



KEYSTONE WG2: Activities and Results Overview on Keyword Search

Julian Szymański¹ and Elena Demidova²

¹ Department of Computer Systems Architecture, Faculty of Electronic Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland

`julian.szymanski@eti.pg.gda.pl`

² L3S Research Center, Leibniz Universität Hannover, Hannover, Germany
`demidova@L3S.de`

Abstract. In this chapter we summarize activities and results achieved by the Keyword Search Working Group (WG2) of the KEYSTONE Cost Action IC1302. We present the goals of the WG2, its main activities in course of the action and provide a summary of the selected publications related to the WG2 goals and co-authored by WG2 members. We conclude with a summary of open research directions in the area of keyword search for structured data.

1 WG2 - Keyword Search - Objectives

The amount of structured data published on the Web, including entity-centric Web Data, Linked Open Data cloud (LOD)¹ and Knowledge Graphs is constantly growing. These data comes from various sources and domains and has a potential to foster creation of new services and businesses for political, social and commercial activities. In this context, it becomes crucial to enable end users to easily retrieve relevant data from heterogeneous distributed structured sources.

One of the most flexible techniques enabling novice users to access structured data is keyword search. Currently, semantic keyword search over structured sources such as Web Data, LOD cloud, Knowledge Graphs, relational databases and other kinds of structured sources faces severe limitations. This includes insufficient dataset profiling techniques, such as systematic assessment of dataset characteristics [1], ambiguity of keyword queries, scalability problems, as well as lack of query routing techniques that take into account both, query semantics and structured dataset profiles.

Keyword search pipelines over structured data, such as those addressed by WG2, typically include several building blocks, each bringing in its own challenges, namely:

1. Data preprocessing and indexing,
2. Query understanding and interpretation,

¹ LOD cloud diagram: <http://lod-cloud.net/>.

3. Federated search,
4. Retrieval models and ranking, and
5. Integration and fusion of search results.

During the COST Action IC1302 (October 2013 - October 2017) 68 members² joined the Keyword Search Working Group (WG2) and collaborated on the topics related to keyword search on structured data. These collaborations included a range of activities, including joint research activities as well as dissemination and communication events. Together, we aimed at the development of novel methods that address research challenges along all the components of the keyword search pipeline and enable effective and efficient keyword search over structured data sources. In the following we provide a more detailed overview of these activities in Sect. 2, refer to selected research contributions of the WG2 members in Sect. 3 and summarize future research directions in Sect. 4.

2 Dissemination and Communication Activities

During the KEYSTONE COST Action we conducted a number of dissemination and communication activities. These activities included regular Working Group meetings and organization of workshops on the topics related to dataset profiling and federated search at the key venues in the field of Semantic Web such as the Extended Semantic Web Conference (ESWC) in 2014 - 2016 and the International Semantic Web Conference (ISWC) in 2017. Furthermore, we organized a special issue of the IJSWIS in 2016 and conducted a survey closely related to the WG2 topics accepted for publication in the Semantic Web Journal in 2017. Several of these activities were carried out in close collaboration with the WG1 Working Group “Representation of structured data sources” of the KEYSTONE Cost Action. In particular our activities included:

- Organization of several Working Group meetings that aimed at defining research agenda and establishing collaborations within the network. These meeting took place as follows:
 - 24-25 March 2014, Leiden, (NL).
 - 25 May 2014, Hersonissos, Crete, (GR).
 - 17-18 October 2014, Riva del Garda, Trento, (IT).
 - 10-11 May 2015, Kosice, (SK).
 - 22-23 February 2016, Marseille, (FR).
 - 20-21 February 2017, Belgrade, (RS).
- Co-organization of the PROFILES workshop series on Dataset PROFiling and federated Search for Web Data (in collaboration with WG1):
 - PROFILES 2014 workshop at the 11th Extended Semantic Web Conference (ESWC 2014), May 26, 2014, Anissaras, Crete, Greece [2].
 - PROFILES 2015 workshop at the 12th Extended Semantic Web Conference (ESWC 2015), June 1, 2015, Portoroz, Slovenia [3].

² <http://www.keystone-cost.eu/keystone/work-group/wg2/> 68 members.

- PROFILES 2016 workshop at the 13th Extended Semantic Web Conference (ESWC 2016), May 30, 2016, Anissaras, Crete, Greece [4].
- PROFILES 2017 workshop at the 16th International Semantic Web Conference (ISWC 2017), October 22, 2017, Vienna, Austria [5].

The PROFILES workshops series initiated and co-organized by the KEYSTONE COST Action starting from 2014 gathered novel works from the fields of semantic query interpretation and federated search for entity-centric Web Data, dataset selection and discovery as well as automated profiling of datasets using scalable data assessment and profiling techniques. We aimed at promoting research on approaches to analyze, describe and discover data sources - including but not limited to SPARQL endpoints - as a facilitator for applications and tasks such as query distribution, semantic search, entity retrieval and recommendation.

- A Special Issue on Dataset Profiling and Federated Search for Linked Data, published in The International Journal on Semantic Web and Information Systems (IJSWIS) 12 (3), 2016 [6]. In this special issue we included articles performing text analysis and dataset catalog creation. Furthermore, several important aspects of dataset selection were addressed, including their evolution, connectivity, scalable discovery and quality [6].
- A survey on RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications [1]. In this survey, we provided a first comprehensive overview of the RDF dataset profiling features, methods, tools and vocabularies. We organized these building blocks of dataset profiling in a taxonomy and illustrated the links between the dataset profiling and feature extraction approaches and several application domains. This survey aimed towards data practitioners, data providers and scientists, spanning a large range of communities and drawing from different fields such as dataset profiling, assessment, summarization and characterization. Ultimately, this work intended to facilitate the reader to identify the relevant features for building a dataset profile for intended applications together with the methods and tools capable of extracting these features from the datasets as well as vocabularies to describe the extracted features and make them available.

3 Selected Publications Related to the WG2 Objectives

Figure 1 presents a typical keyword search pipeline. In this section, we briefly summarize selected contributions of WG2 according to the pipeline components.



Fig. 1. Keyword search pipeline.

3.1 Data Preprocessing and Indexing

Preprocessing and indexing are the key building blocks to facilitate keyword search. This step becomes particularly challenging in the presence of multilingual context, as well as in large-scale, noisy, domain-specific and non-textual datasets. WG2 contributions that address these challenges include analysis of multilingual entity and event-centric context in [7–10] as well as language-specific fact extraction [11]. The methods for efficient access to large-scale data include efficient URL-based indexing in [12] and extraction of topically coherent interlinked sub-collections from the Web [13, 14] and Web archives [15, 16]. Furthermore, a range of classification and annotation techniques were proposed in [17–20]. This is complemented with techniques for building domain-specific language-resources, e.g. gazetteers in the nutrition domain [21].

3.2 Query Understanding and Interpretation

Query understanding and interpretation over structured data is a challenging task due to the query ambiguity and a large number of possible interpretations. This task becomes even more difficult in presence of non-textual data, e.g. multimedia or sensor data, as well as within complex domains such as scientific literature in life sciences. In WG2 we addressed certain aspects of query understanding and interpretation in these settings. The proposed methods included combining user and database perspective for keyword query interpretation over structured data [22], interactive approaches to query interpretation using ontologies [23, 24], combination of semantic and machine learning techniques for query interpretation [25] and network analysis [26]. Furthermore, specialized domains such as multimedia retrieval in digital libraries [27] as well as techniques for retrieval of specific text passages in life sciences [28] were addressed.

3.3 Federated Search

One of the key challenges in the context of federated search is the lack of reusable datasets and benchmarks for evaluation of federated search results within particular domains. The task of evaluation of federated search results is partially supported by the Text Retrieval Conferences (TREC) within the Federated Web Search Track, where federated search over unstructured data is addressed [29]. The tasks covered by this track include Resource Selection (i.e. selection of the suitable search engine for the query), Vertical selection (i.e. classification of the query in the correct domain) and Results Merging (i.e. the combination of the results of several search engines into a single ranked list). In this context, “FedWeb Greatest Hits” dataset [30] is a test-collection used in the TREC for Resource Selection, Vertical selection, and Results Merging. Federated search techniques are increasingly applied to new domains and kinds of data, which results in new challenges. In WG2, several aspects of such novel applications of federated search techniques including search for workflow fragments [31], context extraction from datasets [32] and semantic mediation for geospatial data [33] were considered.

3.4 Retrieval Models and Ranking

In the context of WG2, a wide variety of retrieval models and ranking algorithms was considered. This includes machine learning approaches for ranking such as learning to rank models [34], approaches to retrieval from small text collections using latent semantic analysis and relevance feedback [35], ontology-based approaches to Information Retrieval [36], entity retrieval for structured data [37] and selectivity-based keyword extraction methods [38]. Furthermore, recommendations approaches developed for specific domains included recommendations of multimedia visiting paths in cultural heritage applications [39] as well as collaborative and content-based recommendation approaches in scientific digital libraries [40].

3.5 Integration and Fusion of Search Results

Search result integration and fusion play an important role to provide a comprehensive overview of information retrieved from different sources. One particular approach to integration and fusion search results considered in WG2 addresses unsupervised search results clustering using document titles and snippets [41]. Further related approaches to data integration and fusion considered in WG2 include crowd sourcing-based improvement of data quality through mapping verification using games of purpose [42, 43], fusion of entity-centric Web Markup [44] and mediation between heterogeneous environmental observation datasets [45].

4 Future Research Directions

Keyword-based search is an established paradigm for Information Retrieval, where keyword queries are matched with unstructured documents. When applied to structured data, Information Retrieval methods require significant adaptation, which has been the subject of extensive studies over the past years in a number of communities including Information Retrieval, Databases, Semantic Web, Web and NLP.

With the four V's of Big Data, namely Volume, Variety, Velocity and Veracity new challenges arise in the field. Big structured data, e.g. Web Data, Linked Data and Knowledge Graphs, require development of methods that: (1) significantly improve scalability and efficiency of indexing and search to address an increased Volume, (2) enhance actuality of data within index structures to cope with the Velocity aspects, (3) account for an increased Variety of data with advanced integration and fusion as well as (4) address the Veracity dimension e.g. with quality analysis and provenance verification.

Further important challenges are related to keyword search in specialized domains containing large-scale, multilingual, scientific (e.g. in the life sciences domain) or non-textual data (e.g. multimedia, spatio-temporal and sensor data). Research directions in these domains can include e.g. domain-specific semantic data representation to enable application of search techniques and adequate presentation of domain-specific, multilingual or non-textual search results.

Finally, semantic dataset descriptions i.e. dataset profiles [1] become increasingly important in the context of federated query and search in particular in the emerging application areas. Generation of dataset profiles and, even more importantly, their tight integration with federated search and query approaches is an important direction for further research.

Acknowledgements. This chapter was supported by COST (European Cooperation in Science and Technology) under Action IC1302 (KEYSTONE).

References

1. Ben Ellefi, M., Bellahsene, Z., Breslin, J., Demidova, E., Dietze, S., Szymanski, J., Todorov, K.: RDF dataset profiling - a survey of features, methods, vocabularies and applications. *Semant. Web J.* (2017)
2. Demidova, E., Dietze, S., Szymanski, J., Breslin, J.G. (eds.): Proceedings of the 1st International Workshop on Dataset PROFiling & fEderated Search for Linked Data, co-located with the 11th Extended Semantic Web Conference, PROFILES@ESWC 2014, Anissaras, Crete, Greece, 26 May 2014, vol. 1151. CEUR Workshop Proceedings, CEUR-WS.org (2014)
3. Berendt, B., Dragan, L., Hollink, L., Luczak-Rösch, M., Demidova, E., Dietze, S., Szymanski, J., Breslin, J.G. (eds.): Joint Proceedings of the 5th International Workshop on Using the Web in the Age of Data (USEWOD 2015) and the 2nd International Workshop on Dataset PROFiling and fEderated Search for Linked Data (PROFILES 2015), co-located with the 12th European Semantic Web Conference (ESWC 2015), Portorož, Slovenia, 31 May–1 June 2015, vol. 1362. CEUR Workshop Proceedings, CEUR-WS.org (2015)
4. Demidova, E., Dietze, S., Szymanski, J., Breslin, J.G. (eds.): Proceedings of the 3rd International Workshop on Dataset PROFiling and fEderated Search for Linked Data (PROFILES 2016), co-located with the 13th ESWC 2016 Conference, Anissaras, Greece, 30 May 2016. vol. 1597. CEUR Workshop Proceedings. CEUR-WS.org (2016)
5. Demidova, E., Dietze, S., Szymanski, J., Breslin, J.G. (eds.): Proceedings of the 4th International Workshop on Dataset PROFiling and fEderated Search for Web Data (PROFILES 2017) co-located with The 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, 22 October 2017, vol. 1927. CEUR Workshop Proceedings, CEUR-WS.org (2017)
6. Demidova, E., Dietze, S., Szymanski, J., Breslin, J. (eds.): Special issue on dataset profiling and federated search for linked data. *Int. J. Semant. Web Inf. Syst. (IJSWIS)*, **12** (2016)
7. Gottschalk, S., Demidova, E.: MultiWiki: interlingual text passage alignment in Wikipedia. *TWEB* **11**, 6:1–6:30 (2017)
8. Zhou, Y., Demidova, E., Cristea, A.I.: What's new? Analysing language-specific Wikipedia entity contexts to support entity-centric news retrieval. In: Nguyen, N.T., Kowalczyk, R., Pinto, A.M., Cardoso, J. (eds.) *Transactions on Computational Collective Intelligence XXVI. LNCS*, vol. 10190, pp. 210–231. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59268-8_10
9. Zhou, Y., Demidova, E., Cristea, A.I.: Who likes me more?: analysing entity-centric language-specific bias in multilingual Wikipedia. In: *The 31st Annual ACM Symposium on Applied Computing*, Pisa, Italy, 4–8 April 2016, pp. 750–757 (2016)

10. Zhou, Y., Demidova, E., Cristea, A.I.: Analysing entity context in multilingual wikipedia to support entity-centric retrieval applications. In: Semantic Keyword-Based Search on Structured Data Sources - First COST Action IC1302 International KEYSTONE Conference, IKC 2015, Coimbra, Portugal, 8-9 September 2015, pp. 197–208 (2015). Revised Selected Papers
11. Boiński, T., Chojnowski, A.: Towards facts extraction from text in Polish language. In: 2017 IEEE International Conference on Innovations in Intelligent SysTems and Applications (INISTA), pp. 13–17. IEEE (2017)
12. Souza, T., Demidova, E., Risse, T., Holzmann, H., Gossen, G., Szymanski, J.: Semantic URL analytics to support efficient annotation of large scale Web archives. In: Cardoso, J., Guerra, F., Houben, G.-J., Pinto, A.M., Velegarakis, Y. (eds.) KEYSTONE 2015. LNCS, vol. 9398, pp. 153–166. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27932-9_14
13. Gossen, G., Demidova, E., Risse, T.: iCrawl: improving the freshness of Web collections by integrating social Web and focused Web crawling. In: The 15th ACM/IEEE-CE Joint Conference on Digital Libraries, Knoxville, TN, USA, 21–25 June 2015, pp. 75–84 (2015)
14. Demidova, E., Barbieri, N., Dietze, S., Funk, A., Holzmann, H., Maynard, D., Papaïliou, N., Peters, W., Risse, T., Spiliotopoulos, D.: Analysing and enriching focused semantic Web archives for parliament applications. *Future Internet* **6**, 433–456 (2014)
15. Gossen, G., Demidova, E., Risse, T.: Extracting event-centric document collections from large-scale Web archives. In: The 21st International Conference on Theory and Practice of Digital Libraries, TPD L 2017, Thessaloniki, Greece, pp. 116–127 (2017)
16. Gossen, G., Demidova, E., Risse, T.: Analyzing Web archives through topic and event focused sub-collections. In: The 8th ACM Conference on Web Science, Web-Sci 2016, Hannover, Germany, 22–25 May 2016, pp. 291–295 (2016)
17. Szymanski, J., Rzeniewicz, J.: Identification of category associations using a multilabel classifier. *Expert Syst. Appl.* **61**, 327–342 (2016)
18. Szymanski, J.: Comparative analysis of text representation methods using classification. *Cybern. Syst.* **45**, 180–199 (2014)
19. Draszawka, K., Szymański, J., Guerra, F.: Improving css-KNN classification performance by shifts in training data. In: Cardoso, J., Guerra, F., Houben, G.-J., Pinto, A.M., Velegarakis, Y. (eds.) KEYSTONE 2015. LNCS, vol. 9398, pp. 51–63. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27932-9_5
20. Virik, M., Simko, M., Bieliková, M.: Blog style classification: refining affective blogs. *Comput. Inform.* **35**, 1027–1049 (2016)
21. Tagarev, A., Toloşi, L., Alexiev, V.: Domain-specific modeling: towards a food and drink gazetteer. In: Cardoso, J., Guerra, F., Houben, G.-J., Pinto, A.M., Velegarakis, Y. (eds.) KEYSTONE 2015. LNCS, vol. 9398, pp. 182–196. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27932-9_16
22. Bergamaschi, S., Guerra, F., Interlandi, M., Lado, R.T., Velegarakis, Y.: Combining user and database perspective for solving keyword queries over relational databases. *Inf. Syst.* **55**, 1–19 (2016)
23. Demidova, E., Zhou, X., Nejd, W.: Efficient query construction for large scale data. In: The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2013, Dublin, Ireland, 28 July–01 August 2013, pp. 573–582 (2013)

24. Demidova, E., Oelze, I., Nejdl, W.: Aligning freebase with the YAGO ontology. In: 22nd ACM International Conference on Information and Knowledge Management, CIKM 2013, San Francisco, CA, USA, pp. 579–588 (2013)
25. Bergamaschi, S., Guerra, F., Interlandi, M., Lado, R.T., Velegarakis, Y.: QUEST: a keyword search system for relational data based on semantic and machine learning techniques. *PVLDB* **6**, 1222–1225 (2013)
26. Bernabei, C., Guerra, F., Lado, R.T.: Keyword search in structured data and network analysis: a preliminary experiment over DBLP. In: 10th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2015, Trento, Italy, 5–6 November 2015, pp. 1–6 (2015)
27. Bartolini, I., Patella, M.: Multimedia queries in digital libraries. In: Colace, F., De Santo, M., Moscato, V., Picariello, A., Schreiber, F.A., Tanca, L. (eds.) *Data Management in Pervasive Systems*. DSA, pp. 311–325. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20062-0_15
28. Aydin, F., Husunbeyi, Z.M., Ozgur, A.: Automatic query generation using word embeddings for retrieving passages describing experimental methods. *Database* **2017** (2017)
29. Demeester, T., Trieschnigg, D., Nguyen, D., Zhou, K., Hiemstra, D.: Overview of the TREC 2014 federated Web search track. In: *The 23rd Text Retrieval Conference (TREC)* (2014)
30. Demeester, T., Trieschnigg, D., Nguyen, D., Hiemstra, D., Zhou, K.: FedWeb greatest hits: presenting the new test collection for federated Web search. In: *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pp. 27–28 (2015)
31. Belhajjame, K., Grigori, D., Harmassi, M., Ben Yahia, M.: Keyword-based search of workflow fragments and their composition. In: Nguyen, N.T., Kowalczyk, R., Pinto, A.M., Cardoso, J. (eds.) *Transactions on Computational Collective Intelligence XXVI*. LNCS, vol. 10190, pp. 67–90. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59268-8_4
32. Kapitsaki, G.M., Kalaitzidou, G., Mettouris, C., Achilleos, A.P., Papadopoulos, G.A.: Identifying context information in datasets. In: Christiansen, H., Stojanovic, I., Papadopoulos, G.A. (eds.) *CONTEXT 2015*. LNCS (LNAI), vol. 9405, pp. 214–225. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25591-0_16
33. Regueiro, M.A., Viqueira, J.R., Stasch, C., Taboada, J.A.: Semantic mediation of observation datasets through sensor observation services. *Future Gener. Comput. Syst.* **67**, 47–56 (2017)
34. Tax, N., Bockting, S., Hiemstra, D.: A cross-benchmark comparison of 87 learning to rank methods. *Inf. Proc. Manage.* **51**, 757–772 (2015)
35. Layfield, C., Azzopardi, J., Staff, C.: Experiments with document retrieval from small text collections using latent semantic analysis or term similarity with query coordination and automatic relevance feedback. In: Cali, A., Gorgan, D., Ugarte, M. (eds.) *KEYSTONE 2016*. LNCS, vol. 10151, pp. 25–36. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_3
36. Meštrović, A., Cali, A.: An ontology-based approach to information retrieval. In: Cali, A., Gorgan, D., Ugarte, M. (eds.) *KEYSTONE 2016*. LNCS, vol. 10151, pp. 150–156. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_13
37. Fetahu, B., Gadiraju, U., Dietze, S.: Improving entity retrieval on structured data. In: Arenas, M., et al. (eds.) *ISWC 2015*. LNCS, vol. 9366, pp. 474–491. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25007-6_28

38. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Selectivity-based keyword extraction method. *Int. J. Semant. Web Inf. Syst. (IJSWIS)* **12**, 1–26 (2016)
39. Bartolini, I., Moscato, V., Pensa, R.G., Penta, A., Picariello, A., Sansone, C., Sapino, M.L.: Recommending multimedia visiting paths in cultural heritage applications. *Multimedia Tools Appl.* **75**, 3813–3842 (2016)
40. Azzopardi, J., Ivanovic, D., Kapitsaki, G.: Comparison of collaborative and content-based automatic recommendation approaches in a digital library of serbian PhD dissertations. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) *KEYSTONE 2016. LNCS*, vol. 10151, pp. 100–111. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_9
41. Staff, C., Azzopardi, J., Layfield, C., Mercieca, D.: Search results clustering without external resources. In: 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), pp. 276–280. IEEE (2015)
42. Boiński, T.: Game with a purpose for verification of mappings between Wikipedia and Wordnet. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) *KEYSTONE 2016. LNCS*, vol. 10151, pp. 159–170. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_14
43. Jagoda, J., Boiński, T.: Assessing word difficulty for quiz-like game. In: *The International Keystone Conference, IKC 2017* (2017)
44. Dietze, S.: Retrieval, crawling and fusion of entity-centric data on the Web. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) *KEYSTONE 2016. LNCS*, vol. 10151, pp. 3–16. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_1
45. Regueiro, M.A., Viqueira, J.R.R., Stasch, C., Taboada, J.A.: Semantic mediation of observation datasets through sensor observation services. *Future Gener. Comp. Syst.* **67**, 47–56 (2017)