INSEARCH
A platform for Enterprise Semantic Search

Diego De Cao, Valerio Storch, Danilo Croce, and Roberto Basili

Department of Enterprise Engineering
University of Roma, Tor Vergata
00133 Roma, Italy
{decao,storch,croce,basili}@info.uniroma2.it

Abstract. This paper discusses the system targeted in the INSEARCH EU project. It embodies most of the state-of-the-art techniques for Enterprise Semantic Search: highly accurate lexical semantics, semantic web tools, collaborative knowledge management and personalization. An advanced information retrieval system has been developed integrating robust semantic technologies and industry-standard software architectures for proactive search as well as personalized domain-specific classification and ranking functionalities.

1 Introduction

Innovation is an unstructured process in most of Small and Medium Sized Enterprises (SMEs). The so called "Innovation Management Techniques", considered by the European Commission as an useful driver to improve competitiveness, are still underutilized by SMEs. Such techniques include Knowledge Management, Market Intelligence, Creativity Development, Innovation Project Management and Business Creation. However, within these techniques, the Creativity Development techniques are the less used among SMEs[1]. The only activity performed by almost all SMEs is the search for external information, in different sources such as the web, patent databases, in trade fairs or discussing with clients and partners. The main source of information for SMEs is the Internet search [7], an activity realized by more than 90% of SMEs when dealing with innovation. Knowledge and information are often distributed in heterogeneous and unstructured sources across networked systems and organizations. Search for entities (such as competitors or new products) is not always sufficient as search for knowledge, as the one related to novel processes or brands and marketing analysis (whereas connected to large scale opinion mining), is based upon richer information.

The system targeted in the INSEARCH EU project[2] embodies most of the ideas of the currently *en vogue* Enterprise Semantic Search technologies [4]. In

[1] European Commission, DG Enterprise Innovation management and the knowledge driven economy - January 2004

[2] FP7-SME-2010-1, Research for the benefit of specific groups, GA n. 262491

order to determine the core functionalities in the targeted system, an analysis involving 90 SMEs has been performed during the INSEARCH project to understand the process of searching within the innovation process. Most of the SMEs (92% of 90 interviewed SMEs) declared to make use of market and/or technology information when planning a technological innovation. Such informations are used to collect novel information for innovative ideas, performing prior art investigation, acquiring knowledge for technical planning or just gather inspiration and ideas. This search targets product and processes and it is mainly performed on scientific Web Sites and Competitors web site.

In these scenarios, keyword-based search related to product types and functions of the products are still used to retrieve information related to innovation processes. Search is mostly performed through iterative searches, evaluating search results through the very first lines of documents/web sites. Overall, the most requested knowledge extraction features are related to finding patterns within documents to propose possible innovation or customer requirements. This requirements are in line with the INSEARCH proposed approach of making usage of a TRIZ based methodology [1], to abstract functionalities from the specific innovation case under study and search for information through specific patterns (the TRIZ based Object-Action-Tool patterns) that could propose to SMEs possible technology innovations for the system under study.

In this paper the overall INSEARCH framework and its corresponding distributed system will be described, focusing on the advantage of integrating in a systematic fashion the benefits of analytical natural language processing tools, the adaptivity supported by inductive methods as well as the robustness characterizing advanced document management architectures built over interoperability standards in the Semantic Web (such as the iQser GIN Server). In the rest of the paper, section 2 discusses the different involved paradigms used to support semantic search. The overall architecture is presented in Section 3 that also show some typical user interactions with the system. Finally, section 4 derives the conclusions.

2 Integrating Ontological and Lexical Knowledge

2.1 Modeling Knowledge for Enterprise Semantic Search

Ontologies correspond to semantic data models that are shared across large user communities. The targeted enterprise or networked enterprises in INSEARCH are a typical expression of such communities where semantics can be produced, reused and validated in a shared (i.e. collaborative) manner. However, while knowledge representation languages are very useful to express machine readable models, the interactive and user-driven nature of most of the task focused by INSEARCH emphasize the role of natural language as the true user-friendly knowledge exchange language. Natural languages naturally support all the expressions used by producers and consumers of information and their own semantics is rich enough to provide strong basis for most of the meaningful inferences needed in INSEARCH. Document classification aiming at recognizing the interests of a user in accessing a text (e.g. a patent) requires a strongly linguistic basis as texts are mostly free and unstructured, as in [13]. In retrieval, against user queries, document ranking functions are inherently based on lexical preferences models, whose traditional TF-IDF models are just shallow surrogates. Moreover, the rich nature of the patterns targeted by INSEARCH (e.g. Object-Action-Tool triple foreseen by the TRIZ methodology) is strongly linguistic, as the same information is usually expressed in text with a huge freedom, and as for the language variability itself. Consider as an example that if a tool like a *packing machine* is adopted for the manufacturing of coffee boxes, several sentences can make reference to them, e.g. *packing machine applied to coffee*, *coffee is packed through dedicated machines* or *dedicated machines are used to pack small coffee boxes of 10 inch*.

Organizing knowledge through the SKOS concept scheme. Users are able to access, create or refine descriptions of a domain in the form of “tree of topics”, or simply topic-trees (modeled as SKOS [18] concept schemes) which will support their contextual search throughout the system. These topics act as collectors for documents which expose all those textual contents that can be naturally associated to their definition. They are under all aspects a controlled hierarchical vocabulary of tags offered to a community of users. Behind every tag a large term vocabulary is used in order to exploit the corresponding topic semantics during search activities. Topic-document associations may be discovered through information push by the mass: users inside a community contribute their bookmarks to the system. On the other hand, it can be achieved by the system itself, by machine learning from the above information, automatically creating topic associations for massive amount of documents which are gathered through the multichannel multimodal document discovery and acquisition component, as discussed in [13]. Examples of SKOS topic for the specific domain of the *Innovation Engineering* domain are reported in Fig. 1. Main SKOS concepts are `Research and Intellectual Properties` (organizing scientific papers or patents) and `Tecnology`. The latter can be specified with the concept `biotecnology` or `material` and so on. Apart from their role of document containers, topics may be described by enriching them with annotations, comments and multiple lexicalizations for the various languages supported by INSEARCH, so that their usage is informally clarified to human users, possibly enforcing their consistent adoption across the community.

User Management. In INSEARCH, standard models and technologies of the RDF [10] family have been adopted to allow each user to view his own SKOS ontology. It requires to model the information associated to user management, domain modeling and user data. The three different aspects have been physically modularized by partitioning the triples content, and each of these partitions is in turn divided into smaller segments to further account for specific data organization requirements such as provenance and access privileges. The partitions are obtained through the use of RDF named graphs, so that, whenever appropriate, the knowledge server may benefit of a single shared data space, or is able conversely to manage each partition (or set of partitions) as a separate dataset. The

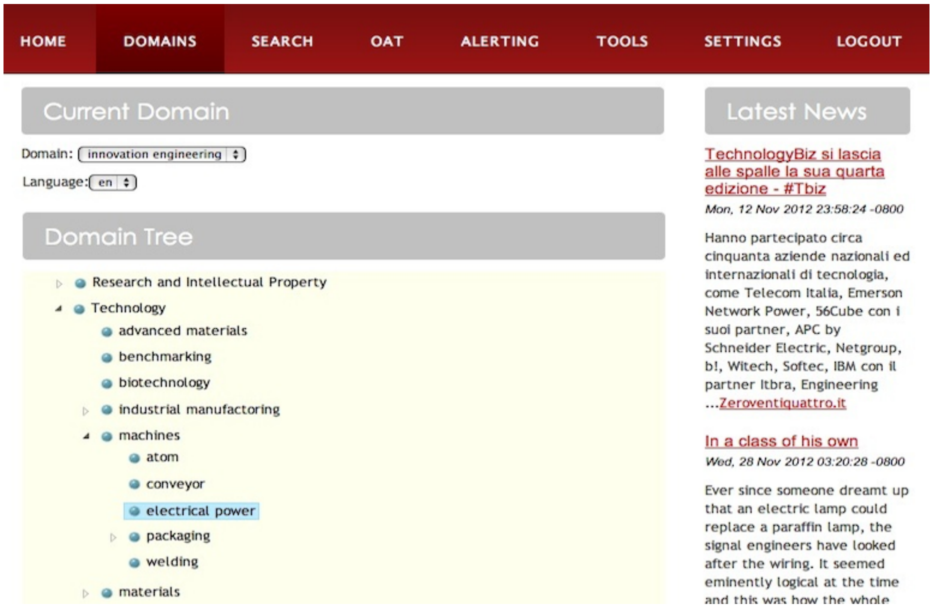

Fig. 1. SKOS topics and bookmarks in the *innovation* domain.

two main categories of users access these partitions in INSEARCH: companies and employees. Companies act like user-groups, collecting standard users (employees) under a common hat and possibly providing shared information spaces (e.g. domain models or reference information) which will be inherited by all of them. Each employee shares with his colleagues common data provided by the company, while at the same time he can be offered a personalized opportunity or a restricted access.

Semantic Bookmarking. In such a scenario, it is crucial to populate the SKOS ontology, thus providing examples for the document categorization process, allowing to link novel documents to existing (or user-defined) SKOS concepts. Semantic Turkey (ST) [14] was born as a tool for semantic bookmarking and annotation, thought for supporting people doing extensive searches on the web, and needing to keep track of: results found, queries performed and so on. Today ST is a fully fledged Semantic Platform for Knowledge Management and Acquisition supporting all of W3C standards for Knowledge Representation (i.e. RDF/RDFS/OWL SKOS and SKOS-XL extension). It is possible to extend it, in order to produce completely new applications based on the underlying knowledge services. The underlying framework allows access to RDF (and all modeling vocabularies already mentioned) through Java API, client/server AJAX communication (proprietary format, no Web service) and client-side Javascript API (hiding TCP/HTTP details). The ST offers among the others functionalities for editing a reference (domain) ontology (i.e. a SKOS-compliant topic taxonomy), bookmarking pages according to the taxonomy as well as organizing query results according to the hierarchical structure the SKOS taxonomy. Users may surf the web with a standards compliant web browser, associating information found on web documents to concepts from the current knowledge organization systems (KOS). The core framework of ST has been totally reused in INSEARCH without specific customization. However, novel dedicated services have been developed and plugged, flanking the main ones, to meet the specific INSEARCH requirements (see also the discussion in next section on architecture). In particular, the annotation mechanism is merged into the multiuser environment of the INSEARCH platform, so that the system may exploit contributions from different users, whenever the power of mass-contribution is exploitable.

2.2 Robust Modeling of Lexical Information

Computational models of natural language semantics have been traditionally based on symbolic logic representations naturally accounting for the meaning of sentences, through the notion of compositionality (as the Montague's approach in [12] or [3]). While formally well defined, logic-based approaches have limitations in the treatment of ambiguity, vagueness and other cognitive aspects such as uncertainty, intrinsically connected to natural language communication. These problems inspired recently research on **distributional models of lexical semantics** (e.g. Firth [8] or Schütze [15]). In line with Wittgenstein's later philosophy, these latter characterize lexical meanings in terms of their context of use [17]. Distributional models, as recently surveyed in [16], rely on the notion of Word Space, inspired by Information Retrieval, and manage semantic uncertainty through mathematical notion grounded in probability theory and linear algebra. Points in normed vector space represent semantic concepts, such as words or topics, and can be learned from corpora, in such a way that similar, or related, concepts are near to one another in the space. Methods for constructing representations for phrases or sentences through vector composition have recently received a wide attention in literature (e.g. [11]). While, vector-based models typically represent isolated words and ignore grammatical structure [16], the so-called **compositional distributional semantics** (DCS) has been recently introduced and still object of rich on-going research (e.g. [11, 5], [9], [2]). Notice that several applications, such as the one targeted by INSEARCH, are tight to structured concepts, that are more complex than simple words. An example are the TRIZ inspired Object-Action-Tool (OAT) triples that describe *Object*(s) that receive(s) an *Action* from *Tool*(s), such as those written in sentences like "… $[$*the coffee*$]_{Object}$ *in small quantities* $[$*is prepared*$]_{Action}$ *by the* $[$*packing machine itself*$]_{Tool}$ …" or "… *for* $[$*preparing*$]_{Action}$ $[$*the coffee*$]_{Object}$ *by extraction with* $[$*hot water*$]_{Tool}$, …".

Here physical entities (such as *coffee* or *hot water*) play the role of *Object*s or *Tool*s according to the textual contexts they are mentioned in. Compositional models based on distributional analysis provide lexical semantic information that is consistent both with the meaning assignment typical of human subjects to words and to their sentential or phrasal contexts. It should support synonymy and similarity judgments on phrases, rather than only on single words. The

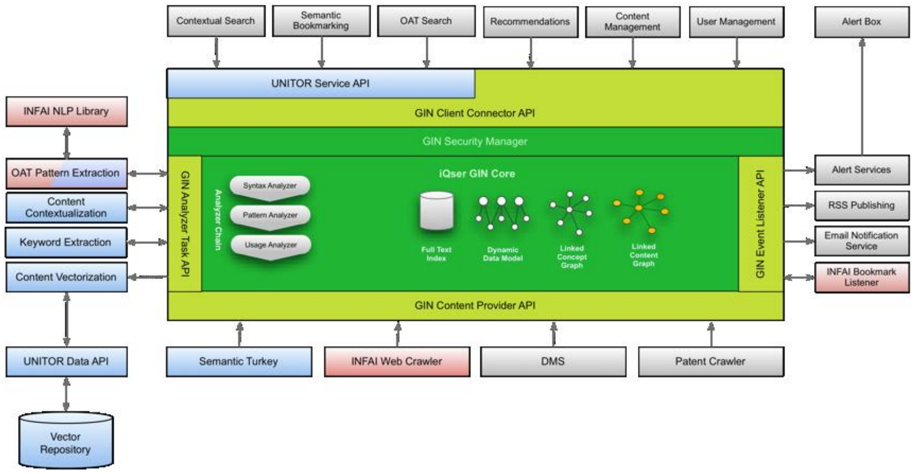

Fig. 2. An high level view of the INSEARCH functionalities and services.

objective should be assigning high values of similarity to expressions, such as "... *buy a car* ..." vs. "... *purchase an automobile* ...", while lower values to overlapping expressions such as "... *buy a car* ..." vs. "... *buying time* ...". Distributional compositional semantics methods provide models to define: (1) ways to represent lexical vectors $\boldsymbol{v}$ and $\boldsymbol{o}$, for words v, o occurring in a phrase (r, v, o) (where r is a syntactic relation, such as verb-direct_object), and (2) metrics for comparing different phrases according to the basic representations, i.e. the vectors $\boldsymbol{v}$, $\boldsymbol{o}$.

While a large literature already exist (e.g. [11]) the user can find more details about the solution adopted in INSEARCH in [2]. Compositional distributional semantic models are used to guide the user modeling of ontological concepts of interest (such as the SKOS topics), feed the document categorization process (that is sensitive to OAT patterns through vector based representation of their composition), concept spotting in text as well as query completion in INSEARCH. The adopted methods are discussed in [2] and [6].

3 The INSEARCH architecture

The INSEARCH overall architecture is designed as a set of interacting services whose overall logic is integrated within the iQser GIN Server for information ecosystems. The comprehensive logical view of the system is depicted in Fig. 2.

The core GIN services are in the main central box. External Analyzers are shown on the left, as they are responsible for text and language processing or, as in the case of the Content vectorization module, for the semantic enrichment of input documents. GIN specific APIs are responsible for interfacing heterogenous content providers and managing other specific data gathering processes (e.g. specific crawlers). Client Connector APIs are made available by GIN for a variety of user level functionalities, such as User Management, Semantic Bookmarking or Contextual searches that are managed via appropriate GIN interface(s). At the client level in fact, the basic search features from web sources and patents, are extended with:

- Navigation in linked search results and Recommendations for uploaded or pre-defined contents through bookmarks or SKOS topics of interest. Recommendations are strongly driven by the semantically linked content, established by the core analysis features of the GIN server.
- Semantic bookmarking is supported allowing sophisticated content management, including the upload of documents, the triggering of web crawling stages, the definition and lexicalization of interests, topics and concepts described in SKOS. Interesting information items are used for upgrading recommendations, topics and concepts and prepare contextual searches.
- Personalization allows user management functions at the granularity of companies as well as people.

On the backend side, we emphasize that the current server supports the integration with Alfresco[3] as the document and content management system, whereas the defined interests are also managed as Alfresco's content. While the integration of Web sources is already supported by a dedicated crawler, also patents are targeted with an interface to the patent content provider WIPO[4].

Contextual Semantic search is also supported through vector space methods. Vectorization is applied to incoming documents with an expansion of traditional bag-of-word models based on topic models and Latent Semantic Analysis (as discussed in Section 2.2). Moreover, the available vector semantics supports distributional compositional functions that model the representation and inferences regarding TRIZ-like OAT patterns, so that natural language processing and querying based on domain specific patterns are consistently realized. Basic feature extraction services and morphosyntactic analyzers (such as lemmatization and part of speech tagging) are already in place as external GIN analyzers.

The main functionalities currently integrated in INSEARCH are thus:

- **Website monitoring**: Observe changes in given pages/domains, which are added by the user and implemented as bookmarklets
- **Assisted Search**: such as in Query completion, e.g. support the user in the designing proper queries about company's products or markets .
- **Document analysis**: Intelligent Document Analysis is applied to asses their relevance to high-level topics predefined by the user in the SKOS taxonomy. Relevance to individual topics is provided through automatic classification driven by weighted membership scores of results with respect to individual topics.
- **Patent and scientific paper search**: Search for patents and/or scientific papers in existing databases (e.g. European patent office) is supported.
- **OAT-Pattern analysis**: TRIZ-inspired Object-Action-Tool (OAT) triples are searched in documents: these patterns play the role of suggestions for *tools*, which provide a certain function specified by the *object* and the *action*.

[3] `http://www.alfresco.com/`

[4] `http://www.wipo.int/portal/index.html.en`

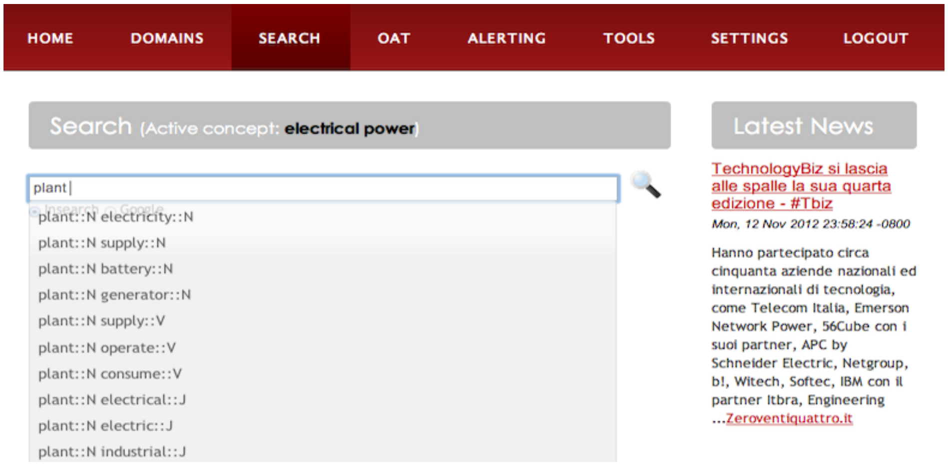

Fig. 3. The INSEARCH front-end and the completion of the Query *plant* when the SKOS concept `electrical power` is selected.

– **Adaptivity**: The system tracks user behaviors and adjusts incrementally its own relevance judgments for the topics and categories of interest.

3.1 Typical user interactions

The system has been recently deployed in its full functional version and provides a unique opportunity to evaluate its application to realistic data sets and industrial processes. The INSEARCH users will be able to quantitatively and qualitatively evaluate the impact of its semantic capabilities, its collaborative features as well as the overall usability of the personalized search environment in a systematic manner.

The front end of the INSEARCH system is shown in an interactive contextual search use-case in Fig. 3 and 4. The main tabs made available here are related to the Domains, Search, Alerting and Tools functionalities. In Domains the user can interact with and refine his own SKOS topics as well as interests and preferences, as shown in Fig. 1. Alerting supports the visualization of the results of Web Monitoring activities: here returned URLs, documents or other texts are conceptually organized around the SKOS concepts thanks to the automatic classification targeted to the ontology categories, made available through the Rocchio Classifiers, as discussed in [13]. In Tools most of the installation and configuration activities can be carried out.

In the Search tab, contextual search and query completion is offered to the user. In Fig. 3 the suggestions related to the ambiguous keyword "*plant*" early provided by the user are shown, where nouns like "*generator*" and "*battery*" (as well verbs like "*generator*" and "*battery*") are the proper continuation of the query, given the underlying domain, i.e `electrical power`. The completion is different when a topic such as `biotecnology` is selected, as shown in Fig. 4.

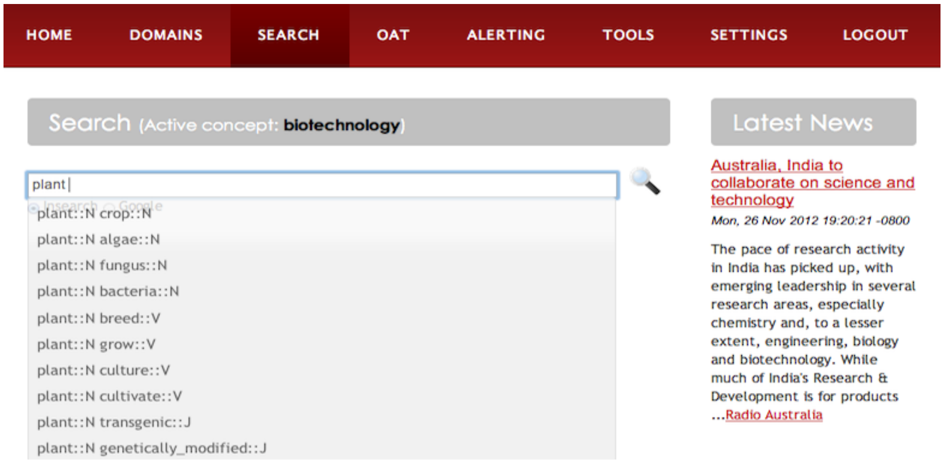

Fig. 4. The INSEARCH front-end and the completion of the Query *plant* when the SKOS concept `biotecnology` is selected.

The different completion is made available by the lexicalization of each concept: these lexical preferences are projected in an underlying Word Space (discussed in Section 2.2) that provides the geometrical representation of all words appearing in the indexed documents. Given the vectors representing all query terms and the lexical preferences of the selected SKOS concepts, the most similar (i.e. nearest) words are selected and proposed for the completion. This adaptivity is achieved also to provide novel information to the final users. In the front-end interface, a list of news is proposed. These are continually downloaded from the web and retrieved using the lexical preferences specified by the user during his own registration as well as the selected SKOS concepts. Notice that news are sensitive to the different SKOS concepts during the session, as in Fig. 3 and 4.

Once the query is submitted, documents are retrieved, automatically classified and clustered with respect to the existing SKOS concepts, as in Fig. 5. This clustering phase allows users to browse documents exploring their relatedness to specific SKOS concepts, such as `electrical power` or `research`. The user interface also allows to implement a *relevance feedback* strategy to improve the quality and adaptivity of text classifiers by simply clicking over the "thumbs up" or "thumbs down" icons. They allow to accept or reject each concept/document association, that reflects the underlying text classification. When the user accepts a classification, the Rocchio classifier associated with the corresponding concept is incrementally fed with the document, that becomes a positive example. On the contrary, the selected document is provided as a negative example, by clicking on the "thumbs down" icon.

Finally, the Object-Action-Tool (OAT) pattern-based search is shown in Fig. 6. The user is allowed to retrieve documents specifying specific actions (*pack*), objects (*coffee boxes*) or tools (*dedicated machine*). During the data-gathering

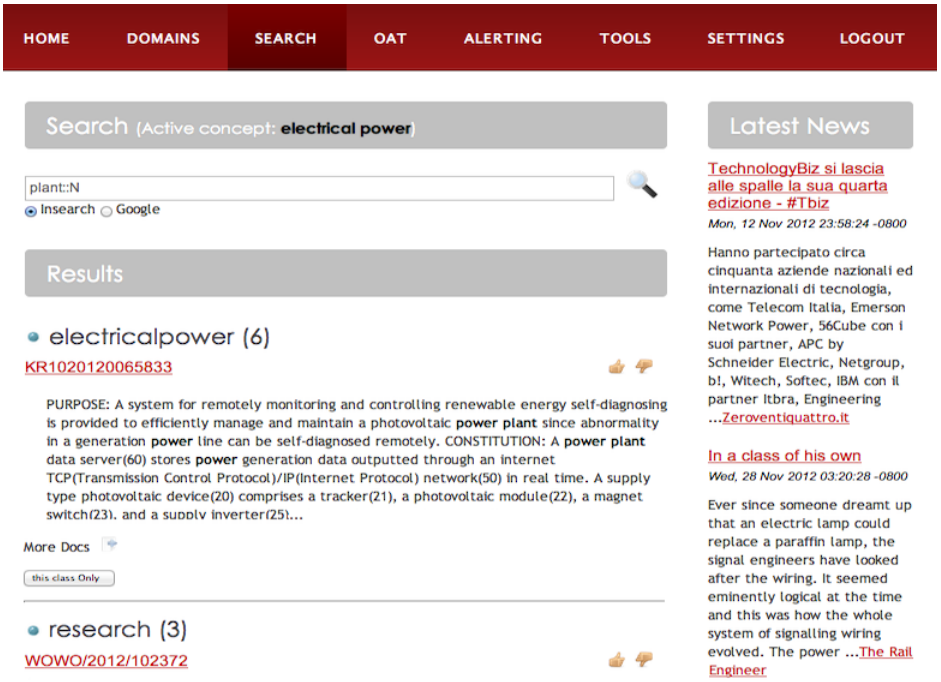

Fig. 5. The INSEARCH front-end the presentation schema of retrieved documents for the query *plant* when the SKOS concept `electrical power` is selected. Here 6 and 3 documents are related to the `electrical power` and `research` concept, respectively. The "Thumbs up"/"Thumbs down" icons allow to implement a relevance feedback strategy.

phase, the *OAT pattern extraction* module (see Fig. 2) extracts all patterns from the documents, by exploiting a set of pre-defined morphosyntactic patterns, such as SUBJECT-VERB-OBJECT. The extracted OAT patterns are used during the indexing phase, thus enabling semi-structured queries through (possible incomplete) OAT patterns. Fig. 6 summarizes a session where the user is interested in documents related to the action *control* and object *nuclear fission*. Initially the system suggests a set of possible tools, such as *method*, *system* or *product*. The user can select one or more tools to browse the related documents.

4 Conclusions

In the innovation process, the search of external information represents a crucial activity for the most of Small and Medium Sized Enterprises. In this paper the system targeted in the INSEARCH EU project is discussed. It embodies most of the state-of-the-art techniques for Enterprise Semantic Search: highly accurate lexical semantics, semantic web tools, collaborative knowledge management and personalization. The outcome is an advanced integration of analytical natural

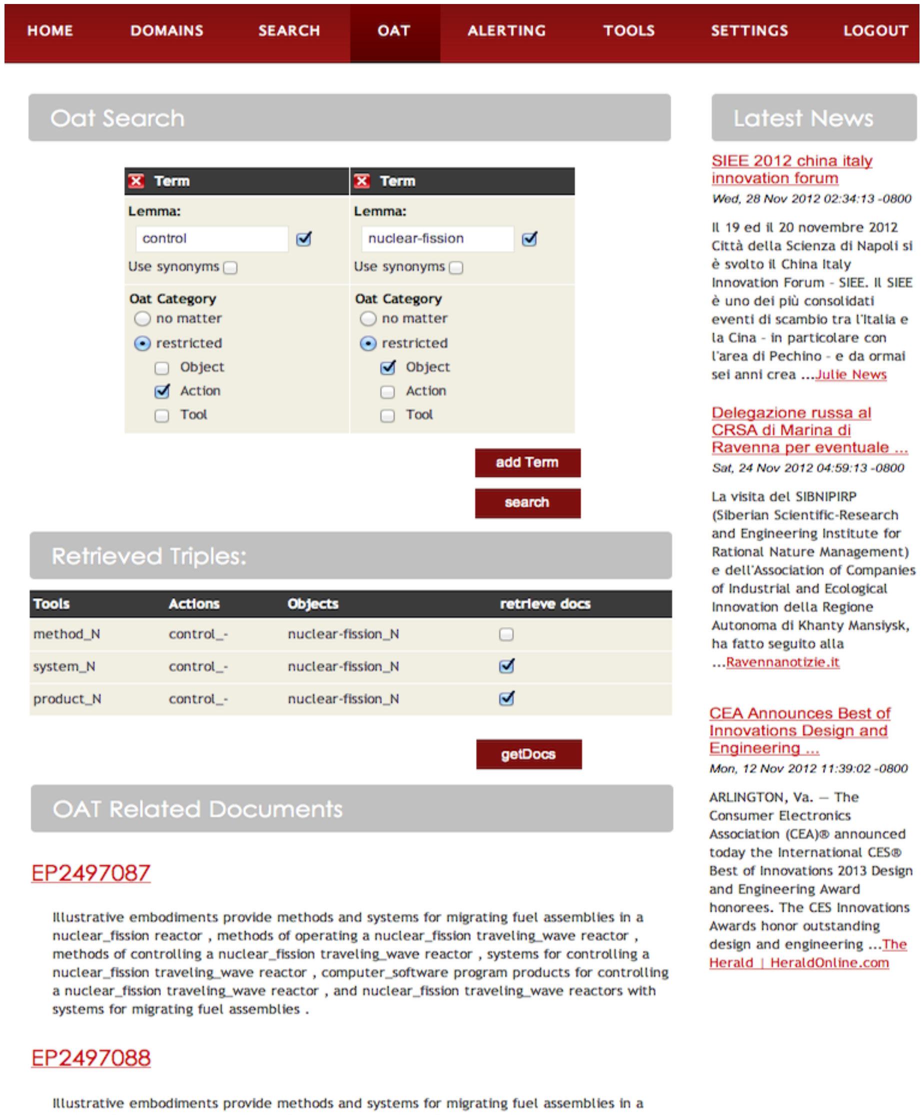

Fig. 6. The INSEARCH front-end for the Object-Action-Tool (OAT) triple-based search schema

language analysis tools, robust adaptive methods and semantic document management systems relying over the Semantic Web standards. The knowledge bases personalization as well as the semantic nature of the recommending functionalities (e.g. query completion, contextual search and Object-Action-Tool triple-based search) will be evaluated in systematic benchmarking activities, carried at the enterprise premises, within realistic and representative scenarios.

Acknowledgment The authors would like to thank all the partners of the INSEARCH consortium as they made this research possible. In particular, we thank Armando Stellato and Daniele Previtali from UNITOR, Jorg Wurzer from iQSer, Paolo Salvatore from CiaoTech, Sebastian Dunninger, Stefan Huber from Kusftein, Antje Schlaf from INFAI, Mirko Clavaresi from Innovation Engineering, Cesare Rapparini from ICA and Hank Koops from Compano.

References

1. Altshuller, G.: 40 principles, TRIZ keys to technical innovation. No. 1 in Triz tools, Technical Innovation Center, Worcester, Mass., 1. ed edn. (1998)
2. Annesi, P., Storch, V., Basili, R.: Space projections as distributional models for semantic composition. In: Gelbukh, A.F. (ed.) CICLing (1). LNCS, vol. 7181, pp. 323–335. Springer (2012)
3. B. Coecke, M.S., Clark, S.: Mathematical foundations for a compositional distributed model of meaning. Lambek Festschirft, Linguistic Analysis 36 (2010)
4. Baeza-Yates, R., Ciaramita, M., Mika, P., Zaragoza, H.: Towards semantic search. Natural Language and Information Systems pp. 4–11 (2008)
5. Baroni, M., Zamparelli, R.: Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In: Proceedings of EMNLP 2010. pp. 1183–1193. Stroudsburg, PA, USA (2010)
6. Basili, R., Giannone, C., De Cao, D.: Learning domain-specific framenets from texts. In: Proceedings of the ECAI Workshop on Ontology Learning and Population. ECAI, ECAI, Patras, Greece (July 2008)
7. Cocchi, L., Bohm, K.: Deliverable 2.2: Analysis of functional and market information. TECH-IT-EASY (2009)
8. Firth, J.: A synopsis of linguistic theory 1930-1955. In: Studies in Linguistic Analysis. Philological Society, Oxford (1957), reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.
9. Grefenstette, E., Sadrzadeh, M.: Experimental support for a categorical compositional distributional model of meaning. CoRR abs/1106.4058 (2011)
10. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax (2004)
11. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: In Proceedings of ACL-08: HLT. pp. 236–244 (2008)
12. Montague, R.: Formal Philosophy: Selected Papers of Richard Montague. Yale University Press (1974)
13. Moschitti, A., Basili, R.: Complex linguistic features for text classification: a comprehensive study. In: Proc. of the ECIR. pp. 181–196. Springer Verlag (2004)
14. Pazienza, M.T., Scarpato, N., Stellato, A., Turbati, A.: Semantic turkey: A browser-integrated environment for knowledge acquisition and management. Semantic Web journal 3(2) (2012)
15. Schütze, H.: Automatic Word Sense Discrimination. Computational Linguistics 24, 97–124 (1998)
16. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37, 141 (2010)
17. Wittgenstein, L.: Philosophical Investigations. Blackwells, Oxford (1953)
18. World Wide Web Consortium: SKOS Simple Knowledge Organization System Reference (Aug 2009)