Searching over Public Administration Legal Documents Using Ontologies

Diego Berrueta[a], Jose Emilio Labra[b] and Luis Polo[a]
[a] *CTIC Foundation, Gijón, Spain*[1]
[b] *Department of Computer Science, University of Oviedo*

Abstract. In this paper, we apply Semantic Web technologies to the creation of an improved search engine over legal and public administration documents. Conventional search strategies based on syntactic matching of tokens offer little help when the users' vocabulary and the documents' vocabulary differ. This is often the case in public administration documents. We present a semantic search tool that fills this gap using Semantic Web technologies, in particular, ontologies and controlled vocabularies, and a hybrid search approach, avoiding the expensive tagging of documents.

Keywords. Semantic Web, Spread Activation Algorithms, Synsets, Concepts, Syntactic Search, Semantic Search

1. Introduction

The vast quantity of information available in the world wide web made indispensable the development of search engines and information localization systems. Some of these systems are daily used tools of the majority of people and have become one of the most profitable companies in the Internet sector. These systems have a general domain and try to offer good results for general searches. They have to tackle with documents written in different languages, about quite different subjects and published by very different people and organizations.

In the case of a Public Administration, the type of published documents is more uniform, with one or a few official languages, about some given subjects and with a more controlled generation process. Nevertheless, Public Administrations publish everyday lots of documents that citizens are supposed to read, such as new laws, announcements, notifications or subventions. This information was usually published in printed bulletins, but is increasingly being published on the web. In such a controlled environment, it could be possible to apply specialised approaches which could improve the search results and user satisfaction.

Most of the Public Administration documents are written in legal and administrative jargon, far from ordinary language, and this represents a hindrance for the communication between citizens and Public Administration. This is a

[1] This work is partially funded by the CTIC Foundation through a contract with the Principality of Asturias and by the CTIC project TIN2004-03453

barrier that renders the traditional syntactic search almost useless, as terms that are considered synonyms by citizens have a clearly different meaning for the expert lawmaker.

The arrival of the Information Society has eased the citizen's access to Administration informative publications. But the published information localization rests chiefly upon identification of certain keywords introduced by the user. This way of accessing documents can become profitless, hence the dearth of citizen's habit regarding legal and administrative terminology.

The Semantic Web initiative powered by the W3C tries to provide an enlargement of the present-day Web. This technology allows the users to access the information easily, efficiently and quickly by means of suitable software services.

Among these technologies, we have decided to use an ontology-based model to ease the access to Public Administration documents. In this way, it is possible to develop new services that remove the language handicap, improving the chances of retrieving the desired information.

The paper is organised as follows. In Section 2 we present some motivation. Section 3 describes our ontology based search approach. In section 4, we describe the architecture and the development process that we have followed. Section 5 describes some related work and finally, Section 6 presents the conclusions and future work.

2. Motivation

CTIC Foundation (Centre for the Development of Information and Communication Technologies in Asturias) is a private non-profit institution founded by the Regional Government of the Principality of Asturias and a Consortium of regional private companies in the field of Information and Communication Technologies. The CTIC Research Department, in collaboration with University of Oviedo, has focused on one of the unidirectional communication channels between Public Administration and citizens, namely the *Boletín Oficial* [1] of the Principality of Asturias (BOPA).

This channel was seen as great opportunity to apply semantics, as it combines a number of technical challenges and a linguistic challenge. Anyone is potentially interested in the information pieces in the BOPA, either driven by personal or professional interests, but only a few know the difference between an order, a law and a decree, and these are only some examples. Additionally, a deep knowledge of Public Administration internal structure is often required to determine which department is responsible for the particular area of interest.

Our primary goal is to provide a practical search tool that enables the user to find the desired information, even if he does not have the knowledge of the administrative domain and vocabulary.

[1] The *Boletín Oficial*, or Official Gazette, is a daily newspaper published by Spanish national and regional Governments. It contains the verbatim listings of new legal texts and compilations of all the Administration announcements.

3. Ontology-based search approach

The standard initiative in semantic web rests upon the use of metadata to describe any resource, in this case, legal documents. However, it is too expensive to tag a vast amount of documents (increasing daily), and it is also difficult to maintain the metadata over changes or additions in the domain. From the very start of the project, we gave up this approach, and we decided to introduce semantics in a different way.

3.1. Ontologies

We have built two kinds of OWL-DL[1,2] ontologies with different purposes:

1. A legal and administrative ontology. This ontology formalises the basic structure of BOPA and Regional Public Administration. It captures the relationships between the different departments of the administration and the type of legal texts they can issue.
2. A set of particular domain ontologies. Each one captures a small and well defined area of general knowledge, actually bringing the human expert knowledge to the system.

Every concept has associated a synonym set, similar to those presented in the WordNet architecture [3] but without such complexity. These synsets are the bridge between the ontology and users on one hand, and the ontology and the document base on the other.

The semantic structure was chosen following the psychological bases of Communication Theory, specially the definition of “context” in Relevance Theory [4], treating each subset of the domain ontologies as a common context to both citizen and search engine. That is to say, the concept set constituting the ontology is relevant both for the user search intentions and for the query process.

3.2. Search Process

The search process uses a hybrid syntactic and semantic search [5]. More precisely, it automatically transforms a semantic query into an equivalent syntactic query, exploiting the relationships in the ontologies and some linguistic knowledge. A user query is composed of a set of concepts from a unique context (see step 1 in Figure 1). Both concepts and context are selected by the user through an interface that hides the complexity of the underlying ontologies. No special abilities are required, since the user interface is not harder to use than conventional search engine.

These user-selected concepts are the input of an spreading activation algorithm [6] applied to the concepts in the ontologies (step 2). This algorithm, developed in the Artificial Intelligence area, works essentially as a graph explorer. Given an initial set of nodes, the algorithm traverses the arcs, activating nodes which are closely related.

In our particular case, the concepts are activated in the ontologies by the query forming the initial set. Each of these concepts receives a score, the top score. Hereafter, the spread activation algorithm scores only the concepts closely

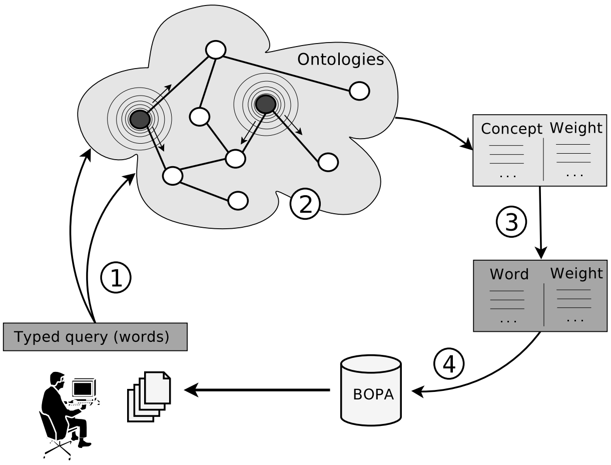

Figure 1. Representation of the semantic search process.

related by semantic links. The more we move away from the initial concepts the more this scoring decreases. The algorithm ends when the concept score is too low or when there are no more concepts to explore.

The output is a larger concept list, sorted by relevance. Some complex techniques are used in order to improve the score of some concepts trying to better reflect the original user search intentions. In this sense, the chain of concepts that forms the most relevant path between user-activated concepts receives an enhanced score.

In the third step, the concept list is then transformed to a word list. Each concept has an associated synset (see Section 3.1.), specially designed to match the legal and administrative vocabulary used in BOPA.

As the fourth and final step, a syntactic search query is built and executed over the XML documents with a conventional search engine. Hopefully, it should bring the desired results.

In spite of using a traditional syntactic search engine, the semantic value of our approach lies in the enrichment of the query by means of the relations between the concepts in our knowledge base. This semantics increments the number of returned documents because of the synsets, merging the results of several queries that users should perform one by one in traditional search engines. For instance, searching for “accessibility” also brings results for “disabilities”, “blindness”, “handicapped” and announces of subventions for the adaptation of public vehicles and buildings, even if there is no syntactical match with the original search term typed by the user.

More interesting results are obtained from multi-concept queries. For instance, combining “chlorine” and “health” returns as result the swimming-pool regulations.

We have implemented this algorithm in Java using OWL-API [7] and Lucene [8].

3.3. Services Provided to Citizens

On the top of these ontologies and this algorithm, we have built three services: semantic search interface, alert subscription and a concept browser.

1. The semantic search service is the bridge service filling the linguistic gap. It provides a simple user interface to perform semantic queries using the algorithm described above.
2. The alert subscription service allows the user to subscribe to any arbitrary search results, and to be notified as new similar information becomes available.
3. Ontologies can be exploited in more ways than just searching pieces of information in the bulletin. As they reflect a model of Public Administration operation, citizens can gain more knowledge about the Administration browsing the contents of the ontologies. We have built two concept browsers, using SVG and plain HTML. Exploring the ontology relations also helps to explain the semantic search results.

4. Architecture and Development

In addition to the research topics described in the previous section, we provide some engineering notes about the application itself and its development process.

4.1. Application Architecture

Our application is a multi-tier application built with J2EE technology, making heavy usage of the industry patterns [9].

The data access tier interacts with multiple data sources, such as a relational database, a XML-native database, some syntactical indexes managed by Lucene and the original web server where the BOPA is still published.

The business tier integrates several vertical subsystems, with low coupling, such as syntactical analysis and search engine, semantical search engine, ontology processing and data fetching.

The main user interface is built using the Struts framework, and communicates with the business tier through Enterprise JavaBeans and SOAP web services.

4.2. Development Process

A team of 5-6 people has been working in this project for nine months. This is a multidisciplinary team of Computer Engineers and Linguistic experts. The development process has been driven by an agile methodology (Extreme Programming [10]), which fits particularly well into projects with a high innovation component. In this way, a working product was already available as soon as three months after the beginning of the project. Since that initial prototype, the team has focused on adding more features, or user stories, in several short (two weeks) iterations.

By following an agile approach, the team has been able to react to frequent changes in the scope of the project (which should be no surprise in a research project) and even changes to the team size.

4.3. Recovering Information from HTML

In the semantic web vision, documents are published in the web with their metadata expressed in an appropriate language, such as RDF. This is not the case nowadays, as a vast amount of the contents of the web are expressed in HTML, focusing in presentation issues and taking small, if any, care of preserving semantics.

The web version of the BOPA is a clear example of the previous problem. By historical reasons, the main interest of its publisher is the page composition of the printed version of the bulletin. The HTML version is generated as a subproduct of the traditional bulletin publishing chain, using the “export to HTML” feature of the word processing software.

As a result, the published HTML has poor quality. At the beginning of our work, some of the main problems we found were non-validating HTML, mixed presentation and data and inconsistent markup of titles and article separators.

We set up a fetching and cleaning process as follows:

1. We fetch the articles from the existing web servers. The file-article correspondence is not one to one, so our robot follows the links and detects the article separators. We gather pages from two different web servers because they contain complementary data. Unfortunately, matching the data from each web server is not straightforward, so we apply some heuristic techniques to determine the mappings.
2. Using the JTidy library (a Java port of the W3C HTML Tidy), we transform the HTML into valid XHTML documents.
3. By applying XSL stylesheets, we extract the data from the XHTML. The data is consolidated in XML documents, and inserted into a native XML database.

4.4. Application Integration

Our application is designed with interoperability in mind. We provide several facilities allowing the integration of new applications:

- The most important features of the application (in particular, the search interface and data access) are exported as SOAP web services. At the moment of this writing, there are two applications consuming these web services: an alternative user interface for blind people and a search interface for the desktop.
- The RSS format is a popular mechanism for syndication. We export a RSS channel which is updated daily with the contents of the new bulletins. We are also considering the generation of RSS channels for the results of the most frequent queries.
- In our system, the user can access the bulletin and article data in HTML and PDF, but also in RDF format. In this way, we feed the semantic web.

4.5. Evaluation of Search Results

It is obvious that the only valid source of feedback about the quality of the search results is the citizen (the end user). Unfortunately, explicitly asking the user for feedback provides only a minimum amount of responses. On the other hand, a lot of indirect feedback can be mined from the server logs and the traces of user activity. Our system implements this in order to evaluate the quality of the search results. We expect to obtain lots of valuable indirect feedback in the next months, as the application is progressively deployed into the production environment.

We also have some training sets, containing the expected results of some popular queries, selected and ranked by human experts. We wrote a simple tool to compare the search results provided by our application and the training sets. This tool provides metrics about the quality of the search results.

5. Related Work

The development of information retrieval methods for web documents has been a topic of research from the beginning of the Web [11,12]. The application of semantic web technologies to improve search engines and knowledge management systems has also been a subject of great interest in the last years [13,14,15,16,17, 18].

In the legal domain there is a lot of ongoing work [19] and specially some approaches that apply semantic web technologies. Gilardoni et al[20] have developed a system which adds a semantic layer to a knowledge management system to assist lawyers in their work. That approach is complementary to our approach in the sense that our focus is to help citizens in general to find legal documents.

Breuker et al[21] proposed a set of ontologies in the e-COURT project. They develop an initial core ontology and several specialised legal ontologies for different local courts. They provide a simple mechanism for query expansion with word sense disambiguation [22]. In their case, the goal of the project is to help in the transcription of hearings of criminal trials. It may be interesting to link the legal concepts in their ontologies with some of the legal concepts in the Public Administration domain or in the European Community Legislative texts[23].

Saias and Quaresma [24] propose a methodology to transform a traditional information system to a semantic aware one and they apply it to legal documents. Although the methodology is similar to our approach, they use a number of different techniques to enrich the queries, to define the legal ontologies and to translate those ontologies to Prolog for inference.

The hybrid syntactic and semantic search approach has been proposed in [5] where they also use the spreading activation algorithm. However, we do not do spreading activation between documents because in our case they are not linked. Our spreading is restricted to concepts in the ontology, knowing nothing about the data in the documents.

In [25], the authors propose another hybrid approach which combines syntactic search using Lucene with some semantic matching using WordNet.

In our approach, we extend the matching to ontology concepts providing more semantic capabilities.

6. Conclusions and Future Work

Combining various techniques (syntactic search engine, spread activation, ontologies and controlled vocabulary thesauri), we have built a semantic search tool over a big document base (over 30,000 documents and almost 150,000 different terms).

We believe that our approach from an ontology based model is fundamentally more efficient than a traditional syntactic search engine. Using the ontology conceptual network, our prototype has the necessary knowledge to find documents which are semantically related to user's intentions. The results we have obtained so far confirm this fact.

This search improving rests upon two factors. On one hand, the combination of semantic exploring techniques (Spread Activation Algorithm) and traditional syntactic processes. On the other hand, the connection between the citizens vocabulary and administrative and legal vocabulary through the ontological bridge.

At the moment of this writing, the application is working with more than 500 different concepts and over 2,000 terms in the controlled vocabulary. We plan to improve our work in several directions:

- Using the tools for search results quality evaluation, we aim to fine-tuning the spreading activation algorithm and the other components of the search process.
- We are also widening the document base to cover generic public administration web pages, and building new ontologies and vocabularies for these new domains.
- Another aim is to improve the user experience bringing the interaction closer to natural language. In order to achieve this goal, WordNet seems to be the most powerful tool [3,26], so we are researching how to link our concepts with WordNet senses.
- There are some query expansion approaches which take into account the system and user history [27,28]. In our case, we are planning to capture a lot of information from the user interaction given that the application domain is localised and that there are a lot of subscribed users.
- Ontology interoperability is one of the keys of the semantic web vision. The DOLCE semantics [29,30] provides a framework to integrate knowledge of different kinds of ontologies. We are planning to align our ontologies under this framework. Another topic for future research is the connection between ontologies [31]

7. Acknowledgements

We want to thank the members of the development: Iván Frade, Jose María Álvarez, Emilio Rubiera and Manuel Cañón and the other researchers from the

University of Oviedo: Agustín Cernuda del Río, Enrique del Teso, Guillermo Lorenzo and Roger Bosch.

References

[1] Deborah L. McGuinness and Frank van Harmelen. Owl web ontology language overview. Technical report, W3C, 2004.

[2] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[3] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*. Harvard University Press and Oxford: Blackwell, 1986.

[5] Cristiano Rocha, Daniel Schwabe, and Marcus Poggi de Aragão. A hybrid approach for searching in the semantic web. In *WWW*, pages 374–383, 2004.

[6] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, (11):453–482, 1997.

[7] Sean Bechhofer, Raphael Volz, and Phillip W. Lord. Cooking the semantic web with the owl api. In *International Semantic Web Conference*, pages 659–675, 2003.

[8] Otis Gospodnetić and Erik Hatcher. *Lucene in Action*. Manning, 2005.

[9] Deepak Alur, John Crupi, and Dan Malks. *Core J2EE Patterns: Best Practices and Design Strategies*. Sun Microsystems, 2003.

[10] Kent Beck. *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional, 1999.

[11] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[12] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. *ACM Computing Surveys*, 32(2):144–173, 2000.

[13] Sara Comai, Ernesto Damiani, and Letizia Tanca. Semantics-aware querying in the www: The wg-log web query system. In *ICMCS, Vol. 2*, pages 317–322, 1999.

[14] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In I. Horrocks and J. Hendler, editors, *International Semantic Web Conference (ISWC 2002)*, 2002.

[15] Sean Bechhofer, Les Carr, Carole A. Goble, Simon Kampa, and Timothy Miles-Board. The semantics of semantic annotation. In *On the Move to Meaningful Internet Systems*, volume 2519 of *Lecture Notes in Computer Science*, pages 1152–1167. Springer, 2002.

[16] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal C Doshi, and Joel Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*. ACM Press, 2004.

[17] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. KIM: a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4):375–392, 2004.

[18] David Huynh, Stefano Mazzocchi, and David Karger. Piggy bank: Experience the semantic web inside your web browser. In *International Semantic Web Conference (ISWC)*, 2005.

[19] R. Benjamins, P. Casanovas, A. Gangemi, and B. Selic, editors. *Law and the Semantic Web*, volume 3369. Lecture Notes in Artificial Intelligence, 2005.

[20] Luca Gilardoni, Chistian Biasuzzi, Massimo Ferraro, Roberto Fonti, and Piercarlo Slavazza. Lkms - a legal knowledge management system exploiting semantic web technologies. In *International Semantic Web Conference*, pages 872–886, 2005.

[21] Joost Breuker, Abdullatif Elhag, Emil Petkov, and Radboud Winkels. Ontologies for legal information serving and knowledge management. In Trevor Bench-Capon, Aspassia Daskalopulu, and Radboud Winkels, editors, *Legal Knowledge and Information Systems*. IOS Press, 2002.

[22] N. Ide and J. Veronis. Word sense disambiguation. *Special Issue on Computational Linguistics*, 24(1), 1998.

[23] Sylvie Despres and Sylvie Szulman. Construction of a legal ontology from a european community legislative text. In T. Gordon, editor, *Legal Knowledge and Information Systems*. IOS Press, 2004.

[24] José Saias and Paulo Quaresma. Semantic enrichment of a web legal information retrieval system. In Trevor Bench-Capon, Aspassia Daskalopulu, and Radboud Winkels, editors, *Legal Knowledge and Information Systems*. IOS Press, 2002.

[25] D. Ravishankar, K. Thirunarayan, and T. Immaneni. A modular approach to document indexing and semantic search. In ACTA Press, editor, *Web Technologies, Applications, and Services*, pages 494–500, 2005.

[26] Piek Vossen. Eurowordnet, general document. Technical report, University of Amsterdam, 1999.

[27] H. Cui, J. Wen, J. Nie, and W. Ma. Query expansion by mining user logs. *IEEE Transaction on Knowledge and Data Engineering*, 15(4):829–839, July 2003.

[28] Erika F. Lima and Jan O. Pedersen. Phrase recognition and expansion for short, precision-biased queries based on a query log. In *ACM SIG Information Retrieval*, pages 145–152, 1999.

[29] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. *The WonderWeb Library of Foundational Ontologies (D18)*. Laboratory for Applied Ontology - ISTC-CNR, 2003.

[30] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweeting ontologies with dolce. In *EKAW*, 2002.

[31] Bernardo Cuenca Grau, Bijan Parsia, and Evren Sirin. Working with multiple ontologies on the semantic web. In *Proceedings of the 3thd International Semantic Web Conference (ISWC)*, volume 3298 of *Lecture Notes in Computer Science*. Springer, 2004.