SHORT TERM SCIENTIFIC MISSION REPORT

STSM details

COST Action	IC1302 - Semantic keyword-based search on structured data sources (KEYSTONE)
STSM Title	Graph-based visualization of query results set based on methods from Formal Concept Analysis
Reference ID	ECOST-STSM-IC1302-220616-078333
STSM dates	from 22-06-2016 to 08-07-2016
Applicant	Dr. Peter Butka
Applicant's institution	Technical University of Kosice, Faculty of Electrical Engineering and Informatics, Department of Cybernetics and Artificial Intelligence, Slovakia
Host	Prof. Dr. Andreas Nürnberger
Host institution	Otto von Guericke University Magdeburg, Faculty of Computer Science, Data & Knowledge Engineering Group (DKE), Germany

Purpose of the STSM

The objectives of this STSM were to develop and implement methods for graph-based visualization of query results set from web documents using the approaches from the area of Formal Concept Analysis (FCA), which are able to produce the hierarchical structures from objects clusters (in this case clusters of query results).
In more details, the goals of the STSM were:

1. Reuse framework for keyword-based search for web results and their analysis, which was developed in DKE group in Magdeburg,
2. Analyze query results set and prepare obtained data for analysis using FCA-based methods,
3. Visualize query results set as hierarchical model of objects clusters (where objects are search results) known as concept lattices, as well as its reductions for better understandability of results,
4. Discuss the possibility for future collaboration of our groups in both preparation of paper on STSM results (if applicable) and collaboration on some other projects (if applicable).

Description of the work carried out during the STSM

First, one of the frameworks of DKE group (CET Search Tool) was analyzed in order to be reused for getting the search results and visualizations of concept lattices based on the results. While we already had some previous tests in this way, we have found graphic part not suitable for more interactive and practical implementations of proposed visualizations based on Formal Concept Analysis (FCA). Therefore, we decided to reuse only backend of the framework for search of results, their first preprocessing and backend for FCA algorithms implementation. In more details, the snippet of every search result was used as input object, analyzed according to the words presented in the snippet, with application of several filters for reduction of terms for creation of object-attribute model for FCA analysis (tokenization, lowcase filter, stopwords, stemming, frequency based filtering, …). In our case, due to the usage of snippets we decided for binary model, where only presence of term in object-attribute models was used.

As next step, we have applied algorithm for FCA model creation (using model known as GOSCL – Generalized One-Sided Concept Lattice), which resulted in concept lattice. This model can be then used directly as hierarchical model for exploration of query result set, or can be reduced in some other way.

For the reduction we have implemented and tested some of the usual approaches or their combinations (which still leads to hierarchical structure based on previously created concept lattice, but it is not usually lattice in mathematical manner, but can be searched and used for navigation):

- Removing the bottom element of concept lattice (is empty set of objects, with all attributes presented)
- Selection of concepts based on the threshold for lowest number of objects (search results) – only concepts with at least N objects are used in resulted hierarchical structure
- Reduction of lattice (or already reduced structure) to tree structure by removing of the edges between concepts based on different conceptual indexes
    - Conceptual indexes are used to define the relevance (score) of edges to parent and select one with best score to retain, for score we can use selection based on stability, support, confidence, similarity (of attributes between parent and child nodes)
- Creation of reduced concept lattice based on the model of Preference-based GOSCL, which is based on selection of different subgroups of attributes with different level of importance – this also leads to the reduction of resulted concept lattice, where more important attributes are also important as differences higher in the lattice structure

Next work was related to the implementation of searchable and interactive visualization of query result sets created using abovementioned methods. We have analyzed some possibilities and prepared two new visualizations based on the *D3.js*[1] and *gojs*[2] frameworks (libraries) for javascript-based graph visualizations. Implemented visualizations were tested on selected queries using Bing Search API.

Additionally, we have discussed the possibility for preparation of joint paper based on the implemented methods and visualizations, as well as other connections to current or expected projects, where FCA-based methods for keyword-based models can be applied.

Description of the main results obtained

As the main results, we could mention design and implementation of tools for preprocessing of query results set, creation of input object-attribute model for FCA-based analysis (based on the application of GOSCL algorithm) and visualization of concept lattices (and their reductions) in interactive and searchable way. It allows users to navigate through query results set in more structured way and understand the differences between particular subsets within the results. We produced (in some way) browser for results where user have hierarchical structure of clusters of results, which he can navigate through and see also particular results related to the selected clusters.

According to the methods or techniques we have:

- Designed and implemented several reduction methods which can be used alone or in combinations – the methods were:
    - Removing of the specifically selected concepts (clusters) with necessary changes in hierarchical graph structure
    - Selection of sublattice from GOSCL based on the expected minimal cardinality of concepts

[1] https://d3js.org/

[2] http://gojs.net/

    - Tree-based reduction of concept lattice – approach which leaves the same number of concepts, but reduce the edges in order to get tree structure based on the conceptual indexes like support, stability (of concepts), confidence, similarity – this method is very useful for creation of more simple navigation menus
    - Implementation of Preference-based GOSCL for creation of one-sided concept lattices with disjunctive subsets of attributes ordered by preferences
- Designed and implemented interactive visualizations for search and navigation through hierarchical structures of concept lattices or their reductions, in all cases there is possibility to see graphical view of concept(s), interact with them (change focus, select other node, etc.) and also see particular links (query results for particular selected / focused cluster), e.g.:
    - Focused visualization – so-called 2-way tree with focus on selected concept (cluster) based on *D3.js* framework – in this case current concept is in the middle and user can see results related to this concept, and he has links to upper and lower concepts (with labeled edges describing the differences between the current node and lower/upper concept)
    - Larger global view on concept lattice – we have used LayeredDigraphLayout graph model from *gojs* framework to see the whole lattice (or reduced hierarchical structure) in standard layout way with most general concept at the top, most specific at the bottom and all other concepts in layers between them (it is usual way for concept lattices visualization)

Future collaboration with the host institution

During the STSM algorithms and visualizations were tested only on smaller examples and query result sets. Our next vision for collaboration is also in testing more precisely the usability of approach and its potential based on the selected users group. Here, DKE can be very helpful due to their experience with the application of use case study approaches. Also, we have discussed other possibilities for usage of FCA-based methods in DKE projects, which will be analyzed in next months. One of the idea is to reuse exactly approaches designed during STSM (or their refinement) for search and navigation of clusters created from technical documents in order to find groups of innovative documents. Another idea discussed for future collaboration was related to the possibility for application of FCA-based methods in analysis of argumentation networks (e.g., structuring of argumentation space based on the concept lattice and extraction of interesting groups of argumentation nodes with expected attributes).

Foreseen publications/articles resulting from the STSM

We wanted to start the preparation of joint paper which will discuss the methods implemented during STSM and their analysis in use case study, which will be defined collaboratively and then realized with the users based on the defined scenarios (currently it is expected that this will be done on DKE site, due to their experience with such studies). After the realization of experiments, we will decide for publication in suitable journal or conference.

Kosice, Slovakia, 14.7.2016

Peter Butka