Searching Big Data under Open World Assumption

STSM Scientific Report - COST Action IC1302 KEYSTONE

STSM Applicant

Dr Mariia Golovianko, Kharkiv National University of Radioelectronics, Ukraine (UA)
golovianko@gmail.com

STSM Host

Prof. Vagan Terziyan, University of Jyväskylä, Finland (FI)
vagan.terziyan@jyu.fi

Period

14/04/2016-25/04/2016

1. The purposes of the STSM

The main purpose of the STSM is to study existing search techniques and develop the new, optimized ones for query answering in Big Data under Open World Assumption.

Current search techniques are mainly developed for the systems working under Closed World Assumption which is usually used in the situations when it`s known that the data (knowledge) base is complete or a query can be effectively processed only on a well-defined finite data subset of an incomplete data (knowledge) base. However gaining and using knowledge is a permanent evolutionary process which is never complete. Heterogeneous data sets (e.g., in the web) are growing very fast creating so called Big Data which implies new challenges and non-standard solutions, including new search techniques. Usually keyword search is interpreted as backward chaining in knowledge-based systems, when keywords point out at the search goal. A shift from the closed world to the open world assumption can create new possibilities for search in Big Data. The idea behind the OWA-driven search is forward-chaining when keywords are used as a starting point for search and the results at the middle stages are used for specification of the request and the goal of the search is unobserved at the beginning.

The obtained results of the research should be described in a draft of a research publication for further submission (e.g., for KEYSTONE Second Open Conference in autumn 2016).

To ensure the possibility of future collaborations around the research topic and further development of the ideas and solutions invented during the STSM one of the tasks for the STSM was to investigate funding possibilities for research project submission (e.g., within HORIZON 2020).

2. Description of the work carried out during the STSM

Dr. Golovianko joined the research group headed by prof. Vagan Terziyan. During the first (kick-off) meeting she got acquainted with the research done by the group, their elaborations and topics of interest. Dr. Golovianko also presented her research results. Possibilities of the results integration into the common research were discussed and the tasks for the STSM period were declared according to the workplan of the STSM.

During the STSM dr. Golovianko carried out several types of activities:

- studied (1) the modern semantics of the big data notion, (2) current techniques of information retrieval and information discovery (exploratory search) and (3) the state of research in problem domain (Big Data search) by reading and analyzing corresponding scientific literature and the elaborations of University of Jyvaskyla in this domain;
- took part in brainstorming sessions, contributed to ideas generation aimed at development of a smart tree-based data structure based on users` search history and algorithms for the data structure construction and learning based on swarm intelligence aimed at big data exploratory search;
- carried out research and drafted a paper on web search under open world assumption based on the idea of discovering hidden users` search behavioral patterns (swarm intelligence) and applying them for search sessions optimization;
- prepared lectures on searching big data and state-of-the-art techniques and tools for home university and other purposes.

3. Main results/conclusions

Collection of extremely large volumes of digital information, typical for a wide variety of today`s domains and environments, e.g., industry, business and academic communities, media and web itself, referred usually to as big data, is both challenging and promising due to the broadly recognized high value of data and at the same time high computational costs of data storage and processing influenced by its volume, velocity and variety. Discovery of hidden models or patterns in big data by data mining techniques can not only provide a basis for multipurpose analytics but also contribute to essential optimization of data processing and more accurate and fast information provisioning to end users. Search is among the data processing procedures most depending on the way information is stored and indexed. Despite the fact that essential effort has already been done to build large-scale search engines, in particular for web, there is still a request for optimization of the web search.

The most popular methods of the Web search optimization make use of: the explicit structure of the web based on the links between pages using so called random surfer model which makes prediction of the page value from theoretical probability of the page to be visited; syntactical, statistical and semantic analysis of web pages content; experimenting with various forms and models of pages content and search queries representations. While all these methods try to evaluate the probable relevance of a page and predict a user`s possible interest to it beforehand, there is a source of information which already contains all needed evaluations for intentional surfer model construction - the history of users` search behavior.

The concept of search as an interactive process has already appeared in several researches. It is noticed that people’s conceptions of their information problems change through their interactions with the search system; there are different kinds of information problems, for which different kinds of interactions might be appropriate.

In our research we consider it important to focus on that kind of search which implies that a user looks for the information initially unobservable: information discovery, typical for open world systems which is opposite to retrieval of the particular output Oj observable at the beginning: information retrieval, typical for closed world systems.
Revealing an underlying data structure which reflects hidden dependencies based on the history of user searches can be applied to creation of data indices facilitating fast and relevant query answering. The structure we use is so called "there-and-back" structure (TB-structure). It is organized as an intersection of trees. The structure is interesting due to the possibility of its self-growth as a result of automatic discovery of implicit links and can be used for search sessions storage in form of queries trails provided by various users. A TB-structure storing queries history contains the root layer filled with nodes denoted by initial user`s queries, the terminal layer including leaves denoted by terminal queries – the ones leading to either satisfaction (the search goal achievement) or disappointment (the search termination because of inability to reach the goal).
The structure learns predicting users` behavior by applying an ant colony algorithm and is capable to self-reconfiguring for self-optimization. The use of the structure for web pages indexing helps search engines providing more accurate results for query answering.

4. Future collaborations

The opportunities for further collaborative participation in EU projects were studied during the study visit. Calls of the HORIZON 2020 programme were analyzed, several of them were selected as the most relevant for application: SwafS-01-2016 "Participatory research and innovation via Science Shops", ICT-20-2017 "Tools for smart digital content in the creative industries", ICT-04-2017 "Smart Anything Everywhere Initiative", ICT-11-2017 "Collective Awareness Platforms for Sustainability and Social Innovation". The analysis of the calls was documented for further consideration.

5. Foreseen publications/articles resulting from the STSM

A paper on web search under open world assumption based on the idea of discovering hidden users` search behavioral patterns (swarm intelligence) and applying them for search sessions optimization was drafted. Its submission for KEYSTONE Second Open Conference is planned.

6. Other comments

During the STSM dr. Golovianko got a possibility to access important research papers brought by the library of University of Jyvaskyla to the staff and the students. She doesn`t have such a possibility in her home university.