1739095.


[57] Edgard Marx et al. "Exploring Term Networks for Semantic Search over RDF Knowledge Graphs." In: *Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*. Ed. by Emmanouel Garoufallou et al. Cham: Springer International Publishing, 2016, pp. 249–261. DOI: 10.1007/978-3-319-49157-8_22; Edgard Marx and Sandro Coelho. "Answering Live Questions from Heterogeneous Data Sources." In: *25th Text Retrieval Conference (TREC 2016), Live QA Track, 15 November 2016, Montgomery County MD, USA*. 2016; Edgard Marx, Tommaso Soru, and André Valdestilhas. "Triple Scoring Using a Hybrid Fact Validation Approach." In: *WSDM Cup 2017: Vandalism Detection and Triple Scoring, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 17)*. ACM, 2017. DOI: 10.1145/3018661.3022762; Edgard Marx, Gustavo Correa Publio, and Thomas Riechert. "CACAO: Conditional Spread Activation for Keyword Factual Query Interpretation." In: *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, 2019, Proceedings 15*. Springer International Publishing. 2019, pp. 256–271; Edgard Marx et al. "SANTé: A Light-Weight End-to-End Semantic Search Framework for RDF Data." In: *The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers 18*. Springer. 2021, pp. 93–97.

Blanco, Roi, Peter Mika, and Sebastiano Vigna. "Effective and Efficient Entity Search in RDF Data." In: *Proceedings of the 10th International Conference on The Semantic Web – Volume Part I*. ISWC'11. Springer-Verlag, 2011, pp. 83–97.

Bollacker, Kurt et al. "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.

Buil-Aranda, Carlos et al. "SPARQL Web-Querying Infrastructure: Ready for Action?" In: *International Semantic Web Conference*. Springer. 2013, pp. 277–293.

Calbimonte, Jean-Paul, Oscar Corcho, and Alasdair J. G. Gray. "Enabling ontology-based access to streaming data sources." In: *Proceedings of the 9th International Semantic Web Conference on The Semantic Web – Volume Part I*. ISWC'10. Shanghai, China: Springer-Verlag, 2010, pp. 96–111.

Cheng, Gong and Yuzhong Qu. "Searching Linked Objects with Falcons: Approach, Implementation and Evaluation." In: *Int. J. Semantic Web Inf. Syst.* 5.3 (2009), pp. 49–70. DOI: 10.4018/jswis.2009081903.

Cheng, Gong, Thanh Tran, and Yuzhong Qu. "RELIN: Relatedness and Informativeness-based Centrality for Entity Summarization." In: *Proceedings of the 10th International Conference on The Semantic Web – Volume Part I*. ISWC'11. Bonn, Germany: Springer-Verlag, 2011, pp. 114–129.

Cheng, Gong, Danyun Xu, and Yuzhong Qu. "Summarizing Entity Descriptions for Effective and Efficient Human-centered Entity Linking." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 184–194.

Craswell, Nick, Hugo Zaragoza, and Stephen Robertson. "Microsoft Cambridge at TREC 14: Enterprise Track." In: *TREC*. Ed. by Ellen M. Voorhees and Lori P. Buckland. Vol. Special Publication 500-266. National Institute of Standards and Technology (NIST), 2005. URL: http://dblp.uni-trier.de/db/conf/trec/trec2005.html#CraswellZR05.

Daiber, Joachim et al. "Improving Efficiency and Accuracy in Multilingual Entity Extraction." In: *Proceedings of the 9th International Conference on Semantic Systems*. I-SEMANTICS '13. Graz, Austria: ACM, New York, NY, USA, 2013, pp. 121–124. DOI: 10.1145/2506182.2506198.

Delbru, Renaud, Stephane Campinas, and Giovanni Tummarello. "Searching Web Data: an Entity Retrieval and High-Performance Indexing Model." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 10 (2012).

Ding, Li et al. "The Semantic Web – ISWC 2005: 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, 2005." In: ed. by Yolanda Gil et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. Chap. Finding and Ranking Knowledge on the Semantic Web, pp. 156–170. DOI: 10.1007/11574620_14.

Dubey, Mohnish et al. "AskNow: A Framework for Natural Language Query Formalization in SPARQL." In: *The Semantic Web. Lat-*

*est Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, 2016*. Springer International Publishing, 2016, pp. 300–316. DOI: 10.1007/978-3-319-34129-3_19.

Fernández, Javier D. et al. "Binary RDF Representation for Publication and Exchange (HDT)." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 19 (2013), pp. 22–41. URL: http://www.websemanticsjournal.org/index.php/ps/article/view/328.

Hasibi, Faegheh, Krisztian Balog, and Svein Erik Bratsberg. "Entity Linking in Queries: Tasks and Evaluation." In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ICTIR '15. New York: ACM, 2015, pp. 171–180. DOI: 10.1145/2808194.2809473.

– "Exploiting entity linking in queries for entity retrieval." In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ACM. 2016, pp. 209–218.

Hogan, Aidan, Andreas Harth, and Stefan Decker. "ReConRank: A Scalable Ranking Method for Semantic Web Data with Context." In: *In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*. 2006.

Hudson, Richard A. *Language networks: The new word grammar*. Oxford linguistics. Oxford University Press, 2007.

Jones, Karen Spärck. "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of Documentation* 28.1 (1972).

Lawrence, Steve. "Context in web search." In: *IEEE Data Eng. Bull.* 23.3 (2000), pp. 25–32.

Lehmann, Jens et al. "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia." In: *Semantic Web Journal* 6.2 (2015), pp. 167–195.

Luhn, Hans Peter. "A statistical approach to mechanized encoding and searching of literary information." In: *IBM Journal of research and development* 1.4 (1957), pp. 309–317.

Marx, Edgard and Sandro Coelho. "Answering Live Questions from Heterogeneous Data Sources." In: *25th Text Retrieval Conference (TREC 2016), Live QA Track, 15 November 2016, Montgomery County MD, USA*. 2016.

Marx, Edgard, Gustavo Correa Publio, and Thomas Riechert. "CACAO: Conditional Spread Activation for Keyword Factual Query Interpretation." In: *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, 2019, Proceedings 15*. Springer International Publishing. 2019, pp. 256–271.

Marx, Edgard, Tommaso Soru, and André Valdestilhas. "Triple Scoring Using a Hybrid Fact Validation Approach." In: *WSDM Cup 2017: Vandalism Detection and Triple Scoring, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 17)*. ACM, 2017. DOI: 10.1145/3018661.3022762.

Marx, Edgard et al. "DBtrends : Publishing and Benchmarking RDF Ranking Functions." In: *2nd International Workshop on Summa-*

*rizing and Presenting Entities and Ontologies, co-located with the 13th Extended Semantic Web Conference*. 2016.

Marx, Edgard et al. "DBtrends: Exploring Query Logs for Ranking RDF Data." In: *Proceedings of the 12th International Conference on Semantic Systems*. SEMANTiCS 2016. Leipzig, Germany: Association for Computing Machinery, 2016, pp. 9–16. DOI: 10.1145/2993318.2993322.

Marx, Edgard et al. "Exploring Term Networks for Semantic Search over RDF Knowledge Graphs." In: *Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*. Ed. by Emmanouel Garoufallou et al. Cham: Springer International Publishing, 2016, pp. 249–261. DOI: 10.1007/978-3-319-49157-8_22.

Marx, Edgard et al. "KBox: Transparently Shifting Query Execution on Knowledge Graphs to the Edge." In: *11th IEEE International Conference on Semantic Computing, 2017, San Diego, CA, USA*. 2017.

Marx, Edgard et al. "SANTé: A Light-Weight End-to-End Semantic Search Framework for RDF Data." In: *The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers 18*. Springer. 2021, pp. 93–97.

Marx, Edgard et al. "Torpedo: Improving the State-of-the-Art RDF Dataset Slicing." In: 2017. DOI: 10.1109/ICSC.2017.79.

Marx, Edgard et al. "Towards an Efficient RDF Dataset Slicing." In: *International Journal of Semantic Computing* 7 (2013), p. 455. DOI: 10.1142/S1793351X13400151.

Matthijs, Nicolaas and Filip Radlinski. "Personalizing Web Search Using Long Term Browsing History." In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: ACM, 2011, pp. 25–34. DOI: 10.1145/1935826.1935840.

Meij, Edgar, Krisztian Balog, and Daan Odijk. "Entity Linking and Retrieval for Semantic Search." In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. WSDM '14. New York: ACM, 2014, pp. 683–684. DOI: 10.1145/2556195.2556201.

Moussallem, Diego et al. "MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach." In: *K-CAP 2017: Knowledge Capture Conference*. ACM. 2017, p. 8.

Page, Lawrence et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999–66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/.

Pérez, Jorge, Marcelo Arenas, and Claudio Gutierrez. "Semantics and Complexity of SPARQL." In: *ACM Trans. Database Syst.* 34.3 (Sept. 2009), 16:1–16:45. DOI: 10.1145/1567274.1567278.

Reisburg, Daniel. *Cognition: Exploring the science of the mind*. New York: Norton, 1997.

Robertson, Stephen E et al. "Okapi at TREC-3." In: *Nist Special Publication Sp* 109 (1995), p. 109.

Rocha, Cristiano, Daniel Schwabe, and Marcus Poggi Aragao. "A Hybrid Approach for Searching in the Semantic Web." In: *Proceed-*

*ings of the 13th International Conference on World Wide Web*. WWW '04. New York, NY, USA: ACM, 2004, pp. 374–383. DOI: 10.1145/988672.988723.

Saussure, Ferdinand de. *Course in General Linguistics*. Translated by Wade Baskin. New York: McGraw-Hill, 1959.

Shekarpour, Saeedeh et al. "SINA: Semantic Interpretation of User Queries for Question Answering On Interlinked Data." In: *Journal of Web Semantics* 30 (2015), pp. 39–51.

Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of documentation* 28.1 (1972), pp. 11–21.

Spearman, C. "The Proof and Measurement of Association Between Two Things." In: *American Journal of Psychology* 15 (1904), pp. 88–103.

Usbeck, Ricardo et al. "HAWK–hybrid question answering using Linked Data." In: *European Semantic Web Conference*. Springer. 2015, pp. 353–368.

Verborgh, Ruben et al. "Querying Datasets on the Web with High Availability." In: *International Semantic Web Conference*. Springer. 2014, pp. 180–196.

Virgilio, Roberto De and Antonio Maccioni. "Distributed Keyword Search over RDF via MapReduce." In: *The Semantic Web: Trends and Challenges*. Berlin Heidelberg, Germany: Springer, 2014, pp. 208–223. DOI: 10.1007/978-3-319-07443-6_15.

Vrandečić, Denny and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase." In: *Communications of the ACM* 57.10 (2014), pp. 78–85.

Wang, Haofen et al. "Semplore: A Scalable IR Approach to Search the Web of Data." In: *Journal of Web Semantics* 7.3 (Sept. 2009). DOI: 10.1016/j.websem.2009.08.001.

Yuan, Pingpeng et al. "TripleBit: A Fast and Compact System for Large Scale RDF Data." In: *Proc. VLDB Endow.* 6.7 (May 2013), pp. 517–528. DOI: 10.14778/2536349.2536352.

Zhang, Lei et al. "Semplore: An IR Approach to Scalable Hybrid Query of Semantic Web Data." In: *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*. ISWC'07/ASWC'07. Busan, Korea: Springer-Verlag, 2007, pp. 652–665.

Zhiltsov, Nikita, Alexander Kotov, and Fedor Nikolaev. "Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data." In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. Santiago, Chile: ACM, New York, NY, USA, 2015, pp. 253–262. DOI: 10.1145/2766462.2767756.

Zhuang, Ziming and Silviu Cucerzan. "Re-ranking Search Results Using Query Logs." In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. CIKM '06. Arlington, Virginia, USA: ACM, 2006, pp. 860–861. DOI: 10.1145/1183614.1183767.