WG2 Session
Keyword Search

chaired by Elena and Julian

Meeting in Belgrade, 21 February 2017

goo.gl/ncYIgW

A map of the KEYSTONE members in the area

http://www.keystone-cost.eu/keystone/work-group/wg2/

68 members

WG2 - Keyword search - Objectives:

The amount of **structured data published on the Web** is constantly growing. These data come from various sources and domains and can foster creation of new services and businesses for political, social and commercial activities. In this context, it becomes very important to enable end users to easily retrieve relevant data from these sources. One of the most flexible techniques enabling novice users to access structured data is **keyword search**. Currently, semantic keyword search over the LOD cloud, relational and other kinds of structured sources faces several problems, such as lack of assessment of data quality, increased ambiguity of keyword queries, scalability problems, as well as lack of query routing techniques that take into account both, query semantics and data quality. In this WG, we aim to support development of novel methods and algorithms that address these problems and enable **effective and efficient keyword search over structured data sources**. In particular, we study techniques for matching user keywords with data structures and the domains of selected sources and formulation of the corresponding queries.

Expected Outcomes

(to be discussed in the meeting - progress made towards the outcomes during the action)

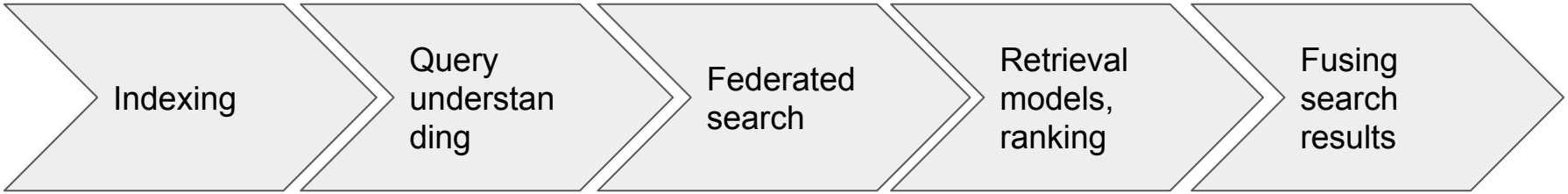
a. Advanced search techniques exploiting statistics, semantics, and metadata.
    i. statistics for query expansion
    ii. metadata for mapping terms to database for relevant or diverse results
    iii. semantic relatedness of terms changing over time
b. Techniques for graph-based search and query interpretation in multi-source search scenarios.
    i. deep web sources (sources with restricted access)
    ii. sensor data with restricted query capabilities, data integration issues, addressed by mediation and relations
    iii. integration of results from deep web
c. Scalable keyword search techniques for large scale structured data.
    i. as a part of interactive search techniques
    ii. freebase data and schema
    iii. still an open research area

Results achieved in the WG2 area

(to be discussed in the meeting)

In which areas do we have achieved significant progress during the action?

statistics, semantics, multi-source, scalability



Results achieved in the WG2 area

(to be completed by WG2 members)

Joint publications (involving KEYSTONE authors from at least 2 countries)

Joint projects (involving organisation from 2 KEYSTONE countries)

Research visits (STSMs)

Published software tools / components / prototypes

Published datasets

What are the open challenges in the WG2 area?

(topics to be discussed in the meeting)

Where do the most challenging problems still reside?

Keyword search in challenging domains: multilingual data, non-textual data, multimedia. Discussion notes: Numerical arrays - what the result of keyword search should be? How to identify the concepts contained in the data? Need of annotations. Domain-specific representations of data. Bag of visual words for searching images. Semantics of terms changes over space and time. Combining visual and textual search for multimedia search. Multilingual search with more advanced techniques for semantic similarity, ontologies, and more languages.

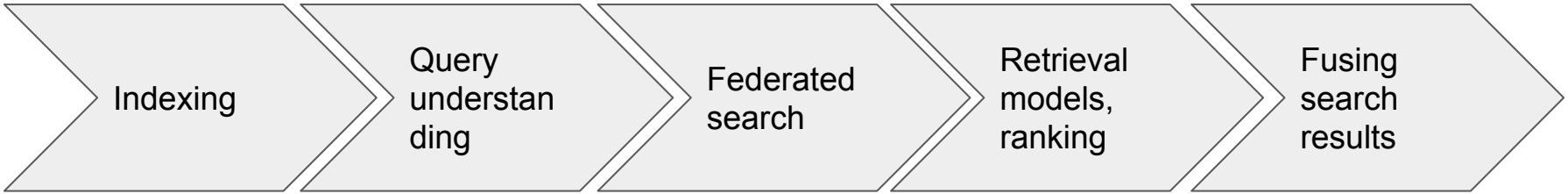
Keyword search in challenging data sources: Big Data, Linked Data, …., scalability, efficiency, quality, trust. Discussion notes: Size of the data, automatically generated numerical data. Verification of search results, dominating sources, linking of library sources to popular search results (e.g. Wikipedia), enriching results with other sources of information.

Search methodologies / applications. Discussion notes: productive systems are still missing - information need of the user and its representation in keyword search results is not alway clear.

A map of open challenges in the WG2 area

(to be discussed in the meeting)

- Challenges along the processing pipeline



A map of open challenges in the WG2 area

(to be discussed in the meeting)

- Indexing, index compression
- Query understanding / interpretation / processing
  - Query intent detection
  - Query reformulation, suggestion, expansion
  - Query representation, languages
- Federated search / using dataset profiles (from WG1)
- Retrieval models and ranking for structured data / similarity measures
- Combination and fusion of results

Dissemination & communication activities

(to be completed by WG2 members)

PROFILES workshop series @ Extended Semantic Web Conference (ESWC) 2014 - 2016

Special Issue on Dataset Profiling and Federated Search for Linked Data, The International Journal on Semantic Web and Information Systems (IJSWIS) 12 (3), 2016

Joint publications: Books and book chapters

(to be completed by WG2 members)

I. Bartolini and M. Patella. Multimedia Queries in Digital Libraries. In Data Management in Pervasive Systems, Series: Data-Centric Systems and Applications, ISBN: 978-3-319-20062-0, Springer, November 2015.

Joint publications: Journal articles

(to be completed by WG2 members)

Yiwei Zhou, Elena Demidova, Alexandra Cristea. What's New? Analysing Language-specific Wikipedia Entity Contexts to Support Entity-Centric News Retrieval. Transactions on Computational Collective Intelligence, accepted in November 2016, to appear.

Regueiro M.A., Viqueira J.R.R., Stasch C., Taboada J.A., Semantic Mediation of Observation Datasets through Sensor Observation Services, Future Generation Computer Systems 67, Elsevier, pp. 47-56, 2017

I. Bartolini, V. Moscato, R.G. Pensa, A. Penta, A. Picariello, C. Sansone, and M.L. Sapino. Recommending Multimedia Visiting Paths in Cultural Heritage Applications. In Multimedia Tools and Applications Journal (MTAP), 75(7): 3813-3842, 2016

K. Belhajjame, Daniela Grigori, Mariem Harmasi, Keyword-Based Search of Workflow Fragments and their Composition. Transactions on Computational Collective Intelligence Online Service, 2017 (To appear)

Joint publications: Conference papers

(to be completed by WG2 members)

Tarcisio Souza, Elena Demidova, Thomas Risse, Helge Holzmann, Gerhard Gossen, Julian Szymanski: Semantic URL Analytics to Support Efficient Annotation of Large Scale Web Archives. International KEYSTONE Conference 2015: 153-166.

Karol Draszawka, Julian Szymanski, Francesco Guerra: Improving css-KNN Classification Performance by Shifts in Training Data. International KEYSTONE Conference 2015: 51-63.

Boiński T.: Game with a Purpose for Mappings Verification, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, Polskie Towarzystwo Informatyczne, 2016, s.405-409

Boiński T.: Game with a Purpose for Verification of Mappings Between Wikipedia and WordNet. Proceedings of the 2nd International KEYSTONE Conference IKC 2016, Cluj-Napoca Romania, 8-9 September 2016

Calì A., Mestrovic A. An Ontology-based Approach to Information Retrieval. Proceedings of the 2nd International KEYSTONE Conference IKC 2016, Cluj-Napoca Romania, 8-9 September 2016

Regueiro M.A., Viqueira J.R.R., Stasch C., Taboada J.A., Sensor Observation Service Semantic Mediation: Generic Wrappers for In-Situ and Remote Devices, 35th International Conference on Conceptual Modeling (ER 2016), Gifu, Japan, 14 - 17 November 2016.

Georgia M. Kapitsaki, Giouliana Kalaitzidou, Christos Mettouris, Achilleas P. Achilleos, George A. Papadopoulos, Identifying Context Information in Datasets, 9th International and Interdisciplinary Conference, CONTEXT 2015, Larnaca, Cyprus, November 2-6,, 2015, 214-225.

Joel Azzopardi, Dragan Ivanovic and Georgia Kapitsaki, 2017, "Comparison of Collaborative and Content-based Automatic Recommendation Approaches in a Digital Library of Serbian PhD Dissertations", IKC 2016: Proceedings of the 2nd International KEYSTONE Conference, Cluj-Napoca, Romania, Springer-Velag, pp 25 - 36

Layfield, C., Azzopardi, J., & Staff, C. (2017). Experiments with Document Retrieval from Small Text Collections Using Latent Semantic Analysis or Term Similarity with Query Coordination and Automatic Relevance Feedback. In A. Cali, D. Gorgan, & U. M. (Eds.), *Semantic Keyword-Based Search on Structured Data Sources. KEYSTONE 2016. Lecture Notes in Computer Science* (Vol. 10151). Cluj-Napoca, Romainia: Springer. http://doi.org/10.1007/978-3-319-53640-8_3

Staff, C., Azzopardi, J., Layfield, C., & Mercieca, D. (2015). Search Results Clustering without External Resources. In *Proceedings of 12th International Workshop on Text-based Information Retrieval*. Valencia.

Joint publications: Editorials, proceedings

(to be completed by WG2 members)

Elena Demidova, Stefan Dietze, Julian Szymanski, John Breslin (Eds.). Special Issue on Dataset Profiling and Federated Search for Linked Data. The International Journal on Semantic Web and Information Systems (IJSWIS), 12(3), 2016.

Elena Demidova, Stefan Dietze, Julian Szymanski, John G. Breslin: Proceedings of the 3rd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '16) co-located with the 13th ESWC 2016 Conference, Anissaras, Greece, May 30, 2016. CEUR Workshop Proceedings 1597, CEUR-WS.org 2016

Bettina Berendt, Laura Dragan, Laura Hollink, Markus Luczak-Rösch, Elena Demidova, Stefan Dietze, Julian Szymanski, John G. Breslin: Joint Proceedings of the 5th International Workshop on Using the Web in the Age of Data (USEWOD '15) and the 2nd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '15) co-located with the 12th European Semantic Web Conference (ESWC 2015), Portorož, Slovenia, May 31 - June 1, 2015. CEUR Workshop Proceedings 1362, CEUR-WS.org 2015

Elena Demidova, Stefan Dietze, Julian Szymanski, John G. Breslin: 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, May 26, 2014. CEUR Workshop Proceedings 1151, CEUR-WS.org 2014

A list of the results achieved: STSMs

(to be completed by WG2 members)

http://www.keystone-cost.eu/keystone/outreach/short-term-scientific-missions-stsms/stsms-approved/

Ontology-based information retrieval: a graph-based approach (Ana Mestrovic University of Rijeka, Rijeka (HR), University of London, Birkbeck College, London (UK))

Interactive Language Learning: Recommending keywords with POS tagged multilingual word graphs (Benjamin Bergner, Otto-von-Guericke-University, Magdeburg, Magdeburg (DE))

Defining research framework for automatic recommendation of new PhD publications and personalization of ranking search results (Joel Azzopardi University of Malta (MT) - University of Novi Sad (RS))

Personalization of search on a digital library of PhD dissertations by re-ranking search results based on automatic user personalization and recommendation (Dragan Ivanovic, University of Novi Sad (RS) - University of Malta (MT))

A list of the results achieved: Open Source Software

(to be completed by WG2 members)

MultiWiki: alignment of Wikipedia text passages across languages
http://multiwiki.l3s.uni-hannover.de/demo.html

Tool for Wikification: https://github.com/mnarusze/enrich-your-text

Application for making computational representation of Wikipedia:
http://kask.eti.pg.gda.pl/CompWiki/index.php?page=matrixu

A list of the results achieved: Open Datasets

(to be completed by WG2 members)

Benchmarks for cross-lingual sentence alignment.

http://multiwiki.l3s.uni-hannover.de/benchmark.html

A list of the results achieved: Joint projects

(to be completed by WG2 members)

QROWD H2020 IA (Germany, UK)

IKC 2017 plans

How about turning (some of the) open challenges in submissions to IKC 2017?