KEYSTONE IC1302

White Paper

Collecting and generating new ideas

Raquel Amaro (ed.) John G. Breslin Jorge Cardoso
Francesco Guerra Raquel Trillo-Lado Yannis Velegrakis

October, 2014

1 Introduction

In the last decade, the amount of free, large, and open digital structured data available on the Internet, and the demand for search mechanisms to explore such data, has been continuously increasing thanks to the use of technologies and initiatives such as RDF, Linked Open Data, the evolution of databases, etc. Moreover, users have become used to searching for information through keyword-based search interfaces due to the success of web search engines such as Google or Bing. As a result of these parallel developments, a requirement to provide similar techniques to support keyword-based queries over structured data sources in multi-source scenarios, and in the absence of direct access to the instances, has appeared.

The COST Action "Semantic Keyword-Based Search on Structured Data Sources" (KEYSTONE) aims towards building successful and effective solutions to overcome current limitations with keyword-based search, by creating synergies among researchers, technologists and users from different disciplines that have traditionally only partially overlapped, such as Semantic Data Management, the Semantic Web, Information Retrieval, Artificial Intelligence, Machine Learning, User Interaction, and Natural Language Processing. Thus, the main goal of KEYSTONE is launching and establishing a cooperative network of researchers, practitioners and application domain specialists, and coordinating and fostering collaboration among them to enable research activities and technology transfer in the area of keyword-based search over structured data sources.

In this context, as a first activity of the KEYSTONE Action, the Management Committee decided to organise a meeting in Leiden, Netherlands during March 2014. The goals of that meeting were mainly two-fold: 1) to present

KEYSTONE and encourage people to participate in it, and 2) to gather open research issues for semantic keyword-based search on Big Data. For this paper, the results from the working sessions of that meeting have been gathered and are presented here.

Also, the current paper aims to present and disseminate KEYSTONE's activities and potential, both to researchers and industry, and to strengthen collaboration between members by serving as the basis for further joint papers, to be presented in other scientific and dissemination events.

The paper is organised in five main sections. This section introduces the paper; Section 2 presents a brief description of the KEYSTONE action, its members and skills, and its systems and expected outcomes; Section 3 describes the methodology followed during the working sessions and the questions addressed; Section 4 presents the results for each question, obtained through the analysis of the answers provided by the participants; and Section 5 discusses possibilities for futures directions.

2 KEYSTONE Action

KEYSTONE uses a modular approach, grouping the issues related to keyword-based search into three main areas and working groups:

i) definition of metadata for describing data sources (Working Group 1: Representation of structured data sources);

ii) keyword search processes (Working Group 2: Keyword search); and

iii) user interactions (Working Group 3: User interaction and keyword query interpretation).

A transversal group, Working Group 4: Research integration, showcases, benchmarks and evaluations, is responsible for coupling the techniques developed by each of the previous working groups for the creation of an effective framework and the identification of techniques for the evaluation of the approaches.

The work developed will thus allow for reviewing, designing, developing, implementing and evaluating techniques in all three of the typical phases of keyword-based search: a) user keyword input; b) result computation; and c) result output. For each phase, the most promising approaches will be analysed and new approaches will be proposed for (i) analysing the user keywords by identifying the concepts related to them and generating lexical alternatives that may make sense for each data source; (ii) matching user keywords with the underlying data structures in the sources; (iii) formulating queries in the native data source languages corresponding to the user keywords; and (iv) performing the fusion, cleaning and ranking of the results of the possible queries generated during the previous step.

Besides the tasks addressed within each thematic area - critical review of existing emerging techniques to create an open annotated bibliography of the most important approaches and techniques; and definition of open/closed issues and of a roadmap for proposing solutions and approaches - the action promotes and assures a (i) coordination of research activities; (ii) coordination of short term scientific missions; and (iii) dissemination of the outcomes (results, tools, showcases, benchmarks).

The outcomes of KEYSTONE include publications in the most important

international journals and conferences and a freely-accessible website where an annotated bibliography, reference datasets and query sets, scenarios, benchmarks, prototypes and software libraries are made available[1]. The website is designed to enable communications and discussions among the members and external researchers interested in KEYSTONE's activities and fields.

The coordination effort aims at promoting the development of a new revolutionary paradigm that provides users with keyword-based search capabilities for structured data sources as they currently do with documents. Furthermore, it will exploit the structured nature of data sources in defining complex query execution plans by combining partial contributions from different sources.

Participants skills KEYSTONE currently gathers over 115 members from 26 countries: Belgium, Bulgaria, Croatia, Cyprus, Estonia, Finland, France, Macedonia, Germany, Greece, Ireland, Israel, Italy, Malta, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovenia, Spain, Sweden, Switzerland, Turkey and United Kingdom.

Members include researchers, practitioners, and application domain specialists from many areas and sub-areas of relevant disciplines, as presented in Annex 1[2].

3 Methodology

As stated previously, the main goal of this paper is to present the participants' contributions and an analysis of six key aspects of keyword search: challenges in keyword-based search; benefiting practical scenarios through keyword-search research; methods for supporting a user in keyword queries and result analysis; methods for obtaining optimal results from keyword queries; benchmarking environments and evaluation of keyword search; and KEYSTONE's application fields.

In order to achieve rapid results, while ensuring the active involvement of all participants as well as collecting contributions from each participant's areas of expertise, the methodology that was used to gather insights on the key aspects of keyword search was carried out through a directed 'brainwriting' session. The brainwriting session took place at the KEYSTONE COST Action's second meeting in Leiden on 24 March 2014, with a duration of approximately 30 minutes.

3.1 Brainwriting

Brainwriting is a way to take advantage of group priming effects through writing and reading interaction that reduces traditional brainstorming production blocking due to face-to-face interaction inhibitions ([1]).

In this method, a participant writes his or her ideas down on a piece of paper, passes them on to a second participant, who reads and develops them further by adding his or her own ideas and comments, and then that second participant passes the paper on to yet a third participant. The ideas are passed forward, and

[1] `http://www.keystone-cost.eu/`

[2] For further information on individual members see `http://www.KEYSTONE-cost.eu/KEYSTONE/members/`

thus developed and screened by three different participants, without returning to the original source.

Although based in the same principles, group brainwriting has been proven to be more effective than individual brainwriting on the one hand, and than traditional brainstorming on the other, when it comes to heterogeneous groups whose members have different levels of knowledge about the issue at hand.

3.2 Process and questions

The brainwriting session was attended by 70 participants, and was devised as follows:

1. Each participant was given a form with a question, with all forms distributed as evenly as possible amongst the participants;
2. Each participant had 15 minutes to write his or her contributions to the question given;
3. Each participant passed the form to the participant to his or her right (and received, in turn, the form from the participant to his or her left);
4. The participants were given a 5-minute period to review the ideas on the form they were given and add/combine/modify them with new ones;
5. Again, each participant passed the form to the participant to his or her right (and received, in turn, the form from the participant to his or her left);
6. A final 5-minute period was again given to the participants to review the ideas on the form they were given and add/combine/modify them with new ones.
7. The forms were then collected to be analysed by five members selected for this task.

The brainwriting session was geared around six questions, devised and selected by the Working Groups' coordinators. Given the aim of the meeting, the questions the participants had to answer were:

- Challenges: What are the main three challenges in keyword-based search for structured data sources and data analytics?
- Scenario/Use Case: Which practical scenarios do you think can benefit from keyword-search research on Big Data?
- Methods (I): How a user can be supported in the formulation of keyword queries/analysis of the obtained results?
- Methods (II): What should the result be to a keyword query in Big Data?
- Benchmarking/Evaluation: What kind of benchmarking environments (which include scalability, accuracy and feasibility) can be devised for Big Data analysis?

- Application Fields: How can the results of KEYSTONE be used to add value to open data and foster the creation of data intensive companies?

The answers to each question were briefly summarised and discussed in a subsequent session of the meeting, to collect general ideas and contributions from all participants.

All contributions and comments were analysed by the five appointed members and the results are the ones presented in the following section.

4 Results

4.1 Main challenges in keyword-based search

Query expansion Queries can be expanded by, e.g., finding additional keywords by searching for synonyms. Query expansion has long been suggested as a technique for dealing with the fundamental issue of word mismatch in information retrieval [2]. Carpineto and Romano [3] made an exhaustive survey of automatic query expansion methods. For example, WordNet can be used to expand a query with synonyms, and super-ordinates and part-whole relations. Other knowledge bases, such as Wikipedia, Wiktionary, and Freebase, can also be used to disambiguate or expand queries.

Ontologies can also be explored for query enrichment. One challenge to this approach is that, since there are many ("toy") ontologies on the Web, it is not clear how they could be used effectively by a keyword-based search engine. Compared to general methods applied to information retrieval, the schema of structured data sources can be exploited by aligning them with formal ontologies. Keyword-based queries can then be matched to the concepts of the ontology, which will contain a direct mapping to the data source schema. In other words, ontologies can mediate the query, its expansion, and its connection to the data sources.

Uncertainty, context, and profile Keyword-based search has to deal with the uncertainty and fuzziness of the search process.

The keywords provided can be very imprecise formulations (and sometimes misleading) of informational needs. Therefore, exactly matching those keywords may hamper the results. Ambiguity should also be dealt with at the schema/instance level allowing for semantic disambiguation and expansion of queries and search items.

One way to cope with uncertainty is to exploit as much as possible of the query's contexts. The significance of context-based approaches is that they enable us to greatly improve the relevance of search results [4]. What may be ambiguous in isolation may have a clear meaning when put in context. Fuzziness should be seen as an opportunity to support approximate search results. Over large data sets, under constraints such as personalisation, aiming for exact results at query time is likely to be unfeasible. In some cases, very different keywords may be used depending on the background of the user.

It is very different for someone with an Economics background to search for specific data when compared with someone from the field of Human Resources. And yet, these individuals often need to query the same data sets. In complex

applications, for example Twitter, user profiles and user feedback can be exploited to better address informational needs. Experimental results show that search systems that adapt to each user's preferences can be achieved by constructing user profiles [5]. User experience models can be automatically learned and fine-tuned to each individual user. Statistical models can be constructed to reflect users' domains of work and can be used to filter results or expand queries.

Query results The search process should return data that is amenable and "ready" for data analytics. The returned data should carry quality information. Meaningful ways of presenting, visualising, and interacting with results beyond ranked lists are needed.

For example, in dynamic domains, data that arrives at time t_0 can be superseded by new data arriving at time t_1. Providing an interface which shows users how a set of keywords provides a different set of results over time can bring a new dimension to the visualisation of data.

Retrieving results and their relationships is becoming more important as the synergies of connected data are becoming clearer (e.g., Linked Data). Techniques are needed to explore results visually in order to find relevant patterns and relationships amongst search results [6].

4.2 Practical scenarios that benefit from keyword-search research

We asked 10 national representatives from the KEYSTONE Action to describe some of the practical scenarios that they felt could benefit from keyword-search research on Big Data. After describing their scenario(s), the experts passed on their idea to another person who further explored and refined them. Six of the 10 scenarios were refined two more times, and four were refined only once. Some of the responses were grounded in specific application domains, whereas others chose to identify the research areas that were thought of as needing more work.

In terms of the research topics of interest for keyword search-based scenarios, the experts identified the following:

- Real-time streams search and trend tracking
- Descriptive/predictive analytics
- More precise data discovery and fine-grained search results (since Big Data often implies an aggregate analysis: macro rather than micro)
- Ranked sentiment analysis linking to background information with named-entity recognition
- Data analysis based on a user input that is less structured
- Decision support system improvements based on text summarisation
- Going beyond text search to multimedia search with ranking/relevance
- Combining human interpretations with high-performance computing
- More easily-navigable and browsable results

Some of the experts identified more than one interesting application scenario for keyword-based search. Of these, there was one use case that appeared three times in the respondents' scenarios – that of using past medical case histories to help with a current patient. The scenarios identified were:

- Improved energy/resource usage monitoring
- Processing huge collections of government documents
- Enabling physicians treating a patient to access Big Data that references several past similar cases from multiple heterogeneous EHRs (electronic health records) with the same medical condition (linking these to bibliographies, genetic data, phenotypes and other potential relationships)
- Evidence discovery and fact checking of information brought up in company meetings from thousands or millions of company documents (spreadsheets, logs, ERP, etc.), if possible in real time
- Using social/linked data to improve the personal search experience, e.g. "find me a 'nice' job"
- Comparison of research literature in journals, etc.
- Community health and outbreak tracking using data from public sources
- Searching in content from multiple messaging services
- Searching for criminal data from various police organisations and records
- Searching across heterogeneous review sites
- Searching libraries, document management systems, e-learning resources
- News personalisation (using a person's profile and geotemporal information)
- Finding the best candidate selection for a task across many user profiles (expert finding)

4.3 Methods for supporting the user in keyword queries and results analysis

The connection between the quality of a query and the quality of search results is well defined in the literature [7], and it has motivated a lot of research activity. Some of the main outcomes generated by this effort are summarised in [8], where the information seeking process has been divided into seven steps:

1. Recognising a need for information;
2. Accepting the challenge to take action to fulfil the need;
3. Formulating the problem;
4. Expressing the informational need through a search system;
5. Examining the results obtained by the search system;

6. Reformulating the problem or re-expressing the informational need if the results do not provide sufficient information;

7. Using the information found.

According to this survey, steps 4, 5 and 6 are the ones where the ICT research community has mainly focused its activities, and these are the ones better supported by software applications. The answers provided by our focus group support this assessment and propose some suggestions about how to customise/redefine these processes for querying big data. In particular, the suggestions that can be extracted from the analysis of the answers concern steps 4 and 5. The process for reformulating queries was mentioned in some answers, and considered as crucial for the information seeking process, but no specific technique/tool has been proposed.

First of all, there is an implicit assumption common to all the answers analysed: search systems are queried via keywords or expressions in natural language. No proposal for the definition of a structured language to be adopted for querying Big Data has been suggested. This implies that the reference search system from Big Data should be oriented towards a "best-match" search paradigm, where informational needs are often vague and subjected to a progressive and gradual process of refinement enabled by the search activity itself. Structured queries would assume the development of an "exact match" search paradigm, where a correct specification of the user informational needs exists and answers are perfect.

Moreover, it is possible to roughly classify the approaches proposed by our focus group into two categories: (a) *task-based approaches* that consider query expressions and analysis of the results as two separate processes to be independently improved, and (b) *holistic approaches* that consider both of the tasks as a single combined process to be improved.

Task-based approaches The query expression and the analysis of the results are considered as two separate tasks, that have to be separately adapted/improved when applied to Big Data.

Concerning the query expression, approaches based on typed queries, interactive queries and collaborative formulations have been proposed by the focus group. Approaches based on typed queries aim to match keywords with elements extracted from some reference ontology/knowledge base, thus making the query independent of the specific keywords adopted for expressing the need.

For this purpose, DBpedia, Freebase, WordNet, GeoNames, Yago and other LOD sources have been suggested as possible examples. Ontologies can support NLP tasks for query and acronym disambiguation, semantic enrichment of queries with related terms, detection of misspelled words, etc.

Advanced processes for transforming keywords can be envisaged: an interesting example is provided by entity tracking that allows the development of techniques taking into account how a concept "changes" in the time. For example, if we are interested in some event that occurred in the city of Saint Petersburg between 1924 and 1991, we should ask for "Leningrad", the name of the city during that time period. The use of ontologies also enables the development of a wide range of interaction capabilities to support information seekers in expressing their needs.

The main proposals that emerged concern the development of techniques for query auto-completion based on log analysis, and the discovery of similar/related keywords according to the reference ontology. Suggested completion for a query can be expressed in the form of text (i.e., other keywords), concepts/properties/relations from the reference ontology, and images (the association of images to keywords is definitely an innovation if we consider the standard interfaces of search systems).

Finally, techniques for profiling users can be adopted for identifying similar users (i.e., users with the same informational needs), thus enabling a sort of collaborative system where queries expressed by a user can help to improve the queries of other users. The final goal is the development of a sort of recommendation system for keyword queries taking into account popular queries from similar users. Even if user profiling is an important and well-known research task that can provide interesting results coupled with search systems, an interesting proposal is to attempt at identifying the task the user is carrying out with the query, rather than trying to categorise the user. Ontologies can support the process and allow the suggestion of other keywords related to the same task.

Concerning the analysis of the results, the main suggestions concern the development of techniques for (1) ranking, (2) filtering and (3) summarising the results.

The ranking of the results can be based on the feedback obtained by other users and user profiling techniques for identifying user preferences. Semantic-based, machine learning and heuristic-based techniques can be adopted for this purpose. Faceted search techniques can also be applied to the result set, enabling users to discriminate among the results according to their needs. Techniques for clustering similar results and summarising them can provide users with an insight of the obtained answers.

Moreover, results shown can be enriched with other information generated by the ontologies. They could allow the identification of similar/related answers, provide information about the context, and enable a visualisation of the outcomes with different levels of abstraction. Finally, only one suggestion concerned the need to give an account of the uncertainty of results. Nevertheless, this information could provide an interesting measure for supporting users in the result analysis.

Holistic approaches Holistic approaches consider the expression of the query and the analysis of the results as part of the same process. Techniques similar to the one described in the previous sub-section can be applied. However, in this scenario, the satisfaction of the user's informational needs is achieved through an iterative process where the analysis of the results generated by an initial keyword query (or a partial set of the results if the operation requires a big computational effort) is the basis of the next re-formulation of the query providing results closer to what the user is looking for.

The process can reiterate some times until the user decides to stop it. In particular, keywords and results can be visualised as part of a semantic network (built by exploiting ontologies and reference knowledge bases) and be related to each other. The user should be able to change some keywords or relax some constraints in the query and to visualise in real-time the implications of the changes in the answers provided.

4.4 Methods for obtaining optimal results of a keyword query

The answers provided to this question mainly focused on four topics: (1) the "composition", (2) the format, (3) the granularity of the results, and (4) the goal that users want to achieve with the results.

Concerning the first topic, the focus group thinks that it can be useful to couple the results of a query with some metadata describing the source and the data returned. Possible examples include temporal information, some measure of the quality of the answer, the level of liability of the source and some other information about provenance. An explanation of the motivation for why an answer is related to a query is an interesting new metadata that can be also included.

There are several possible formats adopted for providing and publishing the results of a query: the search systems can return items directly extracted form the data sources, such as text (keywords, snippets of text, etc.) and images, or items which are obtained through a further elaboration of the results, such as fragments of some reference ontology (concepts and properties with associated values), graphs built with some elaboration of the results, and images related to the query answers. The selection of the format adopted depends on the recipients of the answer. A query could be oriented to human and/or machine consumption and these two kind of possible recipients require different specifications of the results. We envisage the need of generating two answers, with different formats, for each query.

The focus group highlighted the need for providing the results with different levels of detail. Some ideas for achieving this purpose include the use of clusters and ontologies. Clustering the results can provide users with a summarised view of the data retrieved. Some further elaborations of the clusters can be envisaged: (a) different clusters of the same data built by means of different distance measures can provide a more complete view of the results; (b) the application of keyword-search techniques over the clusters can support the user in retrieving data of interest; (c) clusters can be categorised in taxonomies, by means of ontologies, thus providing users with results having different levels of abstraction; (d) some measures about the quality of the clusters can be computed and shown (dimension of the cluster, evaluation of its information power, the quality of the results contained).

Finally, the focus group remarked that the results of a query can be the starting point for new informational needs. For this reason, it could be interesting to include in the search system the capability to relate results obtained from different queries performed by the same user.

4.5 Benchmarking environments and evaluation

A fundamental requirement in every computational task is the ability to evaluate the developed solution. This is typically done through benchmarks. A benchmark is a standardised and widely acceptable set of tests that provides a set of metrics on the performance of the solution under different circumstances. Benchmarks help: developers in understanding the performance of the products of their work; practitioners in evaluating the different solutions that the market offers and choosing the best for their tasks at hand; and, finally, researchers

in understanding the limitations of existing solutions and in guiding their work towards the right direction.

Benchmarking requires considerable attention, specifically in the era of Big Data. There is already work that focuses on the evaluation of existing (or future) Big Data systems. Yet, not much work has been performed at algorithmic or service levels, an area that requires particular attention from the keyword-based search community. The challenging question that needs to be answered is, thus, what kind of benchmarking environment needs to be devised for keyword-based search on Big Data.

One of the fundamental components of every benchmark is the kind of dataset that it offers. A benchmark, to be widely acceptable and used, needs to reflect real life scenarios. Popular scenarios that are being explored at present are based around social web and e-commerce applications. This means that sources like Wikipedia, WordNet, social networks, and industry- or manufacturing-related data can play a significant role in providing the required datasets. On top of this, we live in a linked world. Thus, interlinked information, followers, Linked Data, and associated components (even across different data sources) can provide additional properties to the datasets used in the evaluation of the systems. Furthermore, log files from information systems can provide valuable input on the way real users are actually using systems. Integrating this data in the evaluation gives additional value to the produced results.

The second important component of a benchmark is the functionality it offers and the tests, i.e., the set of experiments required in the evaluation task. Traditionally, these experiments are about the efficiency and effectiveness of the tool. However, Big Data has introduced new tasks and functionalities that are important for Big Data systems and that have to find their way in the respective benchmarks. Examples of such tasks are the ability to effectively visualise the form of the data and communicate their properties to the user, or the ability to discover provenance of data in an integration.

Big Data concerns a very broad field and not a specific task. Keyword searching can be performed in many fields, and thus the experiments that a benchmark would require should be able to distinguish between tasks. For instance, it is important to understand: whether search is happening at real time or is performed offline; whether it is performed on a static repository or on a stream of data; or whether the performance results are saying something about the hardware infrastructure or the software components of the system. To this, the fact should be added that the intended task affects significantly the meaning of the results of the evaluation. For different tasks, different factors may have different levels of importance, and even different experiments may play different roles. For this, there is a clear need for understanding and evaluating different software and hardware configurations for specific tasks, which can be facilitated through a centralised repository where the datasets, metrics and experiments to be performed are available and categorised based on the intended task.

The core component of every benchmark is the set of metrics that it uses to characterise the system under evaluation. These metrics are typically standardised. Big Data systems and software are highly diverse not only in operations but also in nature, often making the specification of a given metric a hard task. A benchmark may consist instead of a set of guidelines on how the benchmarking tasks are to be implemented. Naturally, it will always be important to keep in mind the fact that even such guidelines should not leave room for different and

conflicting interpretations, otherwise the comparison among different systems will not be fair, or even understandable.

4.6 KEYSTONE application fields

The main goal of KEYSTONE is to launch and establish a cooperative network of researchers, practitioners, and specialists working in different areas (IR, Semantic Web, Databases, etc.) to promote and foster the development and analysis of techniques for keyword-based search over structured data sources from different perspectives. Therefore, the question at hand was formulated with a dual purpose. First, we wanted to discover which topics were considered most relevant for fostering the development of the emerging Web of Linked Open Data (in contrast to the current Web, which is oriented towards linking documents). Second, we wanted to identify which topics related to KEYSTONE are interesting for the industry (to improve knowledge transfer between research communities and enterprises across Europe, as ERA suggests [9]).

According to the answers provided in the survey, the most desired-for result of KEYSTONE is the creation of standardised vocabularies in different areas (traffic, social, agriculture, etc.) in order to facilitate the federation of data sources and their integration. Moreover, a great percentage of participants also asked for the creation of guidelines and intuitive tools to anonymise structured data from companies, and to create/publish open repositories of structured data. The evaluation of current techniques, tools and services for keyword-based search on structured data sources, by considering different features (their interface, ease of use, efficiency of search, scalability, etc.), was also considered a key aspect. In this context, the creation of a benchmark (similar to TREC for unstructured data) was proposed.

Despite the fact that KEYSTONE is not a research project but a COST Action to promote networking activities and collaboration, some participants also would like software to be generated as a result. In particular, participants are interested in software focusing on semantic search, data mining, sentiment analysis, and machine learning. Thus, a potential task is the development of a compendium of software in a collaborative way, as well as the already-envisaged creation of an annotated bibliography.

Apart from scientific and technical results, a large number of researchers also expect to get in contact with people with similar or complementary interests and knowledge in order to submit joint proposals in future Horizon 2020 calls. Finally, academics/researchers would like to contact real-world enterprises/companies in order to transfer their research results to industry, and in the other direction, enterprises would like to gather the expertise available from diverse research groups.

5 Future trends and final remarks

Over time, the number of digital structured data sources has increased exponentially. This growth has been such that, in some sectors, it is considered that soon there will not be enough space to store such an amount of data in the near future. Under these circumstances, techniques to compress data would be useful. Moreover, some directives to determine which data should be saved and

which data should be discarded could help non-experts to organise and manage their repositories.

Future trends can be analysed from two different perspectives: one in the context of transactional data sources, and another in the context of analytical data sources. Furthermore, mixed scenarios where both types of data sources could be consulted should also be considered. In the transactional context, speed (short response time) is a key feature, while in the analytical context, automatic discovery and integration of information from different heterogeneous data sources is more important. In both contexts, provenance of information/data and the liability of information sources are getting more attention.

References

[1] Vincent R. Brown and Paul B. Paulus. Making group brainstorming more effective: Recommendations from an associative memory perspective. *Current Directions in Psychological Sciences*, 11(6):1:208–1:212, 2002.

[2] Jinxi Xu and W Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM, 1996.

[3] Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, January 2012.

[4] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 406–414, New York, NY, USA, 2001. ACM.

[5] Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 675–684, New York, NY, USA, 2004. ACM.

[6] F.V. Paulovich, R. Pinho, C.P. Botha, A Heijs, and R. Minghim. Pex-web: Content-based visualization of web search results. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 208–214, July 2008.

[7] W. Bruce Croft and R. H. Thompson. I^3r: A new approach to the design of document retrieval systems. *JASIS*, 38(6):389–404, 1987.

[8] Gary Marchionini and Ryen White. Find what you need, understand what you find. *Int. J. Hum. Comput. Interaction*, 23(3):205–237, 2007.

[9] Improving knowledge transfer between research institutions and industry across europe: embracing open innovation. Technical report.

ANNEX I

Areas contributing to KEYSTONE

(collected from the KEYSTONE website, September 2014)

1. Applied computing
 1.1 Arts and humanities
 1.2 Document management and text processing
 1.3 Education
 1.4 Life and medical sciences
2. Computer systems organization
 2.1 Architectures
3. Computing methodologies
 3.1 Artificial intelligence
  3.1.1 Distributed artificial intelligence
   3.1.1.1 Cooperation and coordination
   3.1.1.2 Intelligent agents
   3.1.1.3 Multi-agent systems
  3.1.2 Knowledge representation and reasoning
   3.1.2.1 Causal reasoning and diagnostics
   3.1.2.2 Ontology engineering
   3.1.2.3 Probabilistic reasoning
   3.1.2.4 Reasoning about belief and knowledge
   3.1.2.5 Semantic networks
   3.1.2.6 Temporal reasoning
   3.1.2.7 Vagueness and fuzzy logic
  3.1.3 Natural language processing
   3.1.3.1 Discourse, dialogue and pragmatics
   3.1.3.2 Information extraction
   3.1.3.3 Lexical Semantics
  3.1.4 Search methodologies
 3.2 Distributed computing methodologies
  3.2.1 Distributed algorithms
  3.2.2 Distributed algorithms
  3.2.3 MapReduce algorithms
  3.2.4 Self-organization
 3.3 Machine learning
4. Human-centered computing
 4.1 Collaborative and social computing
 4.2 Human computer interaction
 4.3 Interaction design
 4.4 Ubiquitous and mobile computing
 4.5 Visualization
5. Information systems
 5.1 Data management systems
  5.1.1 Data structures
  5.1.2 Database administration
  5.1.3 Database design and models
  5.1.4 Database management system engines
  5.1.5 Information integration
   5.1.5.1 Data cleaning
   5.1.5.2 Data exchange
   5.1.5.3 Data warehouses
   5.1.5.4 Deduplication
   5.1.5.5 Entity resolution
   5.1.5.6 Extraction, transformation and loading
   5.1.5.7 Federated databases
   5.1.5.8 Mediators and data integration
   5.1.5.9 Wrappers (data mining)
  5.1.6 Middleware for databases
  5.1.7 Query languages
6. Information retrieval
 6.1 Document representation
 6.2 Evaluation of retrieval results
 6.3 Information retrieval query processing
  6.3.1 Query intent
  6.3.2 Query log analysis
  6.3.3 Query reformulation
  6.3.4 Query representation
  6.3.5 Query suggestion
 6.2 Retrieval models and ranking
  6.2.1 Combination, fusion and federated search
  6.2.2 Information retrieval diversity
  6.2.3 Language models
  6.2.4 Learning to rank
  6.2.5 Rank aggregation
  6.2.6 Similarity measures
 6.3 Retrieval tasks and goals
  6.3.1 Business intelligence
  6.3.2 Clustering and classification
  6.3.3 Expert search
  6.3.4 Information extraction
  6.3.5 Near-duplicate and plagiarism detection

6.3.6 Recommender systems
6.3.7 Sentiment analysis
6.4 Search engine architectures and scalability
6.4.1 Distributed retrieval
6.4.2 Peer-to-peer retrieval
6.4.3 Search engine indexing
6.4.4 Search index compression
6.5 Users and interactive retrieval
6.5.1 Collaborative search
6.5.2 Personalization
6.5.3 Search interfaces
6.5.4 Task models
6.6 Information storage systems
6.6 Information systems applications
6.6.1 Collaborative and social computing systems and tools
6.6.2 Data mining
6.6.3 Decision support systems
6.6.4 Digital libraries and archives
6.6.5 Enterprise information systems
6.6.6 Mobile information processing systems
6.6.7 Process control systems
6.6.8 Spatial-temporal systems
6.7 World Wide Web
6.7.1 Web applications
6.7.1.1 Crowdsourcing
6.7.1.2 Internet communications tools
6.7.1.3 Social networks
6.7.2 Web data description languages
6.7.2.1 Markup languages
6.7.2.2 Semantic web description languages
6.7.3 Web interfaces
6.7.4 Web mining
6.7.5 Web searching and
6.7.6 information discovery
6.7.6.1 Collaborative filtering
6.7.6.2 Content ranking
6.7.6.3 Personalization
6.7.6.4 Social recommendation
6.8 Web Services
7. Mathematics of computing
7.1 Discrete mathematics
7.1.1 Combinatorics
7.1.2 Graph theory
7.2 Probability and statistics
7.2.1 Probabilistic algorithms
7.2.2 Probabilistic inference problems
7.2.3 Probabilistic reasoning algorithms
7.2.4 Probabilistic representations
8. Security and privacy
8.1 Human and societal aspects of security and privacy
9. Theory of computation
9.1 Formal languages and automata theory
9.2 Logic
9.3 Semantics and reasoning