2012 IEEE Sixth International Conference on Semantic Computing

Innovation-related Enterprise Semantic Search: the INSEARCH experience

Roberto Basili, Armando Stellato, Previtali Daniele
Dept. of Enterprise Engineering
University of Roma, Tor Vergata, Italy
Email: {basili,stellato,previtali}@info.uniroma2.it

Paolo Salvatore
CiaoTech,
Roma, Italy
P.Salvatore@ciaotech.com

Jorg Wurzer
iQSer
Bassersdorf, Schwitzerland
joerg.wurzer@iqser.net

***Abstract*—Innovation is a crucial process for enterprizes and pushes for strict requirements towards semantic technologies. Large scale and timely search processes on the Web are here often involved in different business analytics tasks. In the European project INSEARCH, an advanced information retrieval system has been developed integrating robust semantic technologies and industry-standard software architectures for Web monitoring and alerting, proactive search as well as personalized domain-specific classification and ranking functionalities.**

I. Introduction

In the current ICT scenarios, innovation in enterprise strictly depends on the timely access to knowledge and information. This is 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 sources for suitable search services are here large scale legacy repositories (such as internal DBs or product catalogues) as well as the open Web. The integration of structured as well as unstructured sources is challenging for most search technologies. On the one side, structured data require fine-grain semantic models in order to be flexibly used in uncertain tasks such as retrieval of partially specified data or ranking. On the other hand the shallow semantics of most IR technologies do not allow to work at the proper granularity level, as required by domain specific aspects (e.g. terminology) or personalization (e.g. specific semantic phenomena such as preference, novelty or reliability of the sources). As a consequence, integrating the flexibility of highly lexicalized models with the precision of knowledge-based technologies can be a very challenging task whose balanced optimization is still an open research issue. The system targeted in the INSEARCH EU project[1] embodies most of the ideas of the currently *en vogue* Semantic Enterprise Search technologies [1], with the specific 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). The overall INSEARCH framework and its corresponding distributed system will be shortly described in this paper. Section II will introduce the motivations and major requirements of the system, as derived from the market analysis carried out in the project. Section III discuss the different paradigms, i.e. knowledge representation and vector models for lexical semantics, used to support semantic search in the system. The overall architecture is finally presented in Section IV that also show some typical user interactions with the system.

II. Search for Innovation: the Enterprise view

Innovation is an unstructured process in most of Small and Medium Sized enterprises. The so called "Innovation Management Techniques", considered by the European Commission as an useful driver to improve competitiveness, are still underutilized by SMEs; in particular, among such techniques (which include knowledge management, market intelligence, creativity development, innovation project management, business creation, etc.) the Creativity Development Techniques are the less used among SMEs[2]

In the innovation process, the only activity that almost all SMEs perform is to 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 [2], which is an activity realised by more than 90% of SMEs when dealing with innovation.

During INSEARCH, an analysis involving 90 SMEs have been performed to understand the process of searching and using information during the innovation process of SMEs.

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 information are used to sought information for innovative ideas, performing prior art investigation, acquiring knowledge for technical planning or just gather inspiration and ideas. The mostly sought

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

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

978-0-7695-4859-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ICSC.2012.61

information are about product and processes, performed on scientific Web Sites and Competitors web site. Papers and scientific publications are in fact usually sought while performing innovation processes by SMEs of all dimensions and sectors, while Patent analysis is mostly of interest of manufacturing SMEs. Patent analysis is mostly realised through Espacenet, using patent classification as the most used feature during patent search. The function of the product, and functions of components / subcomponents are the most used keywords by SMEs in performing patent search. 91% of the interviewed SMEs stated that they make use of Google or similar general search engine (such as Yahoo). While performing searches for information related to innovation processes, they use keywords related to product types and functions of the products. Search is mostly performed through iterative searches, evaluating search results through the very first lines of documents/web sites

In carrying out searches, most of SMEs use bookmarks in their browsers as the main way to check/monitor interesting web sites during innovation/market analysis processes. SMEs are interested in having an IT system supporting the process of finding and filtering relevant information on the web during innovation. During the analysis performed, the following main functionalities have been deemed as important by SMEs (ranked 4 or 5 in a 0 to 5 scale of importance):

- Monitoring web sites of interest,
- Supporting the systematic definition of the set of keywords for searching of product and market information of their specific interest
- Crawling the web to suggest interesting web sites, find possible interesting documents and automatic download
- Ranking documents in order of importance
- Filter patents in order of relevance to specific innovation
- Finding specific patterns, as for example: finding any "*thing*" that performs a certain "*action*" on an "*object*"

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 [3], 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 SEMs possible technology innovations for the system under study.

Overall, the challenge for the IT system to support SMEs requirements is:

- Allow SMEs to apply the Open Innovation main concepts of locating external knowledge to find innovative solutions.
- Being able to filter documents and find relevant patterns following the approaches of structured methodologies such as TRIZ to be able to identify possible innovations in sectors (industrial/market and technology sectors) that are not the same sectors where the system under study is operating. This implies the capability of human and the system to abstract the innovation problem from the system under study and find solution in different spaces of information that are not restricted to the specific sector in which the company is operating in.

The above survey allows to trace the following findings:

- A Timely and accurate access is crucial to innovation practices
- Currently organisations depend on robust autonomous filtering and classification capabilities
- Innovation-related search requires high level abstractions (as the TRIZ models of the innovation processes suggest)
- A proactive role of the search system is important as user is interested in capitalizing its needs and typical searches, and much less into just interactive search
- Personalization is important as experts findings emerge as a set of personal behaviors, knowledge and information
- Interoperability is important in the open Web world

III. Semantic Search: Integrating Ontological and Lexical Knowledge

A. Semantics in Web data and 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) require a strongly linguistic basis as texts are mostly free and unstructured ([4]). 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, as for the language variability itself. As an example, 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*, *dedicated machines are used to pack small coffee boxes of 10 inch*, ... Finally, user interests cannot be captured outside language. Infact, if a user has to express them, he will make it linguistically, through definitions, glosses or lexical expressions (see for example, the widespread use of tags in user generated contents scenarios of YouTube or

Flickr). More formal definitions, such as profiles described in KR languages, can also be used. However, this rises again the issue of matching these formalized profiles in texts, that in turn evokes the linguistic task of matching symbolic patterns in free texts, traditionally referred to as *Information Extraction* through *text understanding*.

We will see later that the INSEARCH solution to the above problems stands in the integration of ontological knowledge (i.e. information expressed through the KR standards of RDF or OWL) with strongly lexicalized meaning representations, i.e. distributional models of the lexicons ([5], [6] or [7]). Vector models, widely used in Information Retrieval, are here used to augment KR languages, as for example in the lexical description of some concepts (such as SKOS-like topic categories or domain concepts, e.g. Actions and Tools), useful to drive statistical inferences during document classification or ranking.

B. Knowledge Modeling in INSEARCH

In INSEARCH, standard models and technologies of the RDF [8] family have been adopted 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.

User Management. The two main categories of users in INSEARCH are: 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, reference information etc..) 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.

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 [9] 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 two main workflows:

1) Information push by the mass. Users inside a community contribute their bookmarks to the system
2) The system, by machine learning from the above information, automatically creates topic associations for massive amount of documents which are ingested through the multichannel multimodal document discovery and acquisition component ([4]).

Examples of SKOS topic for the specific domain of the coffee packing machines is reported in Fig. 4. 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.

C. Semantic Bookmarking technologies in INSEARCH

Semantic Turkey (ST) [10] was born as a tool for semantic bookmarking/annotation, thought for supporting people doing extensive searches on the web, and needing to keep track of: results found, queries performed and so on (see Fig. 1). 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 offer among the others functionalities for editing a reference (domain) ontology (i.e. a SKOS-compliant topic taxonomy as in Fig. 1), 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 with concepts from the current knowledge organisation systems (KOS). The utility of this association is twofold: KOS developers may document a concept by attaching a set of web resources to it, whereas a KOS consumer may categorize information resources tagging them with concepts from the KOS. The nature of the association may also vary: the editor supports both the bookmarking of web pages as a whole and the annotation of portions of text. In the first case, the bookmarked page's metadata are stored together with the link to a `skos:Concept` through the `dcterms:subject` property. In the second case, the annotation of specific portions of text is instead triggered by drag'n'drop actions performed by the user: when a portion of text is selected, dragged and finally dropped over a concept in the tree, several options are presented to the user. The available options depend on the nature of the RDF resource where the text has been dropped on (i.e. classes or instances in the case of OWL, concepts for SKOS). A flow of actions is performed when information is dropped on a `skos:Concept`. First, the user is prompted with a dialog window listing the set of available options, namely:

1) *add an annotation* to the selected concept,
2) *create a new concept* (and annotate it),

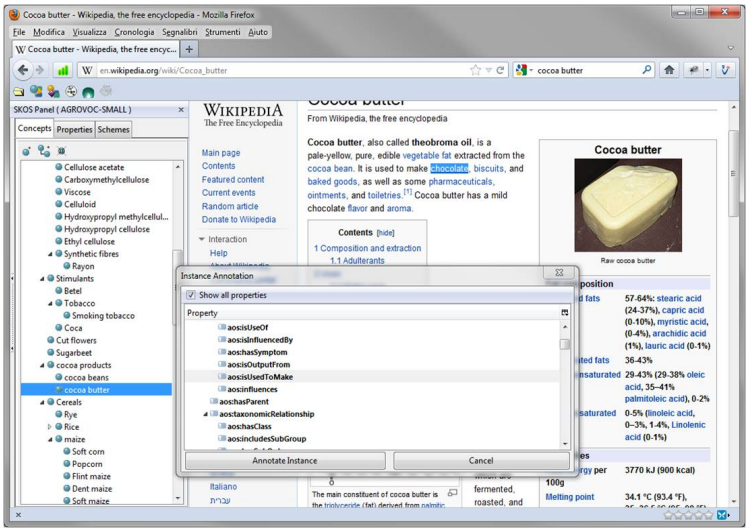

Fig. 1. Ontology-driven semantic bookmarking of a Wikipedia page in Semantic Turkey.

3) *add a new value for a property* of the concept.

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. We achieve thus the potential impact of a large scale collaborative annotation platform. For instance, associations of interest between pages and topics can be imported towards a new user (if a given topic tree is being readopted). This provides a bootstrapping mechanism for shared knowledge. On the contrary, a novel user can decide to start with an empty system, whose adaptation is totally under his control through interactions and learning. This mixed approach solves the *cold start problems* of this class of complex systems, with users that can soon experience how much the system "fits" their expectations and then progressively fill it with their own preferences and contents.

At the same time, security and privacy issues are also of concern in INSEARCH. The above contents, when suitable to be contributed to mass data exploitation, will be used in an anonymous way. Conversely, an authentication based access to data is exploited to ensure that data privacy holds in its proper domain, whether it is information dictated by a company (similar to a user-group) to be shared by all of its employees, or secured at the level of each individual user.

Finally, a specific User Interface for the ST-like interaction has been shipped for INSEARCH. A variety of widgets for data management and search to be hosted on the main INSEARCH portal has been realized. Regarding the annotation and bookmarking functionalities, we recognized that an "on-site" solution would have resulted in an unfriendly solution for the user, requiring them to work from inside the INSEARCH portal to perform tasks which are instead naturally associated to a free browsing and navigation attitude. For this reason, specific INSEARCH bookmarklets have been devised to let the user download them once from the INSEARCH portal, and install them on their favorite browser. These bookmarklets can then be clicked when a user visits a page, to invoke the annotation functions over the visited web page, thus completely skipping the visit to the portal.

D. Robust Modeling of Lexical Information: Word Spaces, Latent Semantics and Compositionality

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 [11] or [12]). 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 [13] or Schütze [14]). In line with Wittgenstein's later philosophy, these latter characterize lexical meanings in terms of their context of use [15]. Distributional models, as recently surveyed in [6], 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. The distance between two points (via angular or Euclidean metrics) represents semantic dissimilarity between concepts. Methods for constructing representations for

phrases or sentences through vector composition has recently received a wide attention in literature (e.g. [16]). While, vector-based models typically represent isolated words and ignore grammatical structure [6], the so-called **compositional distributional semantics** (DCS) has been recently introduced and still object of rich on-going research (e.g. [16], [17], [18], [19]) 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]$_{Tool}$...

... 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 objective should be high values of similarity between expressions, such as "... *buy a car* ..." vs. "... *purchase an automobile* ...", while lower values for overlapping expressions such as "... *buy a car* ..." vs. "... *buying time* ...". This is a stringer benefit as a computational model for entailment, so that the representation for " ... *buying something* ..." is still implied by the expression "... *buying a car* ..." but not by "... *buying time* ...". Distributional compositional semantics methods provide models to define: (1) ways to represent lexical vectors $\vec{v}$ and $\vec{o}$, 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 selected representations, i.e. the vectors $\vec{v}$, $\vec{o}$.

While a large literature already exist (e.g. [16]) the user can find more details about the solution adopted in INSEARCH in [19]. Compositional distributional semantic models are used in INSEARCH to guide the user modeling of ontological concepts of interest (such as the SKOS topic), feeed 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 [19] and [7].

IV. The INSEARCH architecture

The INSEARCH overall architecture is deigned 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 ingestion 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 infact, 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 that allows 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 of Web sources is already supported by an own crawler, also patents are targeted with a native interface to a patent content provider. The integration with the Semantic Turkey supports a push synchronization and event listening to add Web pages of interest to the GIN Server repository.

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 III-D): relevant words, terms and expressions are added even if they do not occur in documents, according to vector semantics. This will support scalability and personalization, as lexical vectors are available to better focus topics, preferences and contexts of interest of the individual user types. A large term vocabulary is used as a lexical interface for every topic. This allows to exploit the corresponding semantics during search activities. A so-called contextualization analyzer and a related search service are already available for the client.

Moreover, the available vector semantics will support distributional compositional functions that model the representation

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

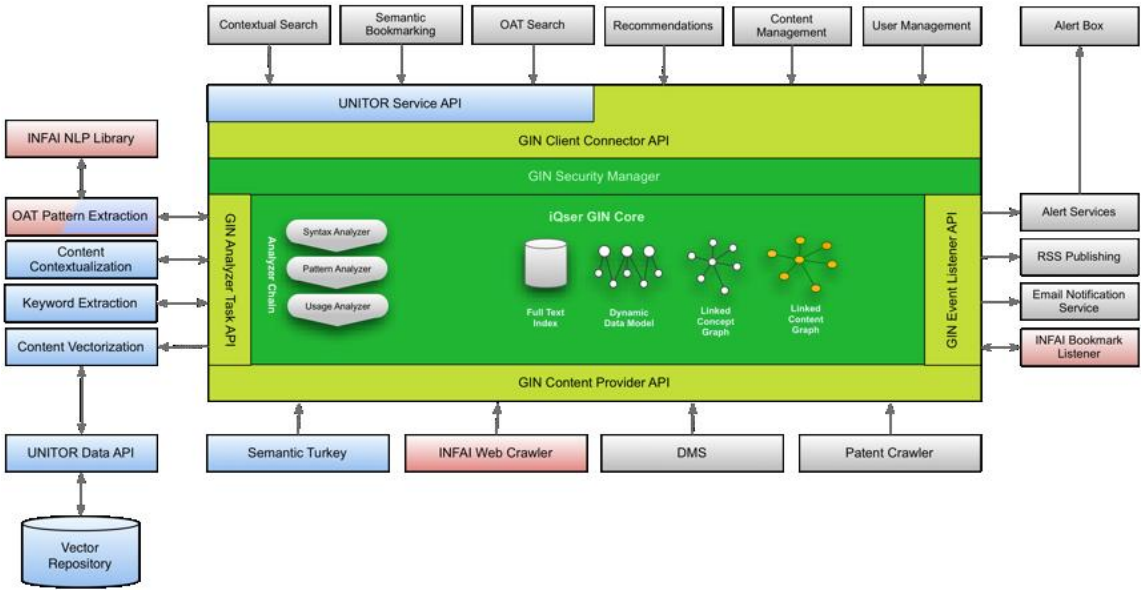

Fig. 2. An high level view of the INSEARCH functionalities and services.

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 such as lemmatization, part of speech tagging and semantic classification 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 .
- **Personalized Web Crawling**: While user defines interesting websites as seeding points, this triggers crawling towards information related to important research topics.
- **Document analysis** Intelligent Document Analysis is applied to asses their relevance to high-level topics predefined by the user. Relevance to individual topics is provided through automatic classification that provide weighted membership scores of individual query/search results to the topics organized as a SKOS taxonomy. The result is a taxonomical organization of retrieved documents according to relevance judgements automatically assigned.
- **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 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*. The INSEARCH system locates OAT-patterns in relevant documents and offers OAT-oriented querying and browsing functionalities.
- **Adaptivity**. The system tracks user behaviors and adjusts incrementally its own relevance judgments for the topics and categories of interest.

A. Knowledge Management Server

In INSEARCH, the core framework of Semantic Turkey has been flanked by dedicated services written specifically for INSEARCH. A service layer for user management has been totally implemented from scratch, and multiuser support has been accounted in many of the traditional ST services. All of these new or customized services are still compatible with (and are based on) the core architecture of Semantic Turkey, hence these have been dynamically added through the ST's extensible service mechanism based on the Open Service Gateway standard OSGi [20].

In Fig. 3, the front end of the INSEARCH system is shown when it is used for an interactive contextual search. The main tabs made available here are related to the "Personalization", "Search", "Alerting" and "Tools" functionalities. In "*Personalization*" the user can interact with and refine his own SKOS topics as well as interests and preferences. "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 abilities targeted to the ontology categories. An hint on the topic hierarchy made available for the ICA "*coffee packaging*" domain is in Fig. 4. In "Tools" most of the installation and configuration activities can be carried out. A very interesting tab is certainly "Search" where contextual search and query completion is offered to the user: in Fig. 3

the suggestions related to the keyword early digited by the user are shown, where "*foodstuff*" as the proper continuation of the "*packaging foodstuff*" query is automatically proposed by the user, given the underlying domain, i.e "*coffee packaging*".

This mechanism is very interestingly based on the notion of current "*context*" of a query. The automatic expansion infact depends on a set of *active SKOS concepts*, that semantically characterize the user focus, here called the "*context*". When the context is changed (i.e. another SKOS hierarchy is selected, or another set of SKOS classes is activated by the user) then suggestions associated to a given word/phrase also change accordingly. In Fig. 3, the "*foodstuff*" keyword is suggested as a side effect of the activation of the INDUSTRIAL PRODUCT concept in the hierarchy, as shown in Fig. 4. In Fig. 4 in fact the currently developed taxonomy for the SKOS topics related to one of project case studies (i.e. the *coffee packaging* domain) is reported.

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 final stage of the project has been planned to support these crucial assessment activities, to which we also look forward as part of our near future research.

V. CONCLUSIONS

In the innovation process, the main activity that most SMEs perform is searching for external information. Their main source of information is the Internet [2], which is the activity more than 90% of SMEs carry out when dealing with innovation. The system targeted in the INSEARCH EU project embodies most of the state-of-the-art techniques for Semantic Enterprise: highly accurate lexical semantics, semantic web tools, collaborative knowledge management and personalization. The outcome is an advanced integration of analytical natural language analysis tools, robust adaptive methods as well as semantic document management over the Semantic Web standards. In the paper, we discussed how an advanced software architecture has been extended to host most of these advanced knowledge management methods. The personalization of knowledge bases as well as the semantic nature of the recommending functionalities (e.g. query completion or contextual search) will be explored in the near future in systematic benchmarking activities that will be 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 the research discussed in this paper possible. In particular, we thank Sebastian Dunninger, Stefan Huber from Kusftein, Antje Schlaf from INFAI, Mirko Clavaresi from Innovation Engineering, Diego De Cao and Valerio Storch from UNITOR, Cesare Rapparini from ICA and Hank Koops from Compano.

REFERENCES

[1] R. Baeza-Yates, M. Ciaramita, P. Mika, and H. Zaragoza, "Towards semantic search," *Natural Language and Information Systems*, pp. 4–11, 2008.

[2] L. Cocchi and K. Bohm, "Deliverable 2.2: Analysis of functional and market information," *TECH-IT-EASY*, 2009.

[3] G. Altshuller, *40 principles, TRIZ keys to technical innovation*, 1st ed., ser. Triz tools, L. Shulyak and U. Fedoseev, Eds. Worcester, Mass.: Technical Innovation Center, 1998, no. 1.

[4] A. Moschitti and R. Basili, "Complex linguistic features for text classification: a comprehensive study," in *Proc. of the 26th European Conference on Information Retrieval (ECIR)*. Springer Verlag, 2004, pp. 181–196.

[5] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological review*, pp. 211–240, 1997.

[6] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, p. 141, 2010. [Online]. Available: doi:10.1613/jair.2934

[7] R. Basili, C. Giannone, and D. De Cao, "Learning domain-specific framenets from texts," in *Proceedings of the ECAI Workshop on Ontology Learning and Population*, ECAI. Patras, Greece: ECAI, July 2008. [Online]. Available: http://olp.dfki.de/olp3/Basili.pdf

[8] G. Klyne and J. J. Carroll, *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C Recommendation, 2004.

[9] *SKOS Simple Knowledge Organization System Reference*, World Wide Web Consortium, Aug. 2009.

[10] M. T. Pazienza, N. Scarpato, A. Stellato, and A. Turbati, "Semantic turkey: A browser-integrated environment for knowledge acquisition and management," *Semantic Web journal*, vol. 3, no. 2, 2012.

[11] R. Montague, *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, 1974.

[12] M. S. B. Coecke and S. Clark, "Mathematical foundations for a compositional distributed model of meaning," *Lambek Festschirft, Linguistic Analysis, vol. 36*, vol. 36, 2010.

[13] J. Firth, "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.

[14] H. Schütze, "Automatic Word Sense Discrimination," *Computational Linguistics*, vol. 24, pp. 97–124, 1998.

[15] L. Wittgenstein, *Philosophical Investigations*. Oxford: Blackwells, 1953.

[16] J. Mitchell and M. Lapata, "Vector-based models of semantic composition," in *In Proceedings of ACL-08: HLT*, 2008, pp. 236–244.

[17] M. Baroni and R. Zamparelli, "Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space," in *Proceedings of EMNLP 2010*, Stroudsburg, PA, USA, 2010, pp. 1183–1193.

[18] E. Grefenstette and M. Sadrzadeh, "Experimental support for a categorical compositional distributional model of meaning," *CoRR*, vol. abs/1106.4058, 2011.

[19] P. Annesi, V. Storch, and R. Basili, "Space projections as distributional models for semantic composition." in *CICLing (1)*, ser. LNCS, A. F. Gelbukh, Ed., vol. 7181. Springer, 2012, pp. 323–335.

[20] "Osgi bundle repository specification," 2005. [Online]. Available: http://www2.osgi.org/Download/File?url=/download/rfc-0112_BundleRepository.pdf

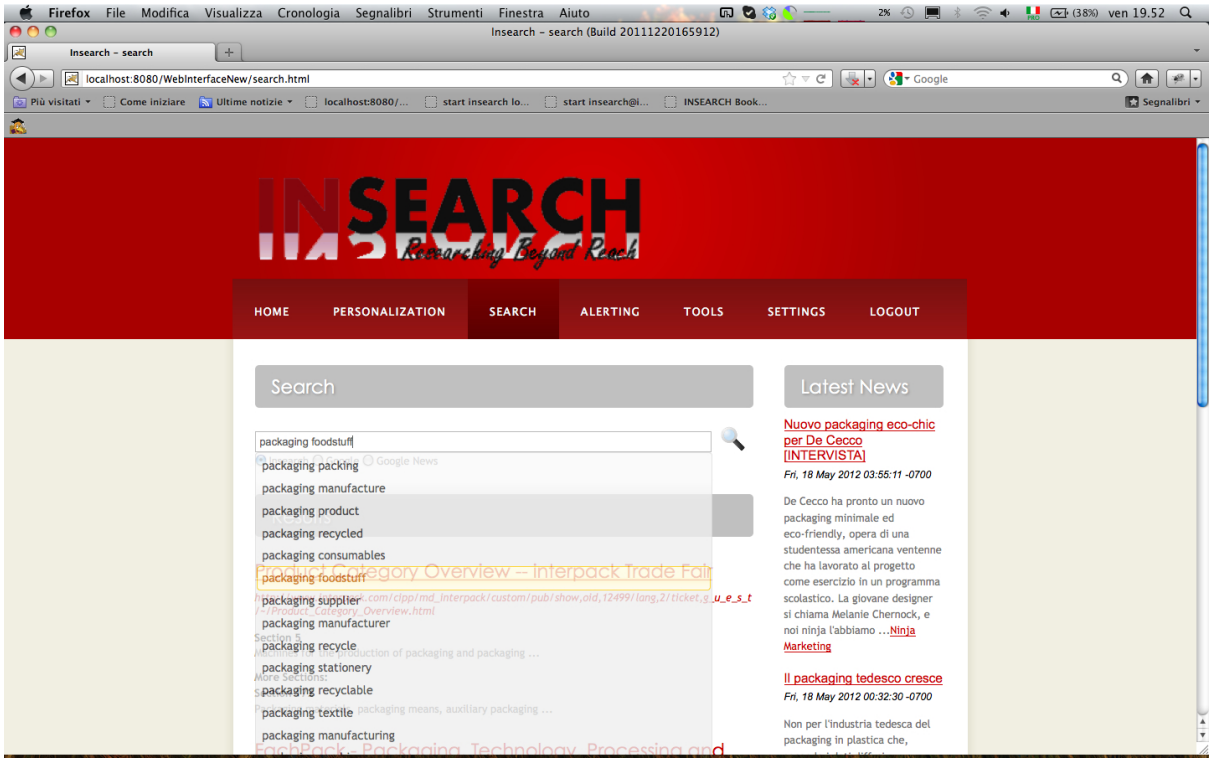

Fig. 3. The INSEARCH front-end and the completion of the Query *packaging*.

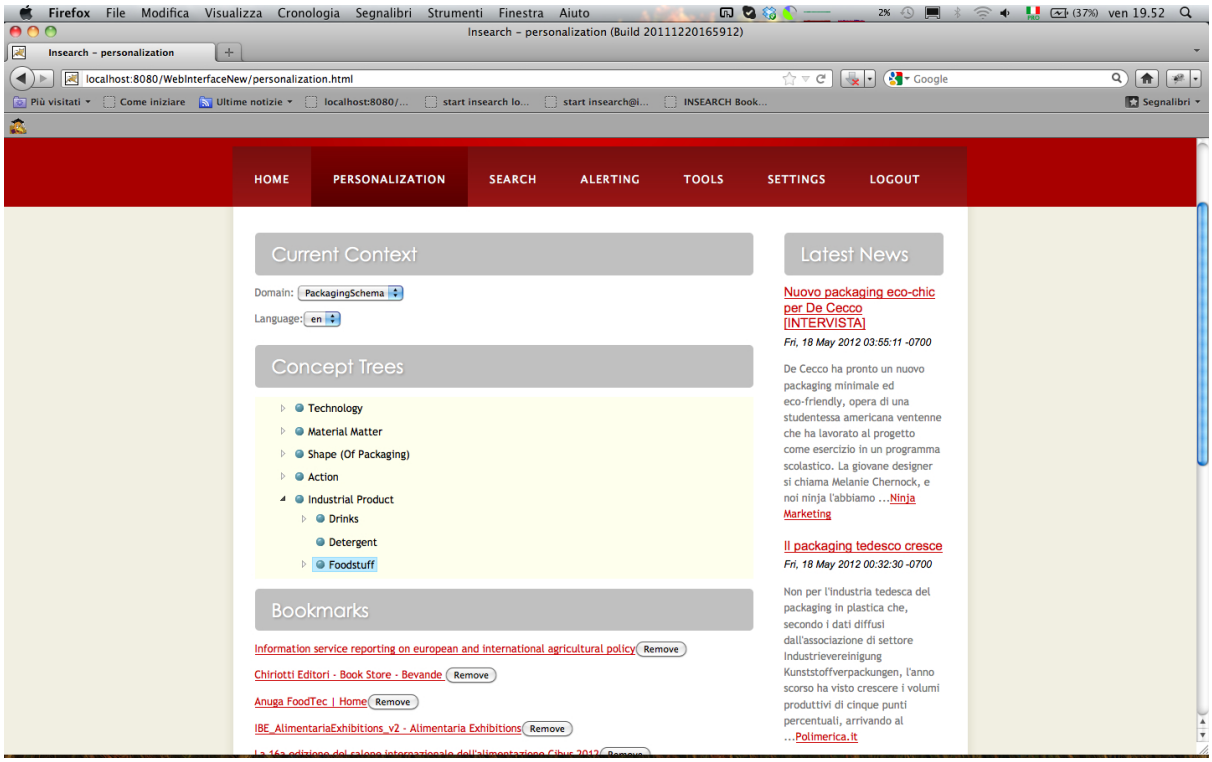

Fig. 4. SKOS topics and bookmarks in the *coffee packaging* domain.