

Integrated Agent-Based Approach for Ontology-Driven Web Filtering

David Sánchez, David Isern and Antonio Moreno

Universitat Rovira i Virgili (URV)
Computer Science and Mathematics Department
Artificial Intelligence Research Group, BANZAI
43007 Tarragona, Catalonia (Spain)
{david.sanchez,david.isern,antonio.moreno}@urv.net

Abstract. For knowledge-intensive industries it is of paramount importance to keep an up-to-date knowledge map of their domain in order to take the most appropriate strategic decisions. The Web offers a huge amount of valuable information, but its interaction is very hard and time consuming for humans because it requires to filter, analyse all related web pages and integrate it in a knowledge repository. This paper describes an integrated agent-based ontology-driven approach to retrieve web pages that contain data relevant to each of the main concepts of the domain of interest in a completely automatic, unsupervised and domain independent way.

1 Introduction

The Web offers a huge amount of valuable information, but it is scattered, unstructured and impossible to analyse manually. It is usually searched by means of keyword-based search engines, allowing a user to retrieve information by stating a combination of keywords. The results of this type of search usually suffer from two problems derived from the nature of the query and the lack of structure in the documents: low precision and recall ratios. Furthermore, while search engines provide support for automatic information retrieval, the tasks of extracting relevant data and its further processing remain to be done by the user.

In the last years it has been argued that the performance of a search engine can be improved by using ontologies [1]. Ontologies allow organizing and centralizing knowledge in a formal, machine, and human understandable way, making themselves an essential component to many knowledge-intensive environments like the Semantic Web, knowledge management, and electronic commerce. In consequence, they provide a semantic ground that can help to sort out web pages with relevant information about a concept from web pages that contain data with just syntactic similarities to the concept. However, ontologies are traditionally built entirely by hand and, in consequence, their creation and management requires a significant amount of human effort that can compromise the performance and applicability of knowledge based tools.

To tackle those problems, this paper presents an integrated approach for web information retrieval and filtering, providing a tool for obtaining web pages that contain relevant information about the main concepts of a domain, expressed by an automatically obtained domain ontology. This ontology has a hierarchical tree structure that contains the basic classes (concepts) of the domain and the main characteristics (attributes) of each concept.

The system uses two previously developed tools for knowledge acquisition and information retrieval. Concretely, in [2] an automatic, domain independent, web based ontology learning method is presented. Its results (machine readable ontologies for any domain) can be used as input for the system described in [3], which implements methods and techniques that allow the use of the information contained in the domain ontology in order to move from a purely syntactic keyword-based web search to a semantically grounded search. The final result is a structured representation (in an ontological fashion) of the main concepts for a certain domain, which is used to retrieve, filter and classify the most relevant web resources available in the Web. As the processing required to treat with a huge repository like the Web is a very time consuming task, the full system is presented in a distributed approach. More concretely, in order to provide a scalable solution, the agent paradigm is a promising technology for information retrieval. *Multi-agent* systems provide advantages with respect to traditional systems such as scalability, flexibility and autonomy [4] and they are very suitable for implementing dynamic and distributed systems like the presented.

As a summary, the main features of our contribution are: *i*) unsupervised operation during the analysis, learning process and filtering of Web resources. This is important due to the amount of resources available, avoiding the need of a human expert on the searched domain. *ii*) automatic operation, allowing to perform easily executions at any time in order to maintain the results updated. This characteristic fits very well with the dynamically changing nature of the Web. *iii*) domain independent solution, because no domain related assumptions are formulated and no predefined knowledge is needed. This is interesting when dealing with technological domains where specific concepts may appear.

The rest of the paper is organised as follows. Section 2 introduces the learning methodology used to obtain basic ontologies. Section 3 describes the ontology driven web information retrieval and filtering tool. Section 4 presents the agent-based integration of those two complementary approaches and discusses the evaluation of the results. The last section discusses related approaches and introduces some lines of future research.

2 Ontology Learning

The base for obtaining the basic ontologies for a domain, is the intensive use of a methodology [3] for acquiring knowledge from the Web. The most important characteristic of the method is that the whole process is performed in an automatic, unsupervised and domain independent way, allowing to obtain results without user's intervention.

The algorithm is based on analysing a significant number of web sites in order to find important concepts for a domain by studying the neighbourhood of an initial keyword. Concretely, in the English language, the immediate anterior word for a keyword is frequently classifying it (expressing a semantic specialization of the meaning) [5]. So, the *previous word* for a specific *keyword* is used for obtaining the taxonomical hierarchy of terms (*e.g. breast cancer* will be a subclass of *cancer*). The process is repeated recursively in order to create deeper-level subclasses (*e.g. metastatic breast cancer* will be a subclass of *breast cancer*).

In order to extract and select the most relevant concepts for a domain from the Web, the method relies on a search engine for accessing the available web resources. It constructs dynamically the appropriate queries for the search engine, obtaining the most adequate corpus of web resources at each time. Moreover, the Web search engine is also used to select the most appropriate concepts, checking their relevance for the specific domain (the strength of the taxonomical relationship) through a statistical analysis based on the number of estimated results returned by the search engine. This approach allows obtaining robust statistical measures (as they are based in the whole Web) in a very efficient and scalable way, and has been proved to be an effective strategy for inferring the degree of relatedness between concepts [6].

As an additional step, the taxonomy is filtered in order to detect if some of those selected concepts can be considered as properties or attributes of a specific class. More concretely, we consider that those terms that appear in several branches of the obtained hierarchical structure, are a common property of the immediate superclass. For example, if we find that among the different types of discovered cancers (*e.g. breast cancer, lung cancer, etc.*), many of them can have a common attribute (*e.g. metastatic breast cancer, metastatic lung cancer*), we consider this attribute as a property of the common superclass (*cancer*).

The result of the process is a hierarchical organization of the main concepts available for a domain according to the information contained in the Web, enriched with some automatically discovered attributes (see an example in Fig. 1). This structure is presented in an ontological fashion using a standard machine readable language that allows an easy integration with the rest of the system.

3 Ontology-based Web filtering and ranking

Here, the ontology-based search system that explores the Web to find relevant pages related to the different concepts in an ontology is introduced (see [7, 2]). The retrieved pages are textual instances of the concepts, but conditioned to the meaning of the concept in the whole ontology. It means that the same concept in a different ontology would produce different results because it is in a different context. The content of the pages related to a particular concept are analysed in order to rank them according to a relevance function which takes into account the properties describing the user desired profile of such concept.

The search system wraps traditional search engines by an intelligent system that cuts the domain ontology into pieces (sub-ontologies) according to the

degree of concurrence, and scatters these pieces between the available search processes running in the computers involved in the process. A sub-ontology is formed by concepts (and attributes) from a leaf to the root node. Then, each search process works to obtain as many relevant pages as possible.

As the pages are retrieved, their contents are analysed and the relevance of the page is calculated in terms of the sub-ontology concepts and properties appearing in the documents. The relevance value is used to rank the pages and also to discard those pages which are not of the expected quality. The most relevant ones will populate the domain ontology, and will be joined asynchronously later in order to be sorted and filtered appropriately (see an example in Fig. 1).

In some situations, the amount of ontological information available could be too restrictive for the keyword based search engine (*e.g. microinvasive endobronchial non-small cell lung cancer*). In these cases, a complementary component was designed in order to modify these problematic sub-ontologies by removing the least representative concepts and give alternative queries less constrained.

4 Agent-based ontology-driven web filtering

As has been introduced, due to the computational cost of the described Web based methodologies, we have opted for modelling them into a distributed agent-based platform that can be executed on the server side to which the user can access via a web interface. The full process has been divided into several tasks that are mapped into different types of agents. In this manner, those agents can be executed in parallel among nodes of a network, cooperating to achieve the common goal. Moreover, as shown in [4], agents provide the high degree of flexibility required for the dynamic management of the platform, and implement complex communication skills necessary for coordination.

As shown in Fig. 1, the multi-agent system is divided in two parts that correspond to each described methodology: the output of the ontology learning methodology is used as input of the web filtering process. As a consequence, different types of agents are created and managed dynamically (created, configured and finalized) according to the execution requirements at each moment, ensuring that the available computational resources (nodes of the computer network) are always efficiently maximizing the throughput of the system.

4.1 Ontology learning

The *Ontology Builder* module is composed by two types of agents: an Ontology Builder Agent (OBA) and several Taxonomy Builder Agents (TBA).

The process starts when the OBA receives from the user the concept (*e.g cancer*) that represents the domain to explore (step 1). Optionally, the user can specify some parameters to constrain the search process according to his desires as described in [3]. As OBA's goal is to construct a basic ontology that represents the available knowledge for this domain, it creates a first TBA that starts building a one level taxonomy using the methodology presented in §2

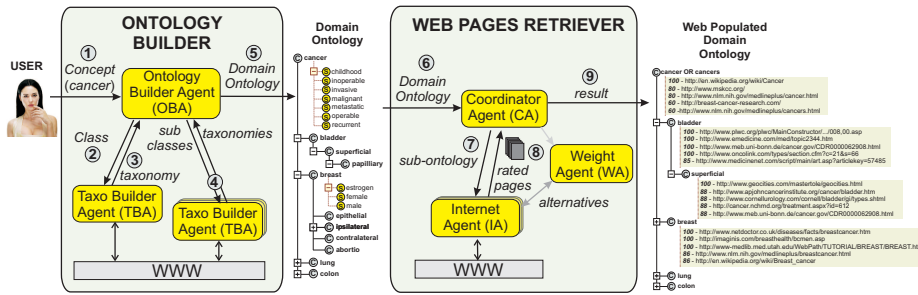


Fig. 1. Integrated agent based ontology driven web recommendation platform. Example results presented correspond to the *Cancer* domain.

(step 2). As a result, a set of immediate taxonomically related subclasses (e.g. *breast*, *lung*, *colon*) is returned to the OBA (step 3). The OBA incorporates this knowledge into the domain ontology and, for each new subclass, it creates a new TBA for exploring it.

Again, several sets of taxonomically related subclasses are returned to the OBA that incorporates them into the ontology (step 4). Repeating this process until no more subclasses are found, the OBA is able to compose recursively an ontology that taxonomically represents the available knowledge in the Web for the domain. As a final step, the OBA refines the ontology in order to detect attributes for each class (e.g. *metastatic cancer*) as described in §2 and outputs the result in a machine readable ontology representation language (step 5).

4.2 Web retrieving and filtering

The *Web Pages Retriever* module is composed by three types of agents: a *Coordinator Agent (CA)*, a *Weight Agent (WA)* and some *Internet Agents (IA)*.

As a result of the execution of the *Ontology Builder* module, the CA receives the automatically acquired *domain ontology* (step 6). Then, the CA divides the domain ontology and distributes the search work among the available IAs (step 7). It uses for this purpose a split operator that creates one ontology per class, keeping the whole path from the root node of the ontology and all the properties inherited from it [2] (*i.e.* all the superclasses of the concept, with all their properties). Each IA uses a standard keyword-based search engine to retrieve a set of web pages that are related to a concept of the domain. The agent uses the semantic information of the its subontology to filter and rank these pages, and sends them to the CA (step 8).

IAs may have problems for retrieving results if its subontology is excessively restrictive, as mentioned in §3. In this case, he can request the help of the WA. This agent is able to find less constrained sets of keywords that can be used by IAs to find more pages. This agent implements a Best First Search with all possible sequences of keywords to be considered by an IA. The agent maintains a sorted queue per IA with the alternatives to be sent when it is needed.

Parallely, the CA waits for all the IAs to supply the results. When those are provided, it incorporates the returned lists of web resources into the domain ontology that is presented to the user as the final result of her request (step 9).

At the end, for each automatically acquired concept, a set of 2-tuple formed by an URL and a rate is presented. This last value indicates the degree of relevance of the particular URL and its associated concept according to the ranking measure employed during the retrieval and filtering process. Note that due to a specificity policy implemented, no redundant results between classes and subclasses are presented.

It is important to note that the full system is not intended to provide an immediate response in a first moment as, depending on the queried domain, the computational resources available on the server side and the search engine response times, the full process can go from minutes to several hours. However, once a domain is explored, results can be consulted immediately and even updated in a very fast way (as previous results as the domain ontology can be reused).

4.3 Evaluation

As the present proposal is an integration of two previously developed tools, the quality of the final results depends on the performance of each methodology. Regarding to the evaluation of the taxonomies obtained by the first module, a discussion is offered in [3], offering a comparison against a gold standard and several taxonomical web search engines. With respect to the second module, in [2] are presented several evaluations against different technological domains starting from ontologies composed manually by experts.

The full platform has been tested in several domains as medicine, biotechnology and computer science. The evaluation has been performed by comparing the results against the web search engine used during the analysis (Google). More concretely, for the list of URLs retrieved for each automatically discovered concept, two users are requested to rate each web site according to their degree of interest for the particular domain with a value between 0 and 100. The same process is repeated for the first web sites returned Google when manually querying the same acquired concept. These ratings will indicate which approach returns, in average, the most interesting set of web resources for the particular domain.

As an example, in Fig. 2, expert's rating for our results against Google for a pair of concepts of the *cancer* domain are presented. One can see that, for the most general concept (*cancer*), the quality of our results overpass significantly, in average, the ones presented by Google. This behaviour have been observed for several tested domains and it is caused by the higher contextualization that the presented approach can apply to the web sites analysis thanks to the automatically acquired knowledge for the domain. Observing the average rating for a more concrete concept (*breast cancer*), we can see that the quality of the returned web sites by each system is very similar. In this case, the search is, in both cases, contextualized enough to retrieve high quality resources.

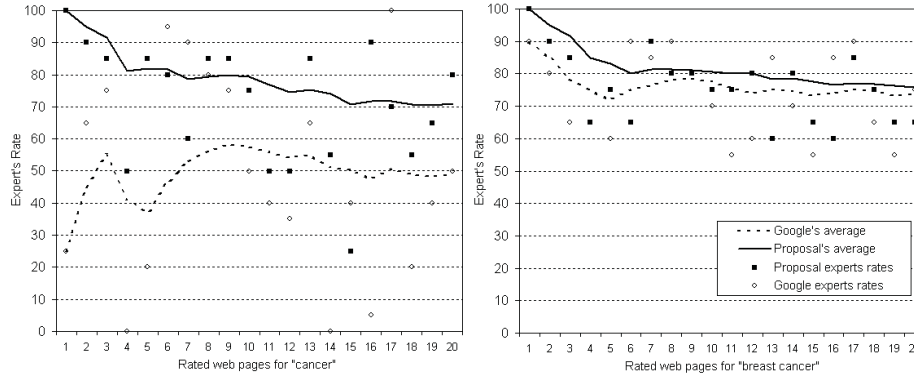


Fig. 2. Expert’s rating for the first 20 web pages returned by our approach against the ones retrieved by Google for the *Cancer* and *Breast Cancer* concepts.

5 Discussion and future work

On the one hand, some authors have been using the Web as a learning corpus for developing [8] or enriching knowledge structures [9], presenting techniques adapted to this environment. On the other hand, the use of knowledge structures (e.g. thesaurus like WordNet) is a common approach for improving the performance of Web information retrieval [10]. However, results for those approaches depend on the domain coverage of the knowledge base used.

On the contrary, the presented proposal does not start from any predefined knowledge and, in consequence, it can be applied over domains that are not typically considered in semantic repositories, but maintaining the semantic context provided by the automatically and unsupervisedly obtained domain ontology. At the end, we can bring the benefits of unsupervised and domain independent and, at the same time, semantically grounded Web information retrieval together into an integrated agent-based approach.

Moreover, the distributed and coordinated agent-based execution, improves the scalability and the throughput of the system, by taking profit of a parallel execution through several nodes of a computer network. In this sense, agent’s flexibility (such as dynamic management and adaptation to the execution requirements of each moment) and communicative skills [4] have been crucial points modelling the presented platform. These facts result in a scalable and suitable method for acquiring knowledge and retrieving relevant web resources from a huge and dynamic repository as the Web.

Applications of the results obtained can be: *a)* the domain ontology builder can be a great tool for structuring automatically the Web’s knowledge. The acquired ontologies are crucial in many knowledge intensive tasks such as electronic commerce, knowledge management or the Semantic Web; *b)* the automatic filtering, ranking and structuring of web resources can be considered as an improvement over the classical way of accessing web sites.

As future lines of research, some topics can be proposed:

- When dealing with natural language resources like web pages, problems about semantic ambiguity may arise. For that reason, we have developed complementary methods for dealing with polysemy and synonymy [11] specially adapted to our working environment. We plan to integrate those techniques into the learning methodology in order to improve the results.
- The multi-agent system versatility can be improved by incorporating communication and negotiation capabilities between agents that can allow them to share intermediate results in order to avoid redundant searches improving the learning performance.

Acknowledgements

The work has been supported by *Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya i del Fons Social Europeu* of Catalonia.

References

- [1] Fensel, D.: *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer Verlag (2001)
- [2] Moreno, A., Riaño, D., Isern, D., Bocio, J., Sánchez, D., Jiménez, L.: Knowledge Exploitation from the Web. In: 5th International Conference on Practical Aspects of Knowledge Management (PAKM'04). Volume 3336 of LNAI., Springer Verlag (2004) 175–185
- [3] Sánchez, D., Moreno, A.: Agent-Based Knowledge Acquisition Platform. In: 9th International Workshop on Cooperative Information Agents (MATES/CIA 2005). Volume 3550 of LNAI., Springer Verlag (2005) 118–129
- [4] Wooldridge, M.: *An Introduction to Multiagent Systems*. John Wiley and Sons, Ltd., West Sussex, England (2002)
- [5] Grefenstette, G.: SQLET: Short Query Linguistic Expansion Techniques: Palliating One-Word Queries by Providing Intermediate Structure to Text. In: Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, RIAO'97. Volume 1299 of LNAI., Springer Verlag (1997) 97–114
- [6] Turney, P.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Twelfth European Conference on Machine Learning. (2001)
- [7] Bocio, J., Isern, D., Moreno, A., Riaño, D.: Semantically Grounded Information Search on the WWW. In: Artificial Intelligence Research and Development. Volume 100., IOS Press (2005) 349–356
- [8] Navigli, R., Velardi, P.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics* **30** (2004) 151–179
- [9] Agirre, E., Ansa, O., Hovy, E., Martínez, D.: Enriching very large ontologies using the WWW. In: Workshop on Ontology Construction of the European Conference of AI (ECAI'00). (2000)
- [10] Abramowicz, W.: *Knowledge-Based Information Retrieval and Filtering from the Web*. Springer Verlag (2003)
- [11] Sánchez, D., Moreno, A.: Development of new techniques to improve web search. In: 9th International Joint Conference on Artificial Intelligence. (2005) 1632–1633