MULTI-DOMAIN RETRIEVAL OF GEOSPATIAL DATA SOURCES IMPLEMENTING A SEMANTIC CATALOGUE

Julio Romeo VIZCARRA
rvizcarrab08@cic.ipn.mx
Instituto Politécnico Nacional, CIC, Mexico

Christophe CRUZ
Christophe.cruz@u-bourgogne.fr@u-bourgogne.fr
Laboratoire Le2i, UMR CNRS 6306, Dijon, France

Abstract. *Nowadays, the expertise of a user plays an important role in the search and retrieval in the information systems that usually combines general and specialized knowledge in the construction of queries. In addition, most of the queries systems are currently restricted on specific domains. Tackling these issues, we propose a methodology that implements a semantic catalogue in order to provide a smart queries system for retrieving data sources on the web by means of the extension of the user expertise. We propose the combination of a query expansion method, and the use of similarity measures and controlled vocabularies. Thus, it allows the system to recommend data sources that are able to fit the need of a user in terms of information. To reach this goal, we exploit standard such as OWL from the W3C and the CSW GeoCatalogue from the OGC.*

Keywords: Semantic catalogue, smart queries, knowledge engineering, multi-domain retrieval, similarity across ontologies.
JEL classification: L86 Information and Internet Services

1. Introduction

Nowadays, the modern society is in a general crisis of knowledge. This term was introduced by Gross [1], it refers to the necessity of understanding an increasing number of concepts produced for the science and technological applications. On this way, the science and the scientific vocabulary have increasingly merged with wider society through applied science that involves the daily life. As a consequence, the borders between the scientific knowledge (specialized), and the knowledge in the real world (general) outside science have become blurred [2][3]. Both kinds of knowledge can be frequently used to refer common objects or situations. On the other hand, some knowledge is produced within a certain domain, but it can be consumed for others, commonly this shared knowledge cannot be easily accessed and known [4]. In this context, the users can face a lack of background, expertise or non-knowledge (opposite of knowledge) within specific fields.

In order to get closer to a solution of the knowledge issues previously described, users of information systems need central tools able to handle general and specialized knowledge, non-knowledge and expertise about different domains. Moreover, another issue has to be taken into account. The information heterogeneities are thematic, semantic, spatial, temporal, etc. As a consequence, the conceptualization of a domain can widely differ from another domain by defining distinct concept, objects, places or circumstances with the same vocabulary or defining the identical concept, objects, places or events with the identical vocabulary. Thus, these heterogeneities are critical factors in the information integration and retrieval [5].

Currently, there is a vast amount of spatial information available on the web though services. This information allows scientists to perform complex analysis. Goodwin [6] used the term smart queries to describe analyses that combine heterogeneous data sources in order to solve complex problems [7][8]. Our field of interest is the use of heterogeneous data sources to perform spatio-temporal smart queries using Semantic Web tools. In previous work [9] we presented our research on spatio-temporal operators, using local data repositories. The next logical step in the evolution of our work is to integrate it to the SDI (Spatial Data Infrastructure). The term SDI was first introduced by the U.S. National Research Council in 1993. It refers to a set of technologies, policies and agreements designed to allow the sharing of spatial information and resources between institutions [10]. The Spatial Data Infrastructure has a service-oriented architecture. In such infrastructure, functionalities such as storage and data search are carried out through Web services. The typical workflow involves: 1) The discovery of a data source, 2) The download of relevant geospatial data, 3) The use of appropriate analytical methods and 4) The visualization of the results on a suitable map.

Today, the OGC services can be storage in a catalogue and include metadata information, which is described in different ways. Those descriptions include problems of heterogeneity which in the process of integration or retrieval becomes complex, time consumption process, ambiguous, etc. It is relevant to get the right meaning of the concepts in such descriptions; on the other hand, the traditional queries have the same problems with their concepts.

As an example, we present two smart queries given to a user who can involve general and specialized knowledge and non-knowledge:

Query 1: What is the population of crows in southwest of France? In this query, the concept crow can be described in two ways:

- Crow described on general knowledge [11, 12] may be a raven, black bird, superstition bird, a butterfly called "common raven”, etc.
- Crow described in specialized knowledge [13], the crow (Corvus corax) may be related semantically with “Birds robin to mallard size”, “Birds medium size”, “Other similar birds in the same category: small corvid , Corvus frugilegus corone (Rook Carrion crow), Pica pica (Magpie), Garrulus glandarius (Jay), Corvus monedula (Jackdaw), large corvid, etc.”

Query 2: Now considering a query requested for a specialist in geology. What are the locations with colluvium in USA during the past 20 years?

- Using the general knowledge the concept Colluvium is unknown for most of the people [14].
- Then it is necessary to describe Colluvium on the specialized knowledge [15], Colluvium is sediment that has moved downhill to the bottom of the slope without the help of running water in streams, gravity in the form of soil creep, and downhill.

After consulting the concept Colluvium on the specialized domain, we are able to understand and infer the meaning on the general knowledge linking this concept with others semantically related on the general domain such as sediment, deposit, alluvial sedimentation, sedimentary clay, etc.

Our work aims to tackle the issues previously described. We provide to the user the capability of navigating through large amounts of information with an expert approach. This is obtained with the inclusion of specialized knowledge in other fields in the search which the user might not have the best expertise. Moreover, the methodology computes semantically user’s queries and returns similar results in an ordered relevance list. The main domains considered in the retrieval of data sources are thematic, spatial and temporal, which can be also described by knowledge specialized and general.

Next section focuses on related work such as Semantic HMC, semantic measures and Cross-Referencing Methods Proposals. Section 3 deals with our approach to improve cross-referencing using the Semantic HMC. The last section concludes this paper.

2. Related work

This section describes some of the most relevant previous works such as projects and publications. Moreover, it is presented projects more technical oriented that can be found on the web. They provided similar solutions the issue tackled.

2.1 Academic purposes

In [16], authors tried to construct a bilingual dictionary from a corpus using the similarity among concepts by polysemy. The contribution of Sra Suvrit [17] was an enhancing on retrieval of information. In the experiments, they showed the usage of learning dictionaries enabling a fast and accurate Nearest Neighbor (NN) retrieval. The dictionaries work on covariance matrix data sets without using semantic features. Pedersen Ted [18] processes the semantic on the concepts constructing a dictionary melded from various sources. They faced the overlapping among dictionaries. In this approach, they implemented Cross–Level Semantic Similarity (CLSS) which is a novel variation on the problem of semantic similarity. The work of Shahriar, Md Sumon [19] proposed a smart query answering architecture oriented to marina sensor data with a data mining approach. They implemented many processes, but no similarity features were provided to process information. Shvaiko [20] implemented an extension of Geonetwork [26] adding a new interface. They included semantic capabilities by using a faceted ontology, but it is limited to semantic matching operations using S-Match [21] between the query and the ontology. A second work with Farazi [22] exploit this work to provide an enhancement that extended the capability of the queries by giving similar answers. It computes the similarity using a nearest neighbor approach without considering a similarity measure that computes the information from an ontology.

2.2 Services on the web

The project “Aonaware Web Services” [23] presents the possibility of consulting dictionaries for the human understanding. Concepts are introduced by providing an extensively definition. The project ontology Lookup Service [24] provides a web service where it is possible to query multiple ontologies instead of only dictionaries. This service is for human recognition and automatic processing in the documents' retrieval domain. The work of Falcons [25] provides a service of consulting through queries. The system looks up for those concepts into their ontologies lexically. As results, the system shows an excerpt of those ontologies. The system does not use specialized domains, and it is impossible to choose relevant ontologies for the search.

2.3 Discusion

Based on the analysis of related work, experimental results show that the information retrieval is improved by using many dictionaries. Word alignment techniques can be applied on shared vocabulary in dictionaries to face the overlapping of a melded dictionary from several dictionaries. The common issue in the related work was the conceptual ambiguity which can be tackled with the semantic processing.

The common factor regarding smart query processing, semantic retrieval and catalogues is the similarity measure necessity to compute the implicit information from the ontology. The cross–level semantic similarity is a feasible solution for processing shared knowledge among ontologies. An important contribution is the extension of the user's expertise in several areas.

Considering the limitations and main features of presented work, the proposed methodology handles the knowledge from different domains in order to improve the retrieval with mechanisms of collaboration among ontologies. Similarity measures are included with the purpose of expanding the expertise of users on general and specialized knowledge about thematic, spatial and temporal domains.

3. Proposition

This section describes our contribution based on three main stages. The three stages are namely the "knowledge analysis", the "data source analysis" and the "query analysis".

- In the "knowledge analysis" stage, a knowledge base is built composed of a set of ontologies from thematic, spatial and temporal domains. The thematic domains consider general ontologies described by a common vocabulary (common domains) and specialized ontologies use specialized concepts (specific domains or domains not commonly used). When the ontologies are loaded in the knowledge base, a necessary semantic pre-processing is executed in two steps. The first step calculates similar concepts in each ontology on the same domain. The second step computes the similarity of the concepts among ontologies.
- In the "data source analysis" stage, the geospatial data sources are stored. The concepts that compose the metadata description are included in the ontology mapping (concept-data source) in order to link concepts in the ontologies to data sources.
- In last stage called "query analysis", queries are introduced and transformed into smart queries. Actually, the query are extended and linked to ontology vocabularies. Users are able to retrieve geospatial data sources semantically related using knowledge specialized and on the thematic, spatial and temporal domains. In this stage, the semantic pre-processing from the conceptualization stage is used in order define the query and get similar concepts related. The semantic pre-processing is also used in the synthesis stage for obtaining the geospatial data sources related using the ontology population mapping.

Figure 1 presents the overall solution proposed. Some parts are widely used in many projects like the pre-processing of the queries. Our approach is unique in the sense that we exploit concept expansion in each ontology using all the ontology using distance similarity measures. And, the result of this expansion is also exploited for the expansion of the queries which make the results much broader and closer to the need of the user. In order to limit the size of the expansion, the principle of intersection allows the systems to select only the vocabularies that are in common to a set of ontologies after the expansion process of the ontologies.

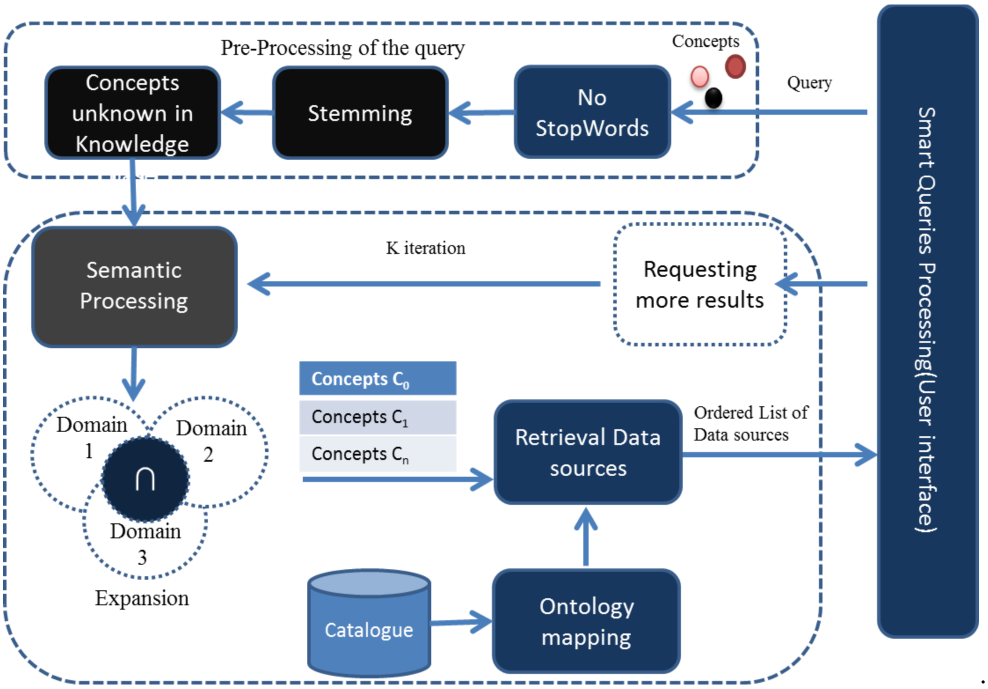

Figure 1. Query analysis stages.

5. Conclusions

We have studied the handling of multi domain knowledge as a way for extending the user expertise in queries processing. This feature can be especially important in the retrieval of data sources in a catalogue. The inclusion of a mechanism of collaborative domains can extend the information in the queries and disambiguating the concepts. The similarity measures provide a semantic approach in the query analysis and in the construction of the catalogue. We have described the work-in-progress that is currently under development. A significant part of the components of the system have been already developed, although they still need to be integrated.

References

[1] M. Gross, “The Unknown in Process Dynamic Connections of Ignorance, Non-Knowledge and Related Concepts”, Sage Publications, Current Sociology, vol 55,5 pp 742-759, 2007.

[2] M. Gross, H. Hoffmann-Riem, “Ecological Restoration as a Real-World Experiment: Designing Robust Implementation Strategies in an Urban Environment”, Public Understanding of Science 14(3): 269–84, 2005.

[3] W. Krohn, J. Weyer, “Society as a Laboratory: The Social Risks of Experimental Research”, Science and Public Policy 21(3): 173–83, 1994.

[4] U. Beck, “World Risk Society”, Oxford: Polity Press,1999.

[5] V. Kashyap, et al, “Semantic heterogeneity in global information”, Cooperative Information Systems: Current Trends and Directions, 1997.

[6] J. Goodwin, “What have ontologies ever done for us - potential applications at a national mapping agency”, In: in OWL: Experiences and Directions (OWLED), 2005.

[7] Md. S. Shahriar, et al, “Smart query answering for marine sensor data”, Sensors, Molecular Diversity Preservation International, vol. 11,3, 2885-2897, 2011.

[8] J. Han, Y. Huang, N. Cercone, Y. Fu, “Intelligent Query Answering by Knowledge Discovery Techniques”, IEEE Trans. Knowl. Data Eng. 8s, 373-390, 1996.

[9] B. Harbelot, H. Arenas, H., C. Cruz, “The spatio-temporal semantics from a perdurantism perspective”, In: Proceedings of the Fifth International Conference on Advanced Geographic Information Systems, Applications, and Services GEOProcessing. February-March, 2013.

[10] ESRI: GIS Best Practices: Spatial Data Infrastructure (SDI), 2010

[11] J. Noragh, “Power of Raven, Wisdom of Serpent”, Floris Books. ISBN 0-940262-66-5, 1995.

[12] A. Borang, et al, “Butterflies of Dihang Dibang Biosphere Reserve of Arunachal Pradesh”, Eastern Himalayas, India, Bullefin of Arunachal Forest Research,vol 24, 41-53, 2008.

[13] Fauna Completeness Ontology. Keith Kintigh. (tDAR ID: 376370); doi:10.6067/XCV8HT2NMV, 2012.

[14] M. Smithson, “Ignorance and Science: Dilemmas, Perspectives, and Prospects”, Knowledge: Creation, Diffusion, Utilization 15(2): 133–56, 1993.

[15] W. E. Dietrich, et al, “Hollows, colluvium, and landslides in soil-mantled landscapes”, Binghamton Symposia in Geomorphology, International Series. Allen and Unwin, Hillslope processes, 361-388, 1986.

[16] L. Xiaodong et al. “Topic models+ word alignment= a flexible framework for extracting bilingual dictionary from comparable corpus”, Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgari, 212-221, 2013.

[17] S. Suvrit et al., “Generalized dictionary learning for symmetric positive definite matrices with application to nearest neighbor retrieval”, Machine Learning and Knowledge Discovery in Databases, Springer.318-332, 2011.

[18] T. Pedersen, “Duluth: Measuring Cross--Level Semantic Similarity with First and Second--Order Dictionary Overlaps”, SemEval 2014, pp.247, 2014.

[19] Md S. Shahriar, et al, “Smart query answering for marine sensor data,Sensors, Molecular Diversity Preservation International ,vol. 11,3, 2885-2897, 2011.

[20] P. Shvaiko et al. “A semantic geo-catalogue implementation for a regional SDI”, University of Trento, 2010.

[21] F. Giunchiglia, P. Shvaiko, M. Yatskevich. “S-Match: an algorithm and an implementation of semantic matching”, In Proc. of ESWS, 2004.

[22] F. Farazi, et al. “A semantic geo-catalogue for a local administration”, Artificial intelligence review, Springer ,vol 40,2.193-212, 2013.

[23] Aonaware Web Services. http://services.aonaware.com/DictService/, consulted on November 2014.

[24] The Ontology Lookup Service. http://www.ebi.ac.uk/ontology-lookup/, consulted on November 2014.

[25] Falcons. http://ws.nju.edu.cn/falcons/objectsearch/index.jsp, consulted on November 2014.

[26] J. Ticheler, et al. “GeoNetwork opensource Internationally Standardized Distributed Spatial Information Management”, OSGeo Journal, vol.2, 1, 2007.