A comparison between Entity-Centric Knowledge Base and Knowledge Graph to Represent Semantic Relationships for Searching as Learning Situations

Marcelo Tibau[1], **Sean W. M. Siqueira**[1], **Bernardo P. Nunes**[2]

[1]Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro, RJ, Brasil

[2]Australian National University – Canberra, Australia

{marcelo.tibau, sean}@uniriotec.br, bnunes@inf.puc-rio.br

Abstract. *Searching the web with learning intent, known as Searching as Learning (SaL), consists on learners to use Web search engines as a technology to drive their learning process. However, it may be difficult to users to find out relevant information online due to an inability to accurately specify their information need, a situation known as Anomalous State of Knowledge (ASK). To minimize the ASK situation, the continuous flow of data gathering and interaction between user and the search results could be used by search engines to tailor learning-intent search experience. It requires Web search engines to identify such intent and they may use linked data, Knowledge Bases and Graph Databases in order to recognize the meaning of query terms and keywords and use them to predict learning intent. In order to explore the possibility of semantic data structures to represent knowledge that could aid a learning-driven Web search engine to recognize learning intention from user's queries, the present paper compared the performance of two different types of data structures based on entity-centric indexing to identify properties and semantic relationships. One was a knowledge base that used a entity-centric mapping of Wikipedia categories and the other was the KBpedia Knowledge Graph. The entity ranking and linking of both were analyzed and we discovered that the knowledge graph could identify about three times more properties and relationships.*

1. Introduction

Web search engine's use could increase in scope from information retrieval tools and techniques to learning technology. For this reason, it appears as an important focus of attention to those who study the nature of knowledge, justification, and the rationality of belief (epistemologists) and those who seek to understand and engineer the Web (scholars, scientists, computer developers, AI practitioners, etc.) [Smart and Shadbolt 2018]. Searching the web with learning intent falls into exploratory search category type. Those searches are motivated by complex information problems and accompanied by misunderstandings about terminology and information structure [White and Roth 2009]. A common feature in an exploratory search scenario is that users usually do not have enough previous information to help them define a structured query [White and Roth 2009] [Marchionini 1995] [Vakkari 2001], a situation know as ASK[1] [Belkin et al. 1982]. Learning from search

[1]ASK, Anomalous State of Knowledge.

DOI: 10.5753/cbie.wcbie.2019.823

tasks, a research area known as Searching as Learning, involves user behaviors not often mapped in Information Retrieval research, such as the abilities of evaluating the usefulness of information critically, differentiating resources, monitoring results, tracking information, prioritizing actions and applying sense-making [Rieh et al. 2016].

The user behavior prompted by ASK indicates a strong influence of significant learning theory [Ausubel et al. 1978] concerning the learning process involved. Significant learning theory considers a person's existing cognitive structure, such as organization, stability, and clarity of knowledge about a specific subject, as being the main stimulus factor for learning and retention of new concepts. The reasons why this study perceives it as a strong influence concerns the user's lack of enough information to structure a proper query, particularly at the beginning of a search session. Lack of information is in itself a major challenge to any search, but users' own searching skills and previous knowledge about the subject searched can play a significant role in narrowing this gap and facilitate a continuous flow of data gathering that can lead to learning. In this sense, users' behavior and the tasks they choose to perform are of great importance to achieve a successful outcome, that is why user's previous search experiences and background are pivotal to understand search intentions and information acquisition patterns. Under these circumstances, exploratory searches can be considered a Knowledge-Intensive Process (KiP) [Tibau et al. 2018] and when applied to Search as Learning context, could be seen within the perspectives of either "searching to learn" and "learning to search" [Rieh et al. 2016] [Vakkari 2016].

As more and more linked data are being integrated and semantics-based search grows, understand and explore the concept, structure and methods to create semantic relationships between concepts indexed by Web search engines is crucial to provide intent recognition capability to them that could be applied to recognize learning intent drawn from user's behavior. In this sense, the paper intends to explore the possibility of semantic data structures to represent knowledge that could aid a learning-driven Web search engine to recognize learning intention from user's queries. It was compared the performance of two different types of data structures based on entity-centric indexing to identify properties and semantic relationships: an entity-centric knowledge base, modeled with the Wikipedia category graph and KBpedia Knowledge Graph, which combined six different knowledge bases into an integrated graph database. The focus of the comparison was the former American President Barack Obama's website[2].

This paper has five more sections besides this introduction. The second section approaches the theoretical background on behavioral and learning theories that grounds this work as a scientific endeavour. The third section presents Entity-Centric Ranking as a path to improve search results. The forth section concerns the Knowledge Graph's usage for semantics purposes. The fifth section compares both approaches and shows the results. Finally, the sixth section concludes the paper with some final remarks.

2. Theoretical Background

The Anomalous State of Knowledge (ASK), when applied in a Searching as Learning context, can be stated as: people has difficulties to find out relevant information online for learning purposes because they cannot accurately specify their information need. ASK is

[2]https://barackobama.com/

part of the Theories of Information Behavior[3], a series of conceptual frameworks used to understand how people search, manage, share, and use information in different contexts.

Vygotsky's zone of proximal development (ZPD), often defined as the difference between what a learner can do with and without help, also provides theoretical background to this work. The ZPD notion were used by Vygotsky in three different contexts [Kozulin et al. 2003], thus it is necessary to explain which was considered. The first is the developmental context, in which ZPD is commonly used for explaining the learner's emerging psychological functions. The second regards an applied context, in which the difference between the learner's individual and aided performances are explained and assessed (much like the definition used at this paragraph's opening). The third is the metaphoric space context, where the everyday concepts used by the learner meet the formalized concepts provided by teachers, mentors or other forms of learning mediators.

Since this work deals with forms to represent concepts and relationships to machines in order to use them as tools to support people's learning processes, our approach tends to blend Vygotsky's second and third ZPD contexts. Considering the ASK situation as an additional component in this curious amalgam, Web search engines provide access to flows concerning both technical formalized and middle-of-the-road contents. Supply them with capabilities to discern useful learning materials is a necessary step of acting as an intermediary. When used as a tool to support learning, Web search engines would need to recognize the learning intent from user's behavior. Indexing Web data in general by the semantic relationships between the concepts presented in the information retrieved seems a reasonable approach to do so.

3. Entity-Centric Ranking to Improve Search Results

Knowledge Base, KB, is the term used to define technologies that stores complex structured, semi-structured and unstructured information used by computer systems. KB serves to represent world facts and is the basis for applying logical rules to deduce new facts and indicate inconsistencies [Hayes-Roth et al. 1983]. A system that processes natural language needs large amounts of represented knowledge to achieve high-level performance [Gesmundo and Hall 2014]. KB fits well in this task because it integrates a large number of entities and allow us to build relationships between these entities. The development of several KBs, including academic projects such as YAGO, NELL, DBpedia and Elementary/Deep-Dive, and private initiatives like those from Microsoft, Google, Facebook and Walmart [Dong et al. 2014], provided the necessary opportunity for semantics. They supply repositories of knowledge capable of storing world facts, including information about people, places, and things [Hayes-Roth et al. 1983].

Entity-centric data management is an area that has received increasing attention as a research field and encompasses a number of disciplines such as Databases, Information Retrieval, and the Semantic Web. Entity ranking has played a key role in information retrieval and used for such tasks as expert finding, where the goal is to find people who have expert knowledge about a particular topic. In recent years, works have looked at how to search for multiple entity types [Tonon et al. 2016]. Standard information retrieval methods based on inverted indices are associated to structured search based on

[3]Also known as TIB.

graphs to connect entities and improve search effectiveness [Tonon et al. 2016]. In Natural Language Processing, entity ranking is used to build up clusters of mentions relying on partially formed clusters produced to make decisions regarding relationships. Then these clusters of mentions are merged if the ranking model predicts they are representing the same entity. In this way, the approach can successfully reject linking [Hillary Clinton] with [Clinton, he] because of the low score between the pair [Hillary Clinton, he], as mentioned in [Clark and Manning 2015].

Another approach applied to Web search engines ranks entity-centric collections in order of relevance to a query, and then identify a set of collections that are likely to contain most relevant entities. It came to be known as Entity-Centric Collection Ranking (EC), in which the central broker, according to their probability of relevance, ranks entities. The top relevant entities contribute to the collection's query-likelihood score [Balog et al. 2012]. Entity-centric ranking is also associated to knowledge bases modeled as graphs to use their representations of entities, along with their related types, to rank the types assigned to entities from the hierarchy created by the graph. In [Tonon et al. 2016], this method was used to select the right granularity of types from the background type hierarchy. Usually, search effectiveness is achieved by combining approaches, especially those combining various ranked lists. The number of entities used for the graph-based search step influences search effectiveness, as seen in [Bron et al. 2013], in which they employed a linear combination of the normalized similarity scores of the text for this purpose.

It was decided to model a specific knowledge base in the form of a graph database. MediaWiki action API was used to capture Wikipedia's data from `Barack_Obama` category - the string can be viewed at Figure 1 - which was saved on a json file - Figure 2. Then, the resource Neo4j Sandbox[4] was deployed to build the datamodel from the loaded data.

Figure 1. String to capture the data from Wikipedia API

56 properties was set and the experiment created 3,080 relationships. The first iteration of the loader, after creating the top-level category, instantiated 25 new categories – Figure 3.

It was observed an entity-centric mapping and an entity ranking, as the category structure of Wikipedia by default identify relevant entity types, looking at query expansion techniques by means of synonyms and other related words, and building on top of the link structure connecting Wikipedia pages to identify alternative entity labels in the anchor text of hyperlinks. Note in Figure 4 that variations and misspellings did became different categories and appeared as nodes in the graph.

[4]https://neo4j.com/sandbox-v2/

Figure 2. Json file with the captured data

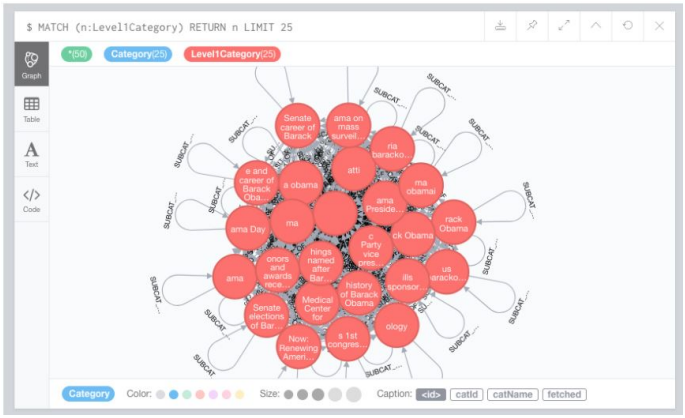

Figure 3. Graph with the instantiated categories

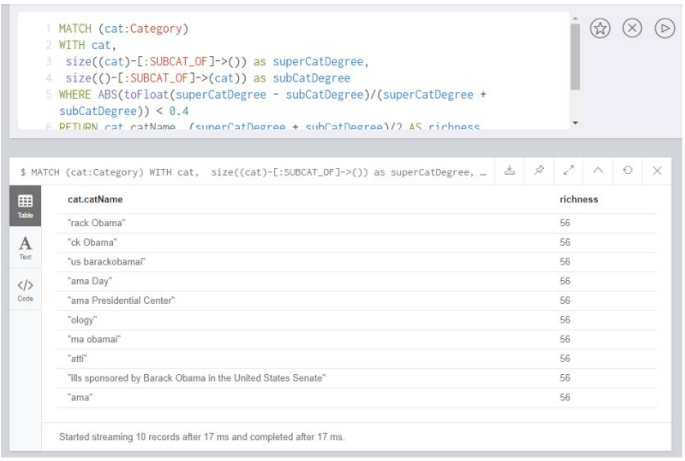

Figure 4. Variations and misspellings became categories

4. Knowledge Graph and its Usage for Semantics

Graph databases use graph structures to represent direct relationships between data items and to create semantic relationships between them. Its main concept is the use of vertices, nodes and edges to model relations between objects, as described in Graph Theory.

Knowledge Graph uses several KBs and models them in the form of a graph database. Google popularized this term in 2012 when it appeared in a text posted on the company's blog[5]. Since then, Knowledge Graphs have been the focus of studies that use graph KB. Applied to search engines, they are used to model knowledge domains aided by data interlinking, subject-topic experts, and Machine Learning algorithms, in particular Learning to Rank algorithms[6]. When Knowledge Graphs are used in Natural Language Processing, they can represent words and phrases and are associated to Machine Learning techniques, such as Skip-gram, to create models that act as semantic algorithms [Mikolov et al. 2013]. The box seen in Google's first page when someone searches for a celebrity, such as the 44th U.S. President, is an example from its implementation.

What connects all these initiatives is the search for semantic integration, made possible by the incorporation of different datasets, commonly known as linked data, and their merging into products and applications based on Artificial Intelligence. The drive behind this specific work is the understanding that the use of Knowledge Graph can improve informational searches[7] results. By working with entities and their relationships, Knowledge Graph structures its information in triples made of subject–predicate–object statements, allowing the query parser to "understand" what is being asked. Syntactic Parsing [Gesmundo and Hall 2014], Knowledge Vault development [Dong et al. 2014], and keyword search improvement [Shan et al. 2017] are examples of its use to provide semantic capability that also could be used to improve search results. Although beyond the scope of the current study, it is our understanding that Knowledge Graphs association to the concept of Search as Learning can significantly boost the use of the Internet, and all the knowledge within, as a huge knowledge flow for learning purposes.

In a Knowledge Graph, entities are nodes, categories are labels associated with each node, and relationships are directed edges between the nodes. One of the methods used to create Knowledge Graph is entity extraction, often done with the help of an information extraction system, such as Never-Ending Language Learner (NELL) [Carlson et al. 2010]. One of the problems derived by this approach is that many textual references which initially seem different actually refer to the same entity, variations and misspellings become different nodes in a graph. Consequently, leading to a need for semantic integration, such as entity resolution, to determine co-referent entities in the Knowledge Graph and to produce a consistent set of labels and relations for each resolved node [Pujara et al. 2013].

The chosen Knowledge Graph was KBpedia's, which itself combines six public knowledge bases –Wikipedia, Wikidata, GeoNames, OpenCyc, DBpedia and UMBEL – and their concepts, entity types, attributes and relations, as stated in their website[8]. An online access to the Knowledge Graph is available, which can be used via an application. The application extracts and analyzes metadata from the webpage and tags concepts and

[5]Singhal, Amit (May 16, 2012). "Introducing the Knowledge Graph: Things, Not Strings". Official Blog of Google. Retrieved November 18, 2017 at http://googleblog.blogspot.com.br/2012/05/introducing-knowledge-graph-things-not.html.

[6]Also known as LTR.

[7]Information-oriented searches are one of the three distinct search categories (transactional, navigational, and informational), and corresponds to the type of search in which the user is looking for certain information. Exploratory searches are one of the informational search sub-types.

[8]http://kbpedia.com

entities based on its combined knowledge bases. Then scores the metadata topics and compares the related topics to provide semantic relationships. As next step, it builds the network and hierarchical layouts – as seen in Figure 5 – as well as the graph. It also allows to download the result as structured data, such as a json file.

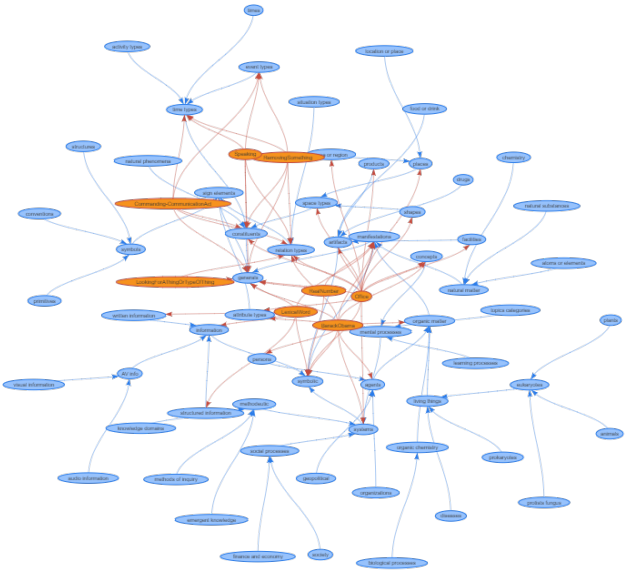

Figure 5. Network layout from KBpedia application

It was observed that the graph structure performed an entity embedding to convey semantics to related entities and entity linking for disambiguation purposes. Again, 56 properties were set, but 9,376 relationships were created – about three times as much as the entity-centric knowledge base.

5. Comparison between Entity-Centric KB and Knowledge Graph

Users with few, or no familiarity at all, with a specific domain face an overwhelming number of challenges. From defining useful terms to formulating queries to explore the options of paths presented by the retrieved data, learning through search is not an effortless task. Proposing ways to aid users expand their domain knowledge while performing searches with learning intent should be a desired result from studies exploring Searching as Learning situations. Exploration paths, data visualization interfaces and semantic-based Information Systems (especially browsers and search engines) [Al-Tawil et al. 2019] are some examples of using linked data to aid minimize the users' burden.

The way linked data is used to promote domain understanding and knowledge expansion is pivotal to a successful approach on the ASK problem. Therefore, it is necessary to determine the proper criterion to perform the comparison. Suitability for capturing data from heterogeneous sources and grasp their relationships is the principle by which we judged our choice. Data enriched with contextual information is demonstrated by the number of properties and relationships created (a property called item context) [Krötzsch 2017]. Given the number of properties and relationships involved, 56/3,080

(Knowledge Base) and 56/9,376 (Knowledge Graph), and based only on these quantity criteria, Knowledge Graphs represent semantic relationships between concepts better than entity-centric knowledge bases.

Knowledge Graphs' capability to connect several KBs in a meaningful way adds to the perception of their suitability to better represent semantic relationships between concepts for this particular context (SaL situations). The fact that Knowledge Graphs explicit meaning through contextual information and relationships between units is strategical, in our view, to help users gain context within the existing data and information provided by the search engine.

6. Concluding Remarks

Searching as Learning research agenda broads the perspective of Web search engines by considering them as possible learning tools. Providing semantics capability could allow a SaL Web searching engine to recognize learning intent and, in a further step, being able to assess if learning occurred from a search. Content type (documents not related to education) and content depth (superficiality or incompleteness) play a pivotal role to organize useful search results to users. The way Knowledge Graphs bring together structured, semi-structured and unstructured information can be used to drive the user experience across his/her learning process by understanding and indexing the content accordingly to the learning necessity. Knowledge Graphs may be the right tool to provide semantic understanding and be helpful to the quest of provide metrics to assess whether the user learned from the retrieved information as well as minimize users' Anomalous State of Knowledge.

A question worth asking is where do we go from here? Some future work paths envisioned by our study give special importance to tailor users' exploration. Using the connections between entities for personalised recommendations; estimating the similarity between users by mapping their search behavior on suchlike domains; and generating exploration paths are all ways to tailoring learning-intent search experience with graph-based approaches.

Limitations to this work include a lack of quality criteria to support the decision made based on the quantity criteria and the choice of not iterate further after creating the top-level category of the entity-centric knowledge-base, subsequently avoiding instantiating its subcategories and producing a next level of node. It could have produced more relationships, although not sufficient to topple the Knowledge Graph position, in our opinion. Also the lack of a dispersion analysis of the entity-centric knowledge-base and the Knowledge Graph's relationships provides a limitation. Relationships more disperse (meaning apart from each other) could indicate weaker links between concepts.

7. Acknowledgment

This study was financed in part by the 'National Council for Scientific and Technological Development (CNPq) - Brazil' - Process 315374/2018-7, Project 'Searching as Learning: the information search as a tool for learning' and by the 'Coordination for the Improvement of Higher Education Personnel' (CAPES) – Brazil – Finance Code 001.

References

Al-Tawil, M., Dimitrova, V., and Thakker, D. (2019). Using knowledge anchors to facilitate user exploration of data graphs. *Semantic Web Journal - Interoperability, Usability, Applicability.*

Ausubel, D., Novak, J., and Hanesian, H. (1978). *Educational Psychology: A Cognitive View*. Holt, Rinehart & Winston, 2nd edition.

Balog, K., Neumayer, R., and Nørvag, K. (2012). Collection ranking and selection for federated entity search. In et al., L. C.-B., editor, *SPIRE 2012*, pages 73–85. LNCS 7608.

Belkin, N., Oddy, R., and Brooks, H. (1982). Ask for information retrieval: Part i. background and theory. In *Journal of Documentation, Vol. 38 Issue: 2*, pages 61–71.

Bron, M., Balog, K., and de Rijke, M. (2013). Example based entity search in the web of data. In Serdyukov, P., Braslavski, P., Kuznetsov, S. O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., and Yilmaz, E., editors, *Advances in Information Retrieval*, pages 392–403, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, Jr., E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1306–1313. AAAI Press.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA*, pages 601–610.

Gesmundo, A. and Hall, K. (2014). Projecting the knowledge graph to syntactic parsing. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*.

Hayes-Roth, F., Waterman, D. A., and Lenat, D. B. (1983). *Building Expert Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Kozulin, A., Gindis, B., Ageyev, V. S., and Miller, S. M. (2003). *Introduction: Sociocultural Theory and Education: Students, Teachers, and Knowledge*, page 1–12. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press.

Krötzsch, M. (2017). Ontologies for knowledge graphs? In Artale, A., Glimm, B., and Kontchakov, R., editors, *Proceedings of the 30th International Workshop on Description Logics (DL 2017)*, volume 1879 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Marchionini, G. (1995). Foundations for personal information infrastructures: information-seeking knowledge, skills, and attitudes. In *in Information Seeking in*

Electronic Environments (Ch.4), pages 61–75. New York NY: Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Pujara, J., Miao, H., Getoor, L., and Cohen, W. (2013). Knowledge graph identification. In Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J. X., Aroyo, L., Noy, N., Welty, C., and Janowicz, K., editors, *The Semantic Web – ISWC 2013*, pages 542–557, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rieh, S., Collins-Thompson, K., Hansen, P., and Lee, H. (2016). Towards searching as a learning process: A review of current perspectives and future directions. In *Journal of Information Science, 42(1)*, pages 19–34.

Shan, Y., Li, M., and Chen, Y. (2017). Constructing target-aware results for keyword search on knowledge graphs. *Data Knowl. Eng.*, 110:1–23.

Smart, P. and Shadbolt, N. (2018). The world wide web. In Coady, J. C. &. D., editor, *Routledge Handbook of Applied Epistemology*. Routledge, New York, New York, USA. 1 edition.

Tibau, M., Siqueira, S., Nunes, B. P., Bortoluzzi, M., and Marenzi, I. (2018). Modeling exploratory search as a knowledge-intensive process. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pages 34–38.

Tonon, A., Catasta, M., Prokofyev, R., Demartini, G., Aberer, K., and Cudré-Mauroux, P. (2016). Contextualized ranking of entity types based on knowledge graphs. *Web Semant.*, 37(C):170–183.

Vakkari, P. (2001). A theory of the task-based information retrieval process. In *J. Doc. 57*, pages 44–60.

Vakkari, P. (2016). Searching as learning: A systematization based on literature. In *Journal of Information Science, 42(1)*, pages 7–18.

White, R. and Roth, R. (2009). Exploratory search: beyond the query-response paradigm. Synthesis lectures on information concepts, retrieval, and services, 1(1): 1-98. Print.