The Human Side of Service Engineering, Vol. 108, 2023, 264–272
https://doi.org/10.54941/ahfe1003128

Fostering Text Mining With Knowledge Graphs: An Approach to Support Business Experts in Defining Domain-Specific Document Sets

Pascal Bratke[1], Roland Zimmermann[1], and Ralph Blum[2]

[1]Technische Hochschule Nürnberg Georg Simon Ohm, D-90489 Nürnberg, Germany
[2]Fraunhofer-Institut für Integrierte Schaltungen IIS, D-91058 Erlangen, Germany

ABSTRACT

Text mining techniques offer an efficient approach to extract information from text-based data, such as online news, for strategic planning tasks like the scenario technique. Data selection and data pre-processing is a crucial step in this process, during which decisive search terms help to sharpen the text corpus for downstream text mining activities. The integration of business knowledge is crucial for the creation of high-quality domain-specific document sets. Manually defining search terms is a time-consuming task for managers. Knowledge graphs accelerate this vital process: Based on a few initial search terms, related domain-specific subgraphs are selected which yield additional search terms that cannot be retrieved via thesauri or mere semantic similarity comparisons. Furthermore, possible biases introduced by subjective expert assessments are avoided and a more objective data selection is realized. The concept is demonstrated in a use case on electric mobility including an online demonstrator (www.digital-scenarios.com).

Keywords: Knowledge graphs, Information retrieval, Search space definition, Text mining, Natural language processing, Scenario technique, Strategic planning

PROBLEMS IN SEARCH SPACE DELIMITATION: BIASES AND SCARCE EXPERT RESOURCES

Using the scenario technique for strategic and tactical planning is a common approach which suffers efficiency due to time-consuming data collection (mainly from text sources) and text analysis (Backhaus et al. 2018). Consequently, methods of automated text analysis are increasingly used for creating foundations for management related tasks like the creation of scenarios (Kölbl et al. 2019). In this context, data retrieval aims to provide domain-specific input documents which aid to focus downstream text analysis on topics relevant for the scenario domain, e.g. in order to ensure fine-tuning through transfer learning to adopt generalized language models to a specific domain. To sharpen the search space in the discovery phase of data mining processes the domain knowledge of experts has been widely used for many years (Anand et al. 1995). Even today, hybrid approaches integrate experts into the foresight process supported by artificial intelligence to improve the quality of data mining results (Bakker and Budde, 2012).

In order to generate domain-specific document sets for text mining algorithms, one possible task of experts could be the definition of hundreds of relevant search terms (Backhaus et al. 2018). Another approach is to let humans mark documents as relevant manually (Kayser and Shala, 2016). Such selection processes which only rely on human inputs can lead to *biased document sets* since they are strongly influenced by the subjective assessments (and knowledge limitations) of the process participants (Mietzner, 2009). The manual definition of hundreds of search terms is not only a potential gateway for biased document sets, but also time-consuming. This relationship is described as a *quality-efficiency trade-off* in machine learning for text processing (Baeza-Yates and Liaghat, 2017). Scarce and expensive expert resources can therefore have a limiting effect on the quality of the domain-specific document set and thus also negatively influence the text mining results.

In this paper, we present an approach to support experts in the creation of domain-specific document sets for downstream text mining that utilizes knowledge graphs. We address the problem of biases and the trade-off between quality and efficiency in text data collection.

Utilizing Knowledge Graphs for Search Space Delimitation

In the field of information retrieval, the importance of knowledge graphs (KG) has increased in recent years (Dietz et al. 2017). A KG allows knowledge to be represented in a machine-processable way. The basic elements of a KG are entities. An entity can represent describable objects such as a *person*, *place*, *technology*, or *company*. Entities can be further characterized with attributes, such as the *number of employees of a company*. A KG links a large number of such entities through relationships, which themselves can have their own attributes, too. For example, a KG can map the relationships between an entity *Electric mobility* as a super category with entities of the subtypes *Electric Car*, *Hybrid Car* and *Hydrogen Car*. The concept of *semantic similarity of entities* allows to select semantically related entities for information retrieval from a KG: Based on a few initially defined entities, further similar, potentially relevant entities can be identified (Hliaoutakis et al. 2006), (Stankovic et al. 2011). Using the example from above, an initial search term *Electric Car* can be used to automatically infer the more general entity *Electric Mobility* via their relationship. From this super category the next relationship to a subordinate category can also be automatically inferred resulting in a new possible search term *Hydrogen Car*. Such similarity inference provides new search criteria that go well beyond a thesaurus-based search: In the example, a hydrogen-based car is certainly not a typical synonym for a battery-based electric car. Employing KGs to infer additional search terms aids to more comprehensively describe the relevant domain for scenario development. We believe that this helps to reduce biases induced by restricted individual knowledge of experts and at the same time increases the efficiency in creating a more comprehensible set of search terms for a domain in question.

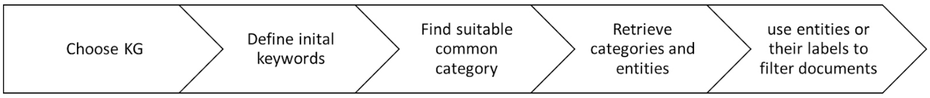

Choose KG
Define inital keywords
Find suitable common category
Retrieve categories and entities
use entities or their labels to filter documents

Figure 1: KG approach to retrieve entities for domain specific document set filtering.

Global knowledge graphs are freely available and contain very many entities. Thus, they offer great potential for semantic similarity searches: For example, DBpedia describes 6,000,000 entities, including 1,500,000 people, 810,000 places, and 270,000 companies (Holze, 2016) or Wikidata that contains over 100.000.000 items (Wikidata, 2023).

The freely accessible global knowledge graph DBpedia was created automatically using Wikipedia as the data basis (Paulheim, 2013). Each entity of the knowledge graph is based on a Wikipedia page. The categorization of pages stored in Wikipedia was also used to classify the entities stored in DBpedia. Thus, entities can be selected based on categories. In addition, the individual categories are linked in the sense of super- and subcategories. Accordingly, thematically narrower categories and their entities, as well as more general categories and their entities can be selected. Relationships between categories are described by *skos:broader* relationships. However, *skos:broader* relationships are not transitive, thus hierarchical inference can only occur between two directly related categories. If there are several hierarchical levels between the categories, such an inference can no longer be made. In order to be able to determine completely hierarchical relationships over several hierarchy levels, the relationship types *skos:broaderTransitive* and *skos:narrowerTransitive* are necessary. However, these relationships are not stored in DBpedia.

Alternatively, other metrics can be used when considering multi-level category relationships. For example, the path length between categories can be used to determine their semantic similarity (Rada et al. 1989). A high correlation with human perception of similarity was found by using distances between DBpedia categories to measure similarity (Senoussi, 2018).

Stankovic et al. used these contexts and, based on DBpedia, identified relevant topics for an open innovation process. For this purpose, initial concepts were first defined on the basis of an innovation problem. Based on these concepts, an algorithm identified further relevant concepts in the next step using *skos:broader* relationships. As a result, relevant topics for the innovation process could be identified (Stankovic et al. 2011).

Extract Domain-Specific Entities From a Knowledge Graph

Our approach to delimit and thus enlarge the search space for documents that serve as input to downstream text-mining activities (which identify and characterize factors relevant to strategic scenario modeling) consists of five steps (Fig. 1).

Choose Knowledge Graph: As indicated, a number of globally available knowledge graphs exist which can be used to identify related keywords. The following section is based on the DBpedia graph.

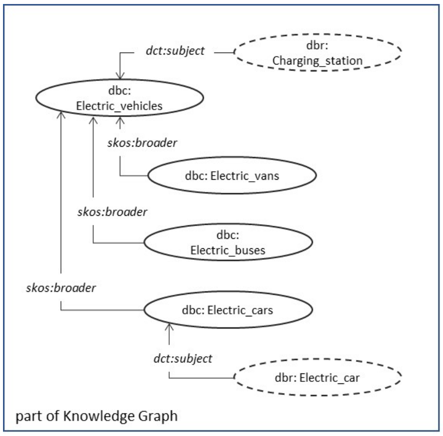
Define initial keywords: An initial search word list L with search words s, which are available, for example, as the result of an expert workshop, is the starting point for narrowing down the search space S. This search word list is used to identify initial entities e_l in DBpedia that represent the search words s. In the example in section 2, the keyword "charging station" can be related to the DBpedia resource "Charging Station".

Find suitable common category: In order to identify further DBpedia entities e, the concept of semantic similarity is used: With the help of the initial list L and the associated categories e_l, the search space S is extended iteratively along *skos:broader* relationships. A maximum permissible distance D between the categories of initial entities e_l and their common category e_c has to be defined. In the example in Figure 2, the common category *dbc:Electric_vehicles* can be identified based on the initial entities *dbr:Charging_station* and *dbr:Electric_car*.

Retrieve categories and entities: Once common DBpedia categories e_c have been identified, it is possible in the next step to find further semantically similar DBpedia categories e_s by following *skos:broader* relationships in the opposite direction: starting from the most common category, narrower categories within the maximum semantic distance are identified via *skos:broader* relations. The entities of all found entities and/or their labels are returned as output.

As a result, a subgraph with reference to the environment is derived from the DBpedia knowledge graph, which represents the enhanced search space S_e. If necessary, this subgraph can be manually curated and further restricted, e.g., to exclude subcategories that are definitely irrelevant for a search environment from an expert's point of view. Furthermore, certain concepts that are not very suitable for environment delimitation, e.g., because they represent people or books, can be automatically excluded from the search space (Stankovic et al. 2011).

Use entities or their labels to filter documents: The entities e_s identified from the KG are used in downstream natural language processing for



Input: Initial list L of search terms l or entities e_l

1. Retrieve categories from the input (entry categories)
2. Find broader categories via *skos:broader* relationships
3. Check if a common category e_c has been found or the maximum semantic distance D has been reached
4. Continue expanding the search space S until a common category e_c is found or the maximum semantic distance D is reached
5. Starting from the common category e_c get narrower categories e_s via *skos:broader* relationships
6. If maximum semantic distance D is not reached, receive iteratively from the found narrower categories their narrower categories via *skos:broader* relationships
7. Get all entities / labels of all found categories to create enhanced search space S_e

Output: enhanced search space S_e with semantically similar DBpedia categories e_s

Figure 2: Identification of similar entities on the topic *electric vehicles* based on two initial entities. *dbc* = DBpedia category; *dbr* = DBpedia resource.

strategic planning to search and filter document sources: Either they can be used as plain keywords creating a list of search terms or in case a larger set of documents has already been pre-analyzed regarding entity types a search based on these entities can be induced.

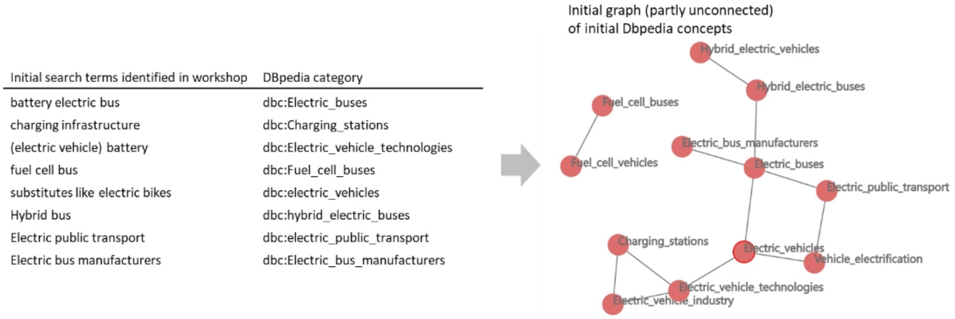
Use Case: Supporting Experts in Building a Document Set for the Electric Mobility Domain

The following case study is based on results of a project of the Fraunhofer IIS Future Engineering group (Fraunhofer, 2023). The aim of the underlying project is to identify factors to support scenario modeling for the topic of e-mobility and to monitor these factors over a longer time frame for an energy provider. The company's management aims to gain insight into the evolving market of electromobility to strategically decide on investments e.g. into charging infrastructure. The documents for analyzing the relevant market domain stem from RSS feeds and scientific publications vaguely related to the topic of electromobility. A detailed selection of texts from this corpus for the scenario development process is based on initial search terms which have been generated in a workshop with experts of the energy provider (see Figure 3). The search terms are mapped to semantically matching DBpedia categories where some of them are directly linked while the concepts *fuel_cell_buses* and *fuel_cell_vehicles* are not directly linked to any of the other concepts (see graph in Figure 3).

Using the semantic similarity approach defined above, more than 100 additional concepts are identified that are of potential interest for the scenario process of the energy provider (see Figure 4 and https://tinyurl.com/Search-Space-Enhancement for an interactive version).

The graph clearly depicts some central, "common" concepts in especially *electric_vehicles* and *electric_rail_transport* that enabled the search algorithm to find relevant subordinate concepts such as *solar_powered_vehicles* or specific variants of *railway_electrification* that enlarged the search space for possibly relevant texts to be retrieved from the text sources.

Based on these newly identified entities, search word lists could have been created. In this project, however, a large document set annotated with



Initial search terms identified in workshop	DBpedia category
battery electric bus	dbc:Electric_buses
charging infrastructure	dbc:Charging_stations
(electric vehicle) battery	dbc:Electric_vehicle_technologies
fuel cell bus	dbc:Fuel_cell_buses
substitutes like electric bikes	dbc:electric_vehicles
Hybrid bus	dbc:hybrid_electric_buses
Electric public transport	dbc:electric_public_transport
Electric bus manufacturers	dbc:Electric_bus_manufacturers

Initial graph (partly unconnected)
of initial Dbpedia concepts
Hybrid_electric_vehicles
Hybrid_electric_buses
Fuel_cell_buses
Electric_bus_manufacturers
Fuel_cell_vehicles
Electric_buses
Electric_public_transport
Charging_stations
Electric_vehicles
Vehicle_electrification
Electric_vehicle_technologies
Electric_vehicle_industry

Figure 3: Initial search terms and their relationships based on mapped DBpedia categories.

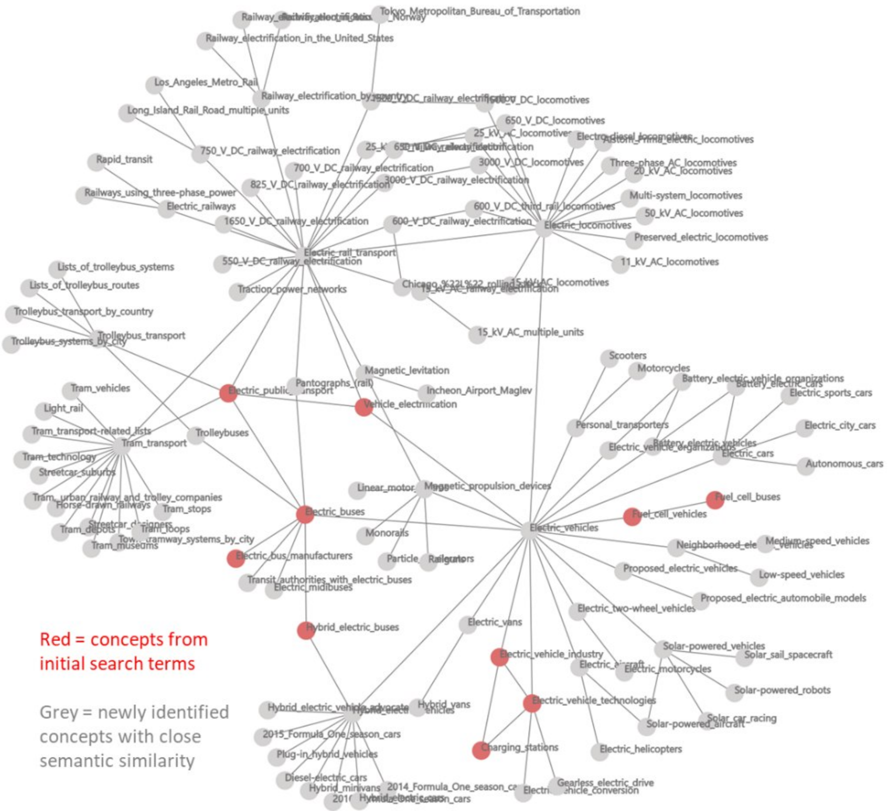


Figure 4: Enlarged search space as a subgraph automatically derived from DBpedia.

DBpedia Spotlight was already available. Thus, the entity definitions were used directly to filter relevant text documents for further processing resulting in a set of more than 18,000 documents for the period from January 2019 to June 2020.

Subsequent processing of these texts consisted of natural language processing tasks based especially on the concept of *Dynamic Topic Modelling* which was used to automatically extract topics of relevance from the source texts (Blei and Lafferty, 2006). Further analysis of the automatically identified topics consisted of calculating indicators such as *media presence* or *temporal intensity* which allow characterization of importance and developments *between* and *within* topics. Qualitative changes over time result in new primary key words in these topics. These changes are identified by calculating *hot* and *cold* words (see Figure 5 bottom left).

Figure 5 additionally illustrates how some newly identified search concepts derived from the KG (e.g. *electric_aircraft* or *tram_transport*) align well with topics that evidently exhibit relevant developments in the analyzed text corpus. Examples are shown with word clouds on *Air taxi* and *Urban transport*. The ten technological topics are subsequently used by the energy provider as potential influencing factors for scenario development. These downstream text mining results can be experienced individually by readers at www.digital-scenarios.com.

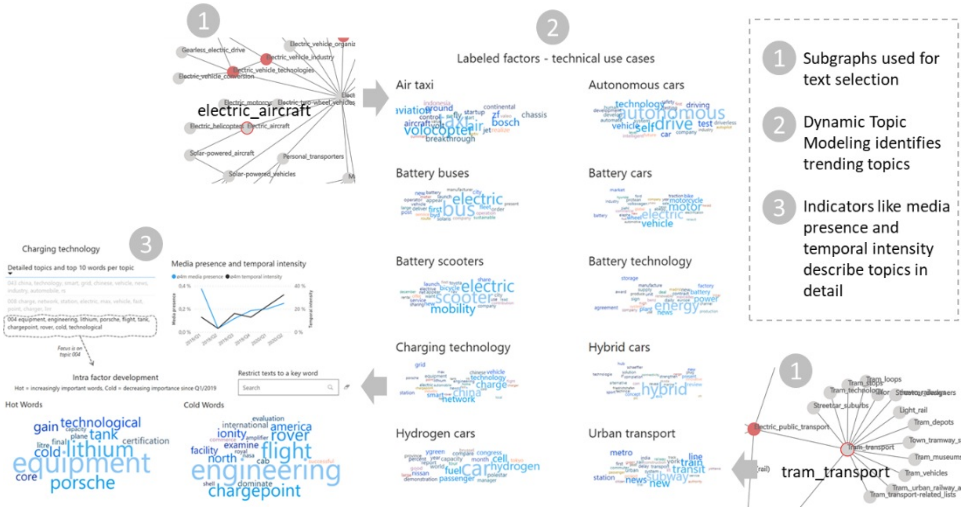

1
electric_aircraft
Gearless_electric_drive
Electric_vehicle_industry
Electric_vehicle_technologies
Electric_vehicle_conversion
Electric_helicopter
Electric_aircraft
Solar-powered_aircraft
Personal_transporters
Solar-powered_vehicles
2
Labeled factors - technical use cases
Air taxi
Autonomous cars
Battery buses
Battery cars
Battery scooters
Battery technology
Charging technology
Hybrid cars
Hydrogen cars
Urban transport
1 Subgraphs used for text selection
2 Dynamic Topic Modeling identifies trending topics
3 Indicators like media presence and temporal intensity describe topics in detail
3
Charging technology
Detailed topics and top 10 words per topic
Media presence and temporal intensity
Focus is on topic 004
Intra factor development
Restrict texts to a key word
Search
Hot Words
Cold Words
gain technological tank cold lithium equipment porsche
international america ionity rover examine north flight engineering chargepoint
volocopter
autonomous
electric bus
electric motor
electric scooter mobility
energy
charge
hybrid
fuel car hydrogen
subway
1
tram_transport
Electric_public_transport
Tram_transport
Light_rail
Tram_depots
Tram_vehicles

Figure 5: Connection between KG-based search term identification and downstream text mining for topic trend-detection.

CONCLUSION

The semantically enhanced search-space delimitation allowed us to identify a large number of search criteria which were not present in the initial set of search strings of the use case. Even with a simple set of initial search terms, environment-specific entities such as range anxiety, vehicle-to-grid concept, or the electric vehicle network as an infrastructure system could be identified. Since there is no standard for defining search strings, the assessment of the quality of additionally identified search criteria remains subjective. Readers are invited to test the sample implementation at www.digital-scenarios.com/search-space-expansion. It allows users to insert their individual key words and then try to find common concepts and semantically similar additional entities. Results are visualized in a graph visualization.

Further exploration is required regarding a viable maximum permissible distance. A higher distance will lower the average relatedness between the entry entity and the identified entity. In our tests we observed that with a maximum permissible distance greater than three the identified entities were partially not similar enough according to human judgment. Other researchers also uncovered this relationship before (Senoussi, 2018).

The procedure for search space expansion described in this paper can be characterized as a basic technical approach. Several advanced approaches for identifying similar concepts from initial search criteria have been described in the literature e.g. (Stankovic et al. 2011) or (Wenige et al. 2019). We therefore assume that the quality of the search space can be further improved by using additional methods for the selection of concepts and by implementing further filter criteria as well as automated cleansing procedures. In addition, it should be examined to what extent other Linked Open Data resources can be used for search space delimitation.

We have addressed that by using knowledge graphs for keyword definition, the bias can be reduced compared to a purely human keyword definition. However, since knowledge graphs such as DBpedia are also affected by human influences, we would like to mention that they are not completely free of bias either (Voit and Paulheim, 2021).

REFERENCES

Anand, S. S., Bell, D. A., Hughes, J. G.: The role of domain knowledge in data mining. In: Pissinou, N., Silberschatz, A., Park, E. K., Makki, K., Nicholas, C. (eds.) Proceedings of the fourth international conference on Information and knowledge management - CIKM '95, pp. 37–43. ACM Press, New York, USA (1995).

Backhaus, K., Paulsen, M.: Szenarioanalyse - Verbesserungen aus der Müsteraner Effizienzwerkstatt. In: Jürgen Gausemeier, Wilhelm Bauer, Roman Dumitrescu (ed.) Vorausschau und Technologieplanung. 14. Symposium für Vorausschau und Technologieplanung, pp. 261–285 (2018).

Baeza-Yates, R., Liaghat, Z.: Quality-Efficiency Trade-offs in Machine Learning for Text Processing. arXiv (2017).

Bakker, S., Budde, B.: Technological hype and disappointment: lessons from the hydrogen and fuel cell case. Technology Analysis & Strategic Management, vol. 24, 549–563 (2012).

Blei, D. M., Lafferty, J. D.: Dynamic topic models. In: Cohen, W., Moore, A. (eds.) Proceedings of the 23rd international conference on Machine learning - ICML'06, pp. 113–120. ACM Press, New York, USA (2006).

Dietz, L., Kotov, A., Meij, E.: Utilizing Knowledge Graphs in Text-centricInformation Retrieval. In: Rijke, M. de, Shokouhi, M., Tomkins, A., Zhang, M. (eds.) Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM '17, pp. 815–816. ACM Press, New York, USA (2017).

Fraunhofer (2023). Fraunhofer Future Engineering Group - Website: https://www.th-nuernberg.de/en/facilities/fraunhofer-research-groups/future-engineering/

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G., Milios, E.: Information Retrieval by Semantic Similarity. International Journal on Semantic Web and Information Systems (IJSWIS), vol. 2, 55–73 (2006).

Holze, J.: YEAH! We did it again ;) – New 2016–04 DBpedia release (2016), https://blog.dbpedia.org/2016/10/19/yeah-we-did-it-again-new-2016-04-dbpedia-release/

Kayser, V., Shala, E.: Generating Futures from Text—Scenario Development Using Text Mining. In: Daim, T. U., Chiavetta, D., Porter, A. L., Saritas, O. (eds.) Anticipating Future Innovation Pathways Through Large Data Analysis. Innovation, Technology, and Knowledge Management, vol. 3, pp. 229–245. Springer International Publishing, Cham (2016).

Kölbl, L., Mühlroth, C., Wiser, F., Grottke, M., Durst, C.: Big Data im Innovationsmanagement: Wie Machine Learning die Suche nach Trends und Technologien revolutioniert. HMD, vol. 56, 900–913 (2019).

Mietzner, D.: Strategische Vorausschau und Szenarioanalysen. Methodenevaluation und neue Ansätze. Gabler Verlag / GWV Fachverlage GmbH Wiesbaden, Wiesbaden (2009).

Paulheim, B.: Type Inference on Noisy RDF Data. In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J. X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) The Semantic Web - ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, p. 516. Springer Berlin Heidelberg (2013).

Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Trans. Syst., Man, Cybern., vol. 19, 17–30 (1989).

Senoussi, H.: Using DBpedia Categories to Evaluate and Explain Similarity in Linked Open Data. In: Fred, A., Filipe, J. (eds.) Proceedings. Volume 1, KDIR, pp. 117–127. Science and Technology Publications, [S. l.] (2018).

Stankovic, M., Breitfuss, W., Laublet, P.: Discovering Relevant Topics Using DBPedia: Providing Non-obvious Recommendations. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 219–222. IEEE (2011).

Voit, M. M., Paulheim, H.: Bias in Knowledge Graphs -- an Empirical Study with Movie Recommendation and Different Language Editions of DBpedia. arXiv (2021).

Wenige, L., Berger, G., Ruhland, J.: SKOS-Based Concept Expansion for LOD-Enabled Recommender Systems. In: Garoufallou, E., Sartori, F., Siatri, R., Zervas, M. (eds.) Metadata and Semantic Research. Communications in Computer and Information Science, vol. 846, pp. 101–112. Springer International Publishing, Cham (2019).

Wikidata (2023), https://www.wikidata.org/wiki/Wikidata:Statistics.