Enterprise Search, Graph Knowledge, Semantic Web

Gianfranco E. MODONI[*], *Marco SACCO*[**], *Walter TERKAJ*[**]

A SEMANTIC FRAMEWORK FOR GRAPH-BASED ENTERPRISE SEARCH

Abstract

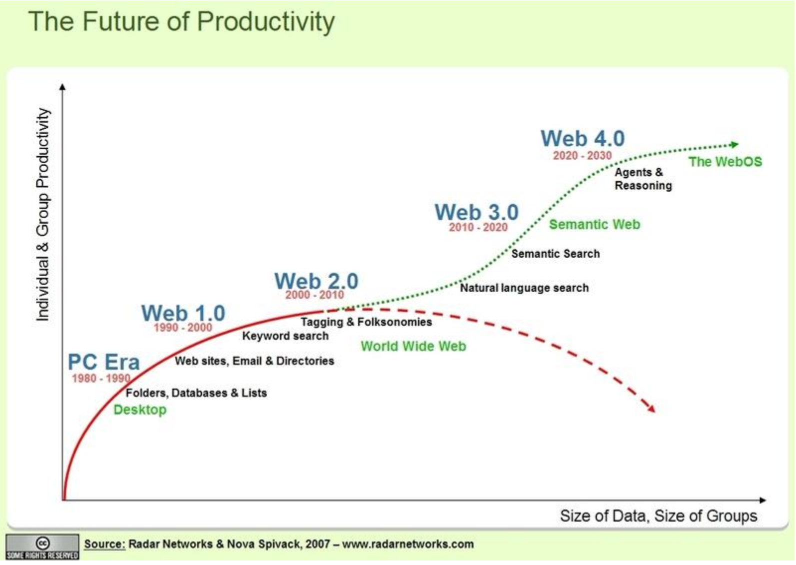
Various recent studies have shown that in many companies workers can spend near half of their time looking for information. Effective internal search tools could make their job more efficient. However, a killer application for this type of solutions is still not available. This paper introduces an envisioned architecture, which should represent the foundations of a new generation of tools for searching information within enterprises.

1. INTRODUCTION

Search engines are important productivity tools that allow users to retrieve desired information through a simple graphical interface, which embodies the one single entry point to search and explore anything related to a specific domain, i.e. within an organization. Companies of all sizes are trying to take advantage and benefit of the increased productivity the search engine technologies can provide. Following this trend, various approaches have emerged and one of the most popular is based on *Knowledge Graph* [1]. This term has been popularized thanks to Google Inc., who enhanced its web search engine results through a graph structure that understands real-world entities and their relationships, providing in this way detailed information about every examined topic in addition to a list of links to other related entities. Graph-based structure is also at the heart of social networks technologies like Facebook, whose semantic search engine, Facebook Graph Search [2], is designed to give answers to user in the form of natural language queries rather than a list of links. According to Spivack [3], such an approach marks the beginning of the third phase (Web 3.0) of the World Wide Web (Fig. 1), which is based on the idea that machines understand the meaning of the exchanged information and are able to make logical relations between them.

[*] Institute of Industrial Technologies and Automation - National Research Council, via Lembo 38 Bari, Italy, gianfranco.modoni@itia.cnr.it

[**] Institute of Industrial Technologies and Automation - National Research Council, via Bassini 15 Milano, Italy, marco.sacco@itia.cnr.it, walter.terkaj@itia.cnr.it



Source: Radar Networks & Nova Spivack, 2007 – www.radarnetworks.com

Fig. 1. The future of Productivity depending on the size of data [source: own study]

Whereas worth Internet search engines are now widely spread, valid enterprise searching tools are still a chimera; thus, the development of this type of applications remains a crucial issue to be faced. It is no coincidence that in many companies "nearly half of a knowledge worker's time is non-productive, spent gathering information, converting formats, unsuccessfully searching or recreating content that already exists" [4]. Moreover, as users are used to retrieve any information they need on the Internet through search, they expect the same behavior within an internal search tool.

Thus, more efforts have to be undertaken by researchers and developers to identify a new generation of internal tools to search, analyze and filter information within an organization [5]. The idea of the research work introduced in this paper goes in this direction, as it addresses the efforts to design the Knowledge Graph Semantic Framework (KGSF), a new interactive and explorative environment for the semantic search of the information within enterprises. Combining and adapting approaches and techniques borrowed from different fields, ranging from artificial intelligence to database management, this work assesses the role of *Knowledge Graph* paradigm to generate an integrated and aggregated view on relevant business knowledge that enhances search operations in real scenarios.

Business of the current enterprises are driven now more than ever by creation and utilization of huge information. Under these conditions, the main key factors for a worth search application are its capability to be integrated with the existing organization's infrastructure and to process large collections of heterogeneous near real-time data in order to convert them into valuable knowledge [6].

Thus, the design of KGSF has to take into account and consider this capability as a mandatory requirement for the development of the underlying architecture.

The remainder of the paper is structured as follows. Section 2 examines the recent technological trend for search engines, whereas Section 3 introduces and illustrates the proposed architecture on the basis of KGSF. Finally, Section 4 draws the conclusions, summarizing the major findings.

2. TOWARDS THE SEARCH ENGINE 3.0

Traditional technologies for search engines, based on keywords, are becoming less valid and effective as the Web increases in size (Fig. 1) [3]. Moreover, they provide a comfortable way for the user to specify information needs, but do not formally capture the explicit meaning of the user input queries [7]. Unlike these, a semantic search engine accesses to the semantics of the information and is oriented to understand the real dynamics between the entities, which represent specific concepts in a certain domain. The input is no longer a list of keywords, but a question, from which the engine will be able to extract the relevant concepts, disambiguate them if necessary and compose a set of queries to build the list of results.

The semantic search can play a crucial role to reduce the errors in the search results that are caused by polysemy (the capacity for a word to have multiple related meanings), synonymy (the capacity for a word to have the same or similar meaning of another word) and malformed queries. In fact, the paradigm shift triggered by Google Inc. towards the search of “things, not strings” [1] goes in the direction to disambiguate the input queries, aiming at understanding the difference between queries like “California” for the state of United States, “California” for the song and “California” for the sports car. This functionality allows the users to avoid wasting time looking through the possible results and also reduces the need to refine the queries. Using instead a syntactic search engine based on keywords, the results are less precise (i.e. many are not related) and incomplete (i.e. not all are relevant to the request).

Thanks to the underlying semantics, a Knowledge Graph approach also contextualizes user searches by creating logical links between the entered data; in this way, it provides the users a summary of key information about what it is looking for and a selection of in-depth details. In this regard, a significant example is represented by the above mentioned search of state “California” that can be defined by its connections to its governor, time zones, population, and so on.

In view of the advantages provided by the use of a semantic-based engine, such an approach has been exploited to design the herein presented KGSF, an interactive and explorative environment for the search of the information within enterprises. An overview of the KGSF architecture is presented in the next section.

3. THE KGSF ARCHITECTURE

The KGSF aims in the first instance at enhancing the semantic integration between the involved heterogeneous sources of data, which contributes in turn to improve the quality of search results. Compared to the state of the art, this solution is expected to provide improvements that can be summarized into four basic points:

- mashup and integration of the corporate knowledge for enabling search within structured data;
- accurate searches through the ability to disambiguate the user input queries;
- presentation of the key facts summaries related to what the end-user is looking for;
- explorative suggestions which allow users to navigate the search space while formulating their queries.

The envisioned architecture on the basis of the proposed KGSF consists of four pillars: (I) Common data model, (II) Backend, (III) Semantic Repository and (IV) Frontend (Fig. 2). This latter (Pillar IV) represents the one single graphical gateway for the search of the information within enterprises. It provides an integrated view on high-quality data derived through an automated analysis process managed by Backend (Pillar II), which plays the role of integrator, transforming and combining structured and unstructured legacy data into valuable information according to the Common data model (Pillar I). After this conversion process, the extracted information are also persisted and stored in the Semantic Repository (Pillar III). The pillars are discussed in details in the following subsections.

3.1. Common data model

In business contexts data come from several sources and can be expressed in a variety of forms. This makes it complex and tricky to find the right information, setting the stage for the first challenge to realize the KGSF: overcoming the problems deriving from the lack of integration between the involved heterogeneous sources of data, thus enhancing their semantic interoperability [8]. In this research work the definition of a common reference model, which embodies a representation of all the organization objects, is proposed as a solution to harmonize data from different sources, providing in this way a systematic manner to classify and integrate the knowledge of the organization. A common data model aggregates and unifies all the information, improving significantly the potential of a search application.

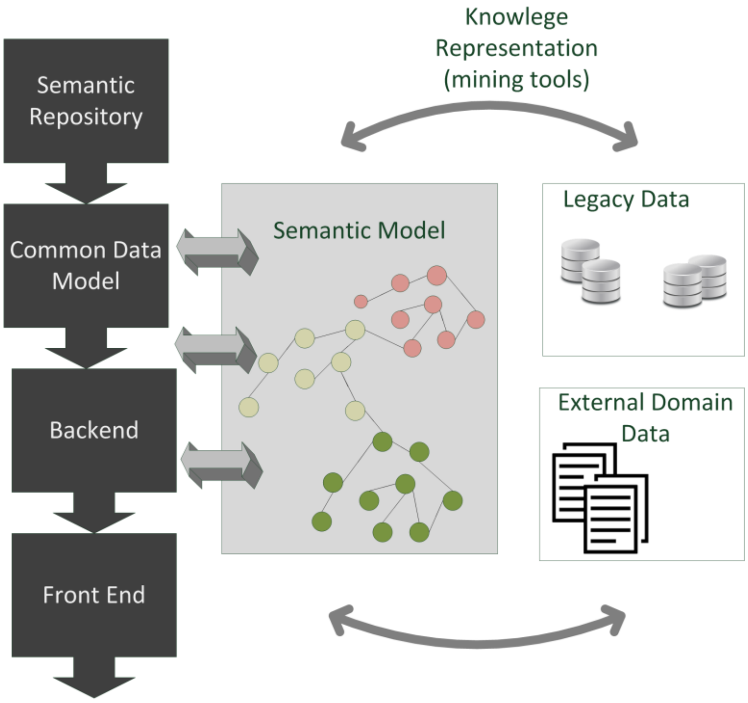


Fig. 2. THE KGSF pillars [source: own study]

Following the approach of recent researches within the factory domain (e.g. the project LinkedDesign [9], the project Virtual Factory Framework [10] and the related Virtual Factory Data Model [11], etc.), the reference model is represented through a set of ontologies by adopting the Semantic Web Technologies [12], which offers the possibility to represent formal semantics. Ontologies can be used in various ways to enhance the effectiveness of search engines. This topic has been deepened in the context of Information Retrieval, where for years researchers have experimented semantic indexing and knowledge representation tools such as thesauri and controlled vocabularies. In comparison to the latter, ontologies can lead to further benefits since they are richer and more formal. Firstly, ontologies allow to compose controlled queries and indexing vocabularies [13]. Secondly, they provide a rich axiomatization of the application domain and can be used to improve the accuracy of traditional search engines. Moreover, thanks to the expressiveness of the underlying semantic languages such as Resource Description Framework (RDF) [14] and Ontology Web Language (OWL) [15], the ontologies are expected to support properly the flexible and schema-less structure of the proposed graph-centric approach.

Finally, it is possible to re-use already existing tools to automatically infer and reason over the available data, thus deriving new knowledge about the concepts and their relationships, in addition to those already asserted and visualized in the graph.

3.2. Backend

Research in the field of information retrieval has traditionally focused more on refining information access rather than analyzing information to discover knowledge [16]. Information access aims at finding the right information for the right users at the right time with less attention on processing or transforming text data. This is the main goal of the data mining which in fact represents "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [17] through the application of specific algorithms for extracting hidden patterns (models) from data.
The overall process is interactive and involves various steps [18]:

- Understanding the domain.
- Collecting the raw data.
- Data pre-processing, which includes data cleaning in order to remove noise, and data transformation.
- Choosing the functions of data mining (clustering, prediction, classification, association, summarization, etc.) needed to be performed to derive the models.
- Storage of discovered knowledge for further reuse and integration in to the existing enterprise system.

In the KGSF the function of data mining is played by the Backend which has the goal to continuously process data streams, in order to extract and derive as much knowledge as possible through specific tools (Fig. 2), that combine and transform large amount of structured and unstructured legacy data [19] into semantic data that are then stored in the Semantic Repository. An information processing algorithm have to be implemented in order to extract large collections of candidate interrelated facts, that include a set of entities, their attributes, and their relations. Unfortunately, converting these candidate facts into useful knowledge represents an arduous challenge that requires to delve into research topics focusing on the problem of knowledge extraction from different sources, including named entity recognition, relation extraction and semantic role labeling.

3.3. Semantic Repository

The RDF is the standard model for data interchange on and off the Web and provides a general method to express data as lists of statements in the form of triples composed of subject-predicate-object. One of its strengths is that, in this format, any type of information can be virtually expressed. As the flip side, this expressivity implies the need to revise some classical data management problems, including efficient storage and query optimization. This is the reason why, as the request of semantic applications in real scenarios are continuously growing, the need to efficiently store and retrieve RDF data is getting more and more relevant [20]. This need has posed significant technological challenges for researchers and developers in terms of efficient and scalable semantic repositories. Several store solutions are currently available (e.g. Virtuoso, AllegroGraph, Stardog, etc.) and each of them may be more suitable in specific cases, depending on the requirements of the scenario.

Thus, a significant study during the plan of the KGSF regards the identification of a valid semantic repository capable to manage and reason on huge amount of Semantic data (also in the form of Big Data [21]) and capable to realize the analyses of intensive data in real-time. In fact this allows to collect and gather billions of real-time bytes of data on the organization resources, that are then processed instantaneously to optimize their utilization. A survey carried out by Modoni et al. [20] has shown how existing semantic repositories only partially support the mentioned capabilities in an effective way. These functionalities are particularly important in order to implement a scalable architecture and their lack represents a relevant technological gap to be addressed during the development of the KGSF.

3.4. Frontend

As thinking in terms of graphs opens new opportunities to allow users to visually navigate their data, another challenging aspect to be faced during the development of the KGSF consists in providing a usable Graphical User Interface (GUI) for exploring and editing the data in the knowledge graph in a natural and comfortable way. Its design, leveraged and driven by underlying ontology, has to be implemented to help users to formulate queries and express constraints for finding resources of the enterprise. The language SPARQL [22] can be used to compose structural queries on the semantic data. This approach may be feasible for experienced users, but are not practicable for users that are not familiar with Semantic Web technologies. To overcome this limitation, more intuitive ways of composing queries and showing results are needed.

A popular visualization of semantic data is based on the Big Fat Graph, which is useful for analyzing clouds of data, also evaluating their shape and density, but are not suitable for formulating flexible queries [23]. The approach on the basis of the KGSF Frontend exploits the technique of the Facets [24], which provides a specific classification of the information, allowing users to navigate a collection of information by applying multiple filters. In fact, a faceted system classifies each element along multiple explicit dimensions (facets), enabling the available items to be accessed in multiple ways rather than in a single and pre-determined order. Through the faceted GUI, the KGSF exposes both the implicit and the explicit interactions between graph entities, allowing knowledge path navigation and enabling the effective involvement of all the stakeholders across the enterprise value chain. Moreover, the proposed framework can provide explorative search suggestion helping users to create and develop new ideas and also discover serendipitous connections.

4. CONCLUSIONS

This paper has introduced a new interactive and explorative architecture for the search of the information within enterprises. Thanks to a semantic graph-based approach, the proposed solution is expected to offer several advantages regarding the quality of results and the usability to formulate the queries. However, the presented work is only a first step of a larger research agenda aiming at realizing a reference model of search engine, based on the Semantic Web technologies.

Future developments will address two main goals. First of all, the identification of a valid solution of repository is needed in order to store and handle large-scale of data processed by the KGSF. In this regard, an analysis of the performance of a set of the most widespread semantic repositories is currently under study, which will be followed by a proof of concept of the selected solution. The second goal will regard the choice of an appropriate data mining algorithm, which aims at extracting hidden knowledge from various sources of data. As many data mining algorithms have been proposed in literature, a survey of these will be conducted in order to select the most appropriate for the implementation of KGSF.

REFERENCES

[1] SINGHAL A.: *Introducing the Knowledge Graph: things, not strings.* Official Google Blog, http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html, May 2012.

[2] DREW O., CONSTINE J., TAYLOR C., LUNDEN I.: *Facebook Announces Its Third Pillar "Graph Search" That Gives You Answers, Not Links Like Google*., TechCrunch. AOL Tech, http://techcrunch.com/2013/01/15/facebook-announces-its-third-pillar-graph-search., 2013.

[3] SPIVACK N.: "http://www.novaspivack.com/technology/diagram-beyond-keyword-and-natural-language-search", 2007.

[4] *Enterprise Findability Without the Complexity*. White paper on Google Search Appliance, Google Inc., http://goo.gl/aFpSD0, 2008.

[5] GUHA R., MCCOOL R., MILLER E.: *Semantic search*. Proceedings of the 12th international conference on World Wide Web, ACM Press , 2003, pp. 700-709.

[6] KLOS S.., *The impact of ERP system on economic situation of enterprise: case study.* Applied Computer Science, Vol. 3, no 2, 2007, pp. 93-101.

[7] TRAN T., CIMIANO P., RUDOLPH S., STUDER R.: *Ontology-based interpretation of keywords for semantic search*. In ISWC/ASWC, 2007, pp. 523-536,.

[8] IEEE Standard Computer Dictionary, *A Compilation of IEEE Standard Computer Glossaries*. IEEE, 1990.

[9] MILICIC A., PERDIKAKIS A., EL KADIRI S., KIRITSIS D., TERZI S., FIORDI P., SADOCCO S.: *Specialization of a Fundamental Ontology for Manufacturing Product Lifecycle Applications: A Case Study for Lifecycle Cost Assessment*. OTM Workshops 2012, 2012, pp. 69-72.

[10] KADAR B., TERKAJ W., SACCO M.: *Semantic Virtual Factory supporting interoperable modelling and evaluation of production systems*. CIRP Annals – Manufacturing Technology 2013; 62(1), 2013, pp. 443-446.

[11] TERKAJ W., PEDRIELLI G., SACCO M.: *Virtual Factory Data Model*. Proceedings of the Workshop on Ontology and Semantic Web for Manufacturing, Graz, Austria, 2012, pp. 29-43.

[12] BERNERS-LEE T., HENDLER J., LASSILA O.: *The Semantic Web*, Scientific American, 284(5), 2001, pp. 34-43.

[13] MENA E., ILLARRAMENDI A., KASHYAP V., SHETH A.: *OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies*. Distrib. Parallel Databases 8(2), 2000, pp. 223-271.

[14] KLYNE G., CARROLL J.J.: *Resource Description Framework (RDF): Concepts and Abstract Syntax*. (W3C Recommendation 10 February 2004), World Wide Web Consortium, 2004.

[15] *W3C OWL 2 Web Ontology Language – Document Overview*. http://www.w3.org/TR/owl2-overview/, 2012.

[16] AGGARWAL C., ZHAI C.: *An introduction to text mining*, In Mining Text Data, Springer, 2012, pp. 1–10.

[17] FAYYAD U. M., PIATETSKY-SHAPIRO G., SMYTH P., UTHURUSWAMY R.: *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI/MIT Press, 1996.

[18] CHOUDHARY A.K., HARDING, J.A., TIWARI M.K: *Data mining in manufacturing: a review based on the kind of knowledge*. Journal of Intelligent Manufacturing, 20 (5), 2009 pp. 501-521.

[19] YUNYAO L., ZIYANG L., HUAIYU Z.: *Enterprise Search in the Big Data Era: Recent Developments and Open Challenges*. PVLDB 7(13), 2014, pp. 1717-1718.

[20] MODONI G., SACCO M., TERKAJ W.: *A survey of RDF store solutions*. Proceedings of the 20th International Conference on Engineering, Technology and Innovation, Bergamo, 2014.

[21] MANYIKA J., CHUI M., BROWN B., BUGHIN J., DOBBS R., ROXBURGH C., HUNG BYERS A.: *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.

[22] W3C: *SPARQL Query Language for RDF*. W3C Semantic Web Activity RDF Data Access Working Group, 2008.

[23] OBITKO M., VRBA P., MARIK V., RADAKOVIC M.: *The impacts of semantic technologies on industrial systems*. In 17th IFAC World Congress, Seoul, Korea, 2008, pp. 13880-13887.

[24] HEARST M.: *Design recommendations for hierarchical faceted search interfaces*, In: ACM SIGIR workshop on faceted search, 2006, pp. 1-5.