A SEMANTIC SEARCH ENGINE FOR A BUSINESS NETWORK
A Personalized Vision of the Web applied to a Business Network

Angioni Manuela, Emanuela De Vita, Lai Cristian, Marcialis Ivan, Paddeu Gavino and Tuveri Franco
CRS4, Center of Advanced Studies, Research and Develepment in Sardinia, Parco Scientifico e Tecnologico Ed.109010 Pula (CA), Italy

Keywords: NLP, Text Categorization, User Profiling, Semantic Search Engine.

Abstract: The Web's evolution during the last few years shows that the advantages from the users' point of view are not so macroscopic. Despite information is still the primal element, is ever more evident the need to redefine the information paradigm so that the net and the information can become really user-centric by an inverse process that brings the information to the user and not more the user to information. Define new tools is needed to create a privileged window of observation on information and knowledge: each user with his specific interest. Not more a single available space of information but shared data for everyone. What each user needs is a specific private space of information according to his point of view, his way to classify and manage the information, related to his network of contacts in the way each person choose to live the Web, the net and the knowledge. In this paper we illustrate a part of a project named *A Semantic Search Engine for a Business Network* where the introduction of Natural Languages, user profiling, automatic information classification according to users' personal schemas will contribute to redefine the vision of information and delineate processes of Human-Machine Interaction.

1 INTRODUCTION

The way of interacting and the modality to access the information is continuously changing. It is going more and more toward tools able to follow and assist the user in its networking activities through the use of technologies related to natural languages, the classification of the information and the user profile (Marcialis and De Vita, 2008). In this scenario the changes carried out by the great innovators in the field of information processing are emerging. Google is still the frontier of search engines, but several efforts have been completed in order to exceed it, such as Bing, who has obtained good results regarding search suggestions and research results with natural language.

Several attempts to reduce the time consuming of online searches have been proposed and tested through meta search engines that simultaneously search on more search engines or with new features specialized in searching on social network (Mislove, 2006).

The introduction of the query in natural language is a common element that is already prefiguring the advent of the Web 3.0 with tools such as the computational knowledge engine Wolfram Alpha, able to answer queries by means of a vast repository of data organized with the help of sophisticated Natural Language Processing algorithms or Aardvark (Horowits and Kamvar, 2010) that allows users, experts on certain topics, to answer to queries made by other users in a more efficient mechanism for online search. Another example is Twine (Wissner and Spivack, 2009), able to improve the relevance of results by means of filters that try to reduce the noise due to less relevant answers.

The passage from the unstructured to the structured information through the use of ontologies has not produced the expected innovation in search engines due to the lack of tagged resources.

New tools able to reduce or even to eliminate the search phase performed by the user are needed, but certainly commercial search engines, that make profit by the number of access to their pages, are not interested in produce them.

The rethinking of search engines involves the emerging of some questions about the method of search through repeated queries and their successive

Manuela A., De Vita E., Cristian L., Ivan M., Gavino P. and Franco T. (2010).
A SEMANTIC SEARCH ENGINE FOR A BUSINESS NETWORK - A Personalized Vision of the Web applied to a Business Network.
In *Proceedings of the 5th International Conference on Software and Data Technologies*, pages 475-480
DOI: 10.5220/0003040304750480

refinement. Someone thinks that search engine should be considered "only a primitive form of decision support" (Spivack, 2010). So, the vision of a Web where search engines are able to provide results without direct questions from users, anticipating their needs, could be now plausible. A Web in service of the user, automatically informed by the system with suggested resources related with his life style and his common behaviour without the need to ask for them.

In this paper we illustrate a section of a project named *A Semantic Search Engine for a Business Network* where some of the ideas previously described are applied. It involves the development of a business network able to create a point of contact between the academic and research world in general and the productive one, with the aim of encourage the cooperation and the sharing of ideas, of different point of views, information material or needs, and in order to support the productive world and decision-making connected with it.

The infrastructure will be designed as a distributed architecture, with regard to information and existing content and by the development of tools thought in order to put users into the center of information, giving them a privileged window of observation on information and knowledge applied to a specific application field.

The remainder of the paper is organized as follows: Section 2 describes the project in a general way, while Section 3 discusses in a detailed way the above questions. Finally, Section 4 draws conclusions.

2 GOALS OF THE PROJECT

In order to focus better on purposes and objectives of the project, some considerations are required.

Web is changing. The way to access the information is not the same of some years ago. Social networks, blogs, RSS and new features in search engines are all news in the ICT context if compared with some years ago. The trend, hopefully, is the definition of new tools developed in order to follow the user in his activity and support him with the automatic generation and delivery of contents without his explicit request and according to his interest. The Web depending on user needs and interests. Not more a single available and shared space of information for everyone, but a specific private space of information available according to the user point of view, his way to organize, classify and manage information, related to his network of contacts in the way each person choose to live the Web, Internet and the knowledge. Currently, the management of information is a key question in the Web.

The automatic categorization of information through a predefined taxonomy, organized in a hierarchical category system, is often a restrictive and forced path. The same resource could be classified in different way from different people and the same user could place the same page under different categories according to the reading context or to the content he is interested in. The classification of a document is, as well, depending by the personal culture, experience and context of life. Moreover, documents are often realized using heterogeneous contents, talk about several topics and are obviously related to several categories.

Otherwise, with the Web 2.0, folksonomy, social tagging and social bookmarking place the user as start point in a categorization work where each user labels resource. This step moves from a hierarchical logic to a more simpler way where all tags are at the same level.

Passing from the user management of information to an automatic one, a classification system should be able to categorize information according to user preferences and to relate his classification to a common set of categories based on a predefined taxonomy.

By means of a such categorization tool, each user manages in a personal way his bookmarks, accedes to a quantity of Web sites, about scientific, news, entertainment or other topics, selecting, choosing and categorizing through the system. The system is able to manage a flow of information coming from a big set of predefined channels and updatable depending on the user preferences. Channels should be social networks, blogs, RSS services, news services, Web sites and search engines too, selected by the user. The system categorizes information from these channels delivering contents that meet user preferences by means of a match algorithm based on user profile and document classification. The user can see categories associated to each resource labelled and ordered according his schema.

The vision of the Web and of search engines, as described below, is applied to a project in starting phase and will converge in a system able to support and follow users in their activities. In particular the idea behind the project is the realization of a business network able to guarantee the match and the cooperation of academic and research world with the productive one in order to sustain related production and decisional processes.

3 THE CLIENT APPLICATION

The semantic search engine is designed as a support tool for the user, an active assistant able to give in "real time" references for the use of the information, reporting as more interesting the information that might match with the personal interest specified in the user profile.

In the project have been identified two types of main users: the company and the generic user, including employees, researchers, professionals, and in general people having specific interests and skills, according to the resources associated with them and emerging by their daily activities, that the system is able to track.

The business network is a point of contact between the academic and research world in general and the productive one. The aim is to encourage the cooperation and the sharing of ideas, of different point of views, information material or needs, and to support the productive world and decision-making connected with it.

The system manages the user profile in order to control how the user preferences evolve during sessions of work. Information is monitored at time interval and new sessions can modify user preferences. The system starts with a predefined user profile and evolves subsequently, using text categorization tools in order to categorize resources that are actually read, saved, commented. Only in these cases the system will modify the user criterion of classification for subsequently analysis.

The system follows step by step the evolution of user interests and suggests him, through the analysis of his profile, topics of interest, documents, contacts, etc, according to his interests. Moreover, the system is able to associate user profiles to companies or project profiles, automatically generating in real-time networks of expertise based on several configurable parameters and requirements.

Figure 1 shows a general description of the client application and the flow of the data coming from several distributed sources, such as social networks, blogs, RSS pages, visited Web pages, etc. The client side of the system is composed by four modules: the User Profiling Module, the Collaborative Filtering and Recommendation System Module, the Classifier and the Matching Module, each responsible of the functionalities described below.
The level of communication between the modules and the distributed information is regulated by a layer that receives the data coming from the sources, and after an analysis and an opportune elaboration, is able to deliver to each module the portion of

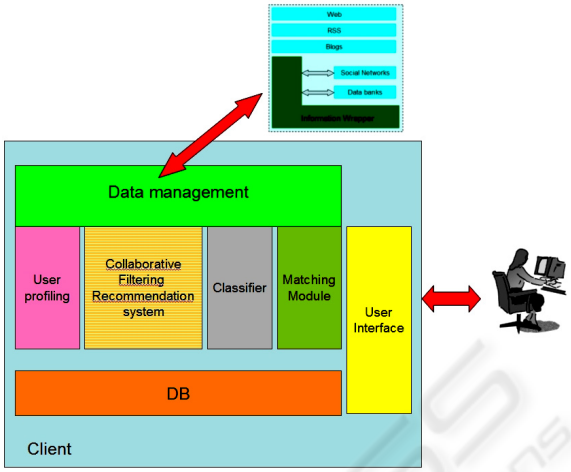

Figure 1: The client application.

information that they are able to manage. Each module performs his activity, sometimes collaborating with other modules, and the result of the process is saved on a database.

The interface allows queries in natural language and presents results according to the user profile and preferences.

More details of the modules involved in the system are described below.

3.1 The Business Network

The business network defines a communication level between users belonging to a community. The business network facilitates the sharing of knowledge, ability, expertise, skills, interests and resources between users belonging to the community that need or are interested in specific topics. In fact, it is not always easy to rise these feature, especially the immaterial expertise. But even publications or ongoing or past projects in which someone is involved, are often dispersed between public databases, or can be found only in the intranet of each company, or sometimes exists only in the head of someone, and it is not easy to explicit them. All the members of the community are linked together by the net of their skills: they are both depository of expertise in the service of users who need it: on the other hand they can need skills (papers, suggestions, projects, contacts) that other members can make available. This can be achieved with the development of an application, running on the computer of the user, that filters his activities and modifies his status, walls and links of the social network that the user subscribed, according to his permissions. Simultaneously records the activities

on the user database.

The application shares this data with the other users that subscribed the community so that each user, according to the settings and the permissions, should know which resources have been visited, from whom and when. The application communicates these information to a plug-in installed on the user's browser that alerts the user and updates the visualization of information according to his preferences.

3.2 Management of the Social Network

As said before, users involved in the project are organized as a community, configured according to their activities, through the management and by reporting organized content, information dynamically updated and personalized according to the specific user profile.

The system will provide access to sources of shared documentation, to monitoring data, to support tools for sharing information between users, to networks of contacts explicitly specified in the community.

The architecture of the general platform is still under discussion.

3.3 Search Engine Module

The search engine module is contained in the Data Management Module, still under definition. The search engine indexes information coming from data sources and manages information related to the users, communities, companies, events, etc.

3.4 User Profiling

User profiling is a crucial process of the system because it has to define the user's interest, allowing the collaborative filtering and the recommendation tools to select and send information useful for the user itself. The module is able to classify and manage user information through the analysis of the resources he visited: the registration to rss resources, blogs, to social networks and the associated map of contacts, the collection of feedback, etc.

A profile for both users, companies and researchers, is defined creating in such a way a history depending on their activities and behaviour. So, the system will be able to identify user requirements and to predict its future behaviour and interests, in order to automatically propose resources useful to its activities without the need to search for them.

Data collected in this way are used by the system to find similarities, complementarities and links between companies and researchers, thus facilitating the match between supply and demand, particularly for intangibles such as interest, expertise, know-how.

The user should be able to access to its profile in order to check the reliability of the image that the system is bringing out, providing a positive or negative feedback to the matching proposed by the system.

3.5 Collaborative Filtering and Recommendation System

During his activities, the user is supported by a module that helps him through two very important features: a collaborative filtering (De Vita et al., 2008) and a recommendation system. This module filters information by means of parameters based on the user preferences and his profile and gives advice to the user for news regarding communities and network activities that should be of interest. Advices are about:

- New activities
- Users having similar interests
- Companies having similar profile
- Researchers having similar profile (based on their *curriculum vitae*)
- Events of the network: workshop, conferences
- Documents, papers, notes, projects, reviews classified that match users interests.
- Announcements of competition, calls, etc

By means of the indications given by the user to the system it is possible to refine the profile.

3.6 Data Categorization

The system, with the user profile module, compares user profiles to company profiles through data categorization. It matches similar profiles, compares curricula of the user with request coming from companies, filters news and contents coming from the search engine working on the semantic of texts.

The classifier is based on a hierarchy of categories proposed by WordNet Domains (Magnini et al., 2002). These categories are the set of starting used by the system for the text categorization of resources. The user has the possibility to confirm the categorization proposed or to redefine it with labels not presents in the original taxonomy or to move the resources to better defined values of the involved categories. The user can also rename categories. The

system needs to keep references with new names and different values given by the user to resources.

In order to modify this kind of information a feedback from the user is necessary. The user instructs the system until the reach of a number of documents big enough to be representative for each category. This allows eventually to pass from a semantic classifier to a faster statistical one.

The classifier performs a semantic disambiguation through the identification of relation between terms in order to identify composed terms, word sense disambiguation, name entities, geographic location.
The main phases are:

- Parsing of the text of resources (Web pages, documents, notes, etc)
- Analysis and syntactic disambiguation (Sleator and Temperley, 1993) (Liu, 2004)
- Semantic disambiguation and identification of real senses of words in sentences by means of a density function (Addis et al., 2009)
- Identification of name entities, geographic locations
- Classification of the textual resource by categories and values (Angioni et al., 2008a)
- Identification of semantic relations between concepts (Angioni et al., 2008b)

3.7 Matching Module

The module is responsible to perform the matching between the information coming from the several data sources and by the users' profile, identifying those of real interest for each user.

It is able to organize data coming from users and companies profile, managing the textual resources, such as notes, papers, comments, profile data, previously analyzed by the classifier and aggregate the information.

Finally it send notifications to users and the information as elaborated by the specific algorithm of matching.

3.8 Data Sources

As we said, the system will be able to retrieve information from several textual and multimedia sources, and from Web services, even if conditionally.

Figure 2 shows in a summary way the data sources and a module named Information Wrapper that uniforms data coming from data banks (DBLP, ACM DL sites or institutional databases) and, under particular conditions, from social networks.

Some sources such as news services, social networks, blogs, RSS feeds, will be selected by the user or they will be automatically proposed by the system, by means of the preferences expressed by default or defined by the user profile and by the interests identified by the viewed pages.

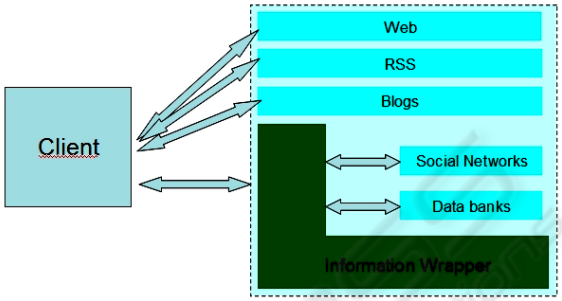

Figure 2: The information wrapper.

Other content will consist of personal and corporate profiles extracted from the HTML home pages, abstracts of scientific publications, bibliographies extracts from data banks.

Initially the contents are classified according to a predefined taxonomy. Then the user requests, aggregates and organizes the information coming from the news services according to its interests.

The system therefore has to be able to manage only the resources having a significant content for the user, eliminating in such a way the redundancy of the received information, the repetitions and the duplications, and avoiding waste of time to read unnecessary and irrelevant contents or to search information among all the resources available on the Web.

Finally, the system has to be able to add "meaning" to the actions performed by the user, creating an area of personalized and organized information, a powerful guide able to predict its tastes and needs.

4 CONCLUSIONS

The introduction of Natural Language Processing in search engines, the user profiling, the automatic classification of information according to the personal schemas of the users are redefining the vision of information on the Web and are delineating new processes of Human-Machine Interaction.

Moreover the deployment of new services and tools of the Web as Social Networks, RSS feed and of new users' supports based on NLP are defining

new evolutionary scenarios and creating new expectations for the Web.

In this paper we illustrated a starting project named *A Semantic Search Engine for a Business Network* that defines a scenario where all above tools converge in a system that, in our intention, will put the user into the center of information giving him a privileged window of observation on information and knowledge.

The proposed approach aims at the development of a business network able to create a bridge between the academic and the research world in general and the productive one, allowing a point of contact between users' needs on one hand and available skills, expertise and ability on the other.

The project aims both at implement the features described and at define and implement the described scenario. A validation to support the value of the expressed ideas will be one of the goal of the above mentioned project, where experimental results will be product.

REFERENCES

Addis, A., Angioni, M., Armano, G., Demontis, R., Tuveri, F., Vargiu, E., 2008. A Novel Semantic Approach to Create Document Collections. *In Antonio Palma dos Reis, editor, Proceedings Of Intelligent Systems And Agents* Pages 53-60, 2008. IADIS Press. Selected for the best paper award.

Angioni, M., Demontis, R., Tuveri, F., 2008a. A Semantic Approach for Resource Cataloguing and Query Resolution. *Communications of SIWN. Special Issue on Distributed Agent-based Retrieval Tools*, 5: 62-66.

Angioni, M., Demontis R., Deriu, M., Tuveri, F., 2008b. SemanticNet: a WordNet-based Tool for the Navigation of Semantic Information. In *A. Tanacs, D.Csendes, V. Vincze, C. Fellbaum, and P. Vossen, editors, Proceedings Of GWC.* University of Szeged.

De Vita, E., Deriu, M., Marcialis, I., Paddeu, G., 2008. Personalization and Collaborative Filtering for Information Retrieval on the Web. *Communications of SIWN.* Special Issue on Distributed Agent-based Retrieval Tools, 5(-): 51-56.

Marcialis, I., De Vita, E., 2008. SEARCHY: An Agent to Personalize Search Results. *A. Mellouk, editor, Third International Conference On Internet And Web Applications And Services.* Volume -. Pages 512-517. IARIA. Institute of Electrical and Electronics Engineers (IEEE). Authorized distributor of all IEEE proceedings.

Horowits, D., Kamvar, S., 2010. The Anatomy of a Large-Scale Social Search Engine. Submitted to WWW2010, Raleigh, NC, USA.

Liu, H., 2004. MontyLingua: An end-to-end natural language processor with common sense, viewed 30 March 2010, <http://web.media.mit.edu/~hugo/montylingua>.

Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., 2002. The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering, special issue on Word Sense Disambiguation,* 8(4), pp. 359-373, Cambridge University Press.

Mislove, A., Gummadi, K., Druschel, P., 2006. Exploiting Social network for Internet Search. In *Proceedings of the 5th Workshop on Hot Topics in Networks,* Irvine, CA.

Sleator, D. D., Temperley, D., 1993. Parsing English with a Link Grammar. in *Third International Workshop on Parsing Technologies*.

Spivack, N., 2010. Eliminating the Need for Search-Help Engines, viewed 30 March 2010, <http://www.novaspivack.com/uncategorized/eliminating-the-need-to-search>

Wissner, J., Spivack, N., 2009. Case Study: Twine. In *W3C, Semantic Web Use Cases and Case Studies*, viewed 30 March 2010, <http://www.w3.org/2001/sw/sweo/public/UseCases/Twine>