KEYSTONE COST ACTION IC1302, SHORT TERM SCIENTIFIC MISSIONS

-

SCIENTIFIC REPORT of COST-STSM-IC1302-20590

STSM Topic: Improvement of automatic formalization processes for thesauri
STSM Applicant: Javier Lacasta
Applicant's affiliation: IAAA, Universidad de Zaragoza
Applicant's address: Departamento de Informática e Ingeniería de Sistemas. María de Luna, 1. CP:50018, Zaragoza, Spain.
e-mail: jlacasta@unizar.es

1. Introduction

COST is an intergovernmental framework for European Cooperation in Science and Technology that allows the coordination of nationally-funded research on a European level. COST Actions are programs whose main objective to establish a cooperative network of researchers, practitioners, and application domain specialists in different technical fields.

Keystone Cost Action IC1302 focus on the area of keyword-based search over structured data sources. It wants to promote the development of a new paradigm that provides users with keyword-based search capabilities for structured data sources as they currently do with documents. To do that, it is needed to provide advances in fields such as semantic data management, semantic web, information retrieval, artificial intelligence, machine learning and natural language processing.

An important part of any COST Action such as IC1302 is to promote Short Term Scientific Missions that allow scientists to travel to an institution or laboratory in another COST country to exchange ideas and knowledge. These collaborations can lead to more successful projects, and put European scientists at the forefront of worldwide technological innovation.

This document describes the results of the Short Term Scientific Mission number COST-STSM-IC1302-20590, funded by Keystone Cost Action IC1302 to stablish a collaboration between researchers of the university of Zaragoza and Geneva in the information retrieval field.

2. Purpose of the STSM

In the information retrieval context (IR), the resources of a collection are frequently classified and searched using concepts from thesauri and other simple knowledge models. However, since they reflect the vision of those who created and maintain them, they are not homogeneous and they may contain heterogeneous concepts and relations [1]. This lack of semantics limits their usability for IR. Thesauri may be used to expand queries by including the narrower concepts of query terms, but are disqualified for logical inference [2]. Lauser [3] highlights the advantages of using formal ontologies with respect to thesauri, but the complexity of creating them has discouraged for years its use in many fields.

A previous collaboration in this field between the IAAA group of the University of Zaragoza and the ICLE in the University of Genève leaded to the definition of proposals for the automatic formalization of thesauri and their conversion into ontologies [4,5]. This STSM has allowed us to continue our collaboration and produce additional advances in this area. Specifically, the work performed has focused on proposing new approaches to improve the semantics of knowledge organization systems used in resource collections to classify and locate

resources. This includes management of multilingualism, the mining of data collections to discover relevant content, or the improvement of user systems required for collection search and faceted browsing.

3. Description of the work carried out during the STSM

Currently there are hundreds of data collections about a great number of different subjects published in the web though IR based system, relational database web interfaces or even as semantic graphs. The content in these collections are usually described through metadata of different degree of completeness and standardization. These metadata are usually a combination of free text and concepts extracted from different knowledge organization models.

The capability of locating information in these collections greatly depends on how well the resources are described. Description of resources through keywords is an effective way to facilitate IR tasks, especially if the keywords are linked to commonly used knowledge organization models. However, in many cases, these resources contain terminology extracted from an unknown source, and they are heterogeneous between collections from different origin. This makes more difficult information retrieval tasks. The problem may not be only the lack of results caused by terminological issues such as polysemy and synonymy, but also the excess of results of heterogeneous quality that contain the search terms, or the inability of crawler systems to index some collections. Web information retrieval systems such as Google or Bing are able to find documents contained in collections publicly accessible to their crawler systems. However, for general information queries (e.g., books of English authors vs. name of the rose book) common term based techniques do not work well. Semantic search systems are better suited for this task, but there are still few semantic collections and the user needs to know which collections to search and their structure.

Along this STSM we have explored how to describe better the content of a collection and to guide the user in the process of locating the information he is interested in. We think the resource descriptions can be improved by using not only the keywords in an isolate way but as part of an ontology that indicates how these keywords are related in the resources. This ontology can facilitate browsing and search. On the one hand, this ontology can provide the concepts the user is interested in, and it allows navigating through the relations in the model. On the other hand, search systems can make use of the ontology relations, if they represent information in the sources (e.g.: a document is about technicians making clocks).

In this context, we have focused on determining how to create a topic ontology that represents the main elements a collection is about. Additionally, we have also explored how to generate better knowledge organization models for classification using existing ones as base, and collections of resources classified using these models.

We have selected the following use cases in which an ontology such as the indicated provide a relevant advantage for information retrieval:

1- **Keyword based query for documents:** The user search for documents about hydrology and environment. The user browses through the ontology to select the two corresponding concepts. As alternative, the terms can be automatically matched with the ontology concepts comparing the labels. In any case, the IR system searches using the concept identifiers and it returns all the documents classified according both terms.

2- **Keyword based query for collections:** The user searches for databases or collections containing a great deal of information about hydrology and environment. It uses a repository containing the ontologies that describe the content of the available collections/databases. The query system searches in the ontologies for the indicated concepts it returns the location of the relevant collections. It can also be used an ontology that integrates the content of all the collection ontologies to facilitate concept selection. Once a relevant collection has been identified, a more precise search can be launch using the specific IR tools provided for the selected collection.

3- **Relation based query for documents:** The user searches for documents about contamination effect in hydrology. For this query a document with a chapter about contamination and other unrelated one about hydrology is not relevant. What the user wants is documents that describe the effect between them. Therefore, it requires that the ontology represents relations that exists in the documents and not general relations between the concepts. As in the keyword based case, the selection of the terms may be performed using the concepts and relations from the ontology, or aligning a free text query to it.
4- **Relation based query for collections:** The user searches for databases/collections dealing about contamination effect in hydrology. This is the most complex case. For databases, as they are a unity, an ontology extracted from the schema can be used to identify a database containing data about the indicated relation in the same way as in the document case. For resource collections is more difficult, a collection may be considered that is relevant respect a concept relation if it contains many documents that deal with this problem. To be able to identify this, it is needed that the collection ontologies contain weights indicating the relevance of each concept and relation in the collection, so the results can be sorted.
5- **Classification of a new collection:** The user wants to classify a new collection adding keywords to their metadata records. In the selected field there are some simple knowledge organization models used by other collections, but they are incomplete in terminology and relations to be really useful in his context. Therefore, the user uses other collections in the field and existing knowledge organization models to generate the desired ontology.

The STSM has focused in evaluating the complexity of the problem of automatically generating an ontology that can be used in the previous situations. To do so, we have been identified the need to perform information extraction from knowledge organization systems, metadata, text documents and databases. For each relevant part of the process that provides an advance in the knowledge management field, we expect to be able to publish a technical paper in a relevant conference or journal.

Figure 1 shows the proposed architectural prototype of our ontology generation process. We have considered three types of sources from which it is possible to extract useful concepts and relations for the ontology: the data collection itself (databases or text files), the metadata describing the collection, and the knowledge models used to create the collection metadata. We have focused on integrating the concepts and relations extracted from these sources through the alignment with a general purpose ontology such as WordNet. The integrated model then can be pruned to adjust it to the source collection or leave at it is for a more general classification use. Finally, to refine the provided relations, the integrated model needs to be aligned with a top level ontology such as Dolce. Following subsections describe in detail the main tasks and objectives to accomplish in each part of the process.

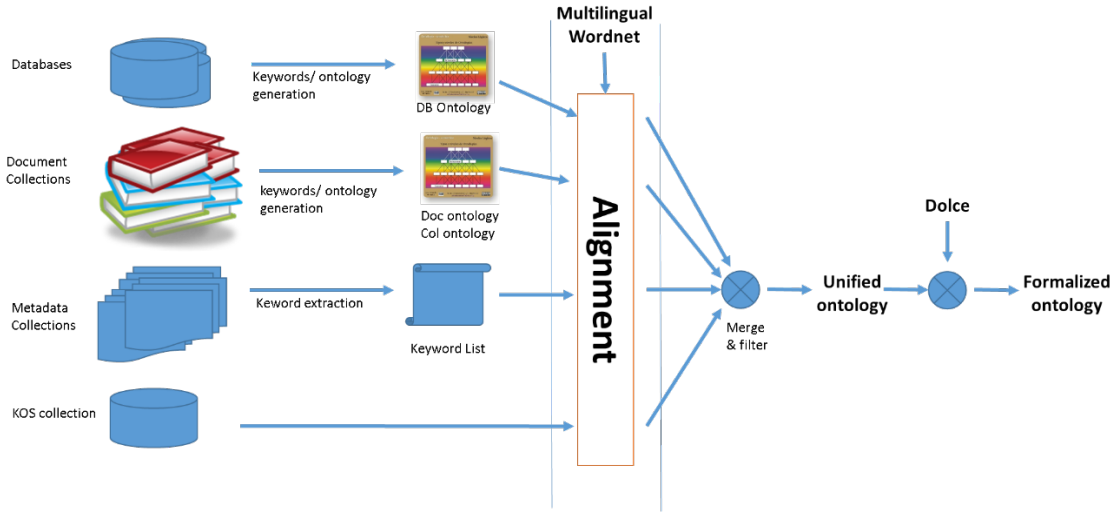

Figure 1: Ontology generation process, extension of [5]

Extraction of keywords and relations from data collections

We have decided to focus on extracting concepts and relations from databases and collections of documents using and improving existent approaches that can be applied to our context.

With respect to extraction of concepts and relations from text, we have identified several relevant works to analyze [6-15]. The host institution already have done some work in this area for generating hyper-books form text documents [16,17] and we want to use it as base for this component. Additionally, we want to compare the differences between applying the selected keywords and relation extractions techniques to complete documents, to parts of it such as chapters or paragraphs, and even to the whole collection. Since the statistical distribution of words is different, results may vary. We want to observe how the results to be able to select the better approach.

The extraction of keywords and relations from databases has to follow a different process. A database is a collection itself, being the resources the rows in each table. In this context, we want to extract concepts and relations from the database schema information and textual fields. To do so, we plan to use existing data mining techniques and IR approaches used to transform keyword searches into SQL queries [18-21].

It is important to note that, using natural language processing on textual documents/fields it is possible to extract relations between the terms [22,23]. These relations may be very valuable in IR context, as they indicate how the terms are related in the document. For example, a document about environment and tornadoes, may describe these elements in a very different ways. For instance, it can indicate how tornados affect environment (floods, dead of animals…), or it can describe environment elements that generate tornadoes (temperature, pressure, climatic change …).

Metadata concept extraction

Metadata describe the content of a resource using a structured fields that contain a combination of free text and terms from controlled vocabularies. While controlled terms are used for keyword based searches, free text fields, such as title, are thought to help users to identify the resources in the search results. However, in many cases, the origin of the used keywords is unknown or heterogeneous. This makes the work of information retrieval systems more difficult as they have to deal in their indexation and search processes with terminological heterogeneity that could be eliminated with better quality metadata. Alignment of keywords with thematic thesauri of the area of interest is a solution for this problem. However, in many contexts, this is not enough. Metadata may not contain suitable keywords or the thesauri used for alignment may not represent the collection content. Therefore, we have decided to explore different lines. On the one hand, we want to process the free text fields in metadata to identify new classification terms (free text extraction techniques are applicable). On the other hand, we want to test different alignment variants between keywords and thesauri and even between the keywords and a general ontology such as WordNet or SUMO to determine which one produce best results [24,25,26].

Knowledge organization model management

Documents collections use thematic terms for classification. These terms may be obtained from thesauri and other simple knowledge organization models. These concepts and their relations are basic to construct the desired collection ontology as, in general, they are already suitably structured and they can be directly used in our process.

Alignment with a general purpose ontology

A solution to integrate the concepts and relation obtained from the different sources (collection, metadata, and knowledge organization models) is to align all these concepts with a general domain ontology such as WordNet or Sumo. We already have performed work in this area proposing a process to align a thesaurus into WordNet [5]. The work in this area is planned to be a continuation of this work. We have decided to focus on improving existing alignment techniques, with specialized processes for each type of source. For example, keywords from documents or databases may have relations that can be used in the alignment; and metadata keywords may have implicit relations that can also facilitate the alignment, between others [27,28]. With respect to knowledge organization models such as thesauri, we have decided to extend the already used alignment techniques to take into account multilingual thesauri and multilingual versions of the ontology to improve alignment [29,30]. Also, we need to study existent representation formats [31,32], as the generated model can be more easily distributed if it is standard.

Generation of topic resource/collection ontology

The objective of this step is to take the general purpose ontology with all the alignments previously identified and generate a new model that describes the content of the collection. Here, we must distinguish between ontologies describing the content of a document, and ontologies describing the content of a collection (or database). Our approach to construct these ontologies is based on integrating all the information obtained in the previous tasks (metadata, documents, and thesauri) and integrating it in a single model that is pruned from the non-relevant parts. The new ontology may have a different structure than the sources depending on occurrences of relations. This model can be then aligned to a top level ontology such as DOLCE to add better semantic to the existing relations [33]. To do this, we want to integrate manual mappings to improve the already created alignment process [5].

A parallel line of work is to use the same approach and techniques to generate an independent ontology about a certain subject. In this case, it requires the use of thesauri focused in the desired theme, and a data collection about the theme. The unpruned ontology is expected to contain a better description of selected subject. Information of different thesauri and resources may complete each other generating a much more suitable model for classification and search in other new collections. To do so, we have decided to explore different alternatives for information integration. A first step in this direction will be to simply extend existing knowledge organization models with concepts extracted from the associated collection. The final objective is to allow us to generate a model with complete is-a or part-of hierarchies, instead of the fragmented ones that current approaches provide. In this context, it will be necessary to determine the quality of an ontology in a numerical way to compare results. For information retrieval, it can be measured in based to the results provided by the system that uses it, for classification or browsing it depends on the user satisfaction.

Finally, we are also interested on how the document and collection ontologies can replace keywords in metadata. This model has to be tested in an IR context to determine its usability and performance for creating a relation based IR system.

4. Future collaboration with the host institution

Our objective for the following months is to begin the implementation and testing of all the different components previously described. Due to the complexity of some parts and the extension of other ones, we have decided to promote end of degree projects and PHD proposals in these subjects in the universities of Zaragoza and Geneva.

Our planning in order to construct the proposed system is the following. We will start with the improvement of the alignment process described in [5] to formalize knowledge organization models using multilingual models and improving WordNet – Dolce alignment through better algorithms and existent partial manual alignments. We will also test the process with other general purpose ontology such as SUMO to determine the differences.

Then, we will continue with the metadata keyword and alignment extraction. This will provides us a context related keyword source that, integrated with the concepts from selected knowledge organization models, will allow us to generate a first version of the desired collection ontology. Here, we will have to give solutions to the problems of distinguish the concepts that best represent the collection from those indicated in the metadata and the existent ones in the used general purpose ontology. The third development phase will focus on the extraction of keywords and relations from a collection of text documents. The used techniques will also be tested with textual metadata fields such as the abstract. The alignment of the extracted concepts with the used general purpose ontology, and the integration with previous metadata keywords and knowledge organization systems is expected to improve the final ontology. In this context, extracted relations will allow us to create an ontology focused on describing the collection content instead of a general ontology in the field. The final step will be to perform the extraction of keywords and relations from a database to determine if they can be classified in an equivalent way as text collections.

To test each process component and measure their performance we will need to use suitable collections. We need complete metadata that use available knowledge organization models, and full access to associated documents or the database storing the information. We have made an initial pre-selection of collections we already have experience working with or that provide easy access to the required information.

- Wikipedia: Online encyclopedia. It provides an easy download mechanism with full text access. As metadata, it provides info-boxes with structured information. Additionally, it provides a folksonomy used for documents classification. Moreover, DBpedia and their categories can be used as context of the Wikipedia to facilitate information extraction (http://en.wikipedia.org)
- Zaguan: Digital repository of the University of Zaragoza. It provides easy access to end of degree and thesis works done by students of the university. Each document is described through metadata and classified using keywords from a controlled vocabulary (http://zaguan.unizar.es).
- Aviation Safety Network: It provides a catalog of aircraft accidents with metadata describing them and full reports detailing their circumstances. It provides controlled vocabularies for types of aircrafts and their parts, operators, and airports (http://aviation-safety.net).

5. Bibliography

[1] D. H. Fischer, From thesauri towards ontologies?, in: Structures and relations in knowledge organization - 5th International ISKO Conference, Lille (France), 18{30, 1998.

[2] D. Soergel, Indexing Languages and Thesauri: Construction and Maintenance, Melville Pub. Company, 1974.

[3] B. Lauser, From thesauri to Ontologies. A short case study in the food safety area in how ontologies are more powerful than thesauri, Agricultural Information and Knowledge Management Paper, Food and Agriculture Organisation of the United Nations, Rome (Italy), 2004.

[4] J. Lacasta, J. Nogueras-Iso, J. Teller, G. Falquet, Transformation of a keyword indexed collection into a semantic repository: applicability to the urban domain, in: International Conference on Theory and Practiceof Digital Libraries, vol. 6966 of LNCS, Berlin (Germany), 372-383,2011. (Best paper in the conference)

[5] J. Lacasta, J. Nogueras-Iso, G. Falquet, J. Teller, F.J. Zarazaga-Soria. Design and evaluation of a semantic enrichment process for bibliographic databases. Data & Knowledge Engineering. 2013, vol. 88, p. 94-107. ISSN 0169-023X.

[6] Cimiano, P and Völker, J, Text2Onto. Natural Language Processing and Information Systems. Lecture Notes in Computer Science. 2005, p. 227-238.

[7] Magnini, B., Negri, M., Pianta, E., Romano, L., Speranza, M., Serafini, L., Sprugnoli, R. (2005). From Text to Knowledge for the Semantic Web: the ONTOTEXT Project. In SWAP (Vol. 166).

[8] Buitelaar, P., & Cimiano, P. (Eds.). (2008). Ontology learning and population: bridging the gap between text and knowledge (Vol. 167). Ios Press.

[9] Brown, E., Epstein, E., Murdock, J. W., Fin, T. H. Tools and Methods for Building Watson.

[10] Fan, J., Kalyanpur, A., Gondek, D. C., Ferrucci, D. A. (2012). Automatic knowledge extraction from documents. IBM Journal of Research and Development, 56(3.4), 5-1.

[11] Balakrishna, M., Srikanth, M. (2008, December). Automatic ontology creation from text for national intelligence priorities framework (NIPF). In Proceedings of 3rd International Ontology for the Intelligence Community (OIC) Conference (pp. 8-12).

[12] Balakrishna, M., Moldovan, D. I., Tatu, M., Olteanu, M. (2010, May). Semi-Automatic Domain Ontology Creation from Text Resources. In LREC.

[13] Booshehri, M., Zamanifar, K., Shariatmadari, S., Zeini, S. M. A New Layer for Automatic Ontology Creation from Text by Using Linked Data and Implied Information.

[14] Biemann, C. (2005). Ontology Learning from Text: A Survey of Methods. In LDV forum (Vol. 20, No. 2, pp. 75-93).

[15] Maedche, A., Staab, S. (2004). Ontology learning (pp. 173-190). Springer Berlin Heidelberg.

[16] Falquet, G., Ziswiler, J. C. (2005). A virtual hyperbooks model to support collaborative learning. International Journal on E-Learning, 4(1), 39-56.

[17] Falquet, G., Nerima, L., Ziswiler, J. C. (2005, September). Augmented hyperbooks through conceptual integration. In Proceedings of the sixteenth ACM conference on Hypertext and hypermedia (pp. 132-134). ACM.

[18] Cullot, N., Ghawi, R., Yétongnon, K. (2007, June). DB2OWL: A Tool for Automatic Database-to-Ontology Mapping. In SEBD (pp. 491-494).

[19] Cerbah, F., Aviation, D. (2009). RDBToOnto: un logiciel dédié à l'apprentissage d'ontologies à partir de bases de données relationnelles. et gestion des connaissances: EGC'2009.

[20] Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., Aumueller, D. (2009, April). Triplify: light-weight linked data publication from relational databases. In Proceedings of the 18th international conference on World wide web (pp. 621-630). ACM.

[21] Cerbah, F. (2008, December). Mining the content of relational databases to learn ontologies with deeper taxonomies. In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on (Vol. 1, pp. 553-557). IEEE.

[22] Fundel, K., Küffner, R., Zimmer, R. (2007). RelEx—Relation extraction using dependency parse trees. Bioinformatics, 23(3), 365-371.

[23] Byrd, R. J., Ravin, Y. (1999). Identifying and extracting relations in text. na.

[24] Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Zhou, J. (2006). Dolce ergo sumo: On foundational and domain models in swinto (smartweb integrated ontology). Submission to Journal of Web Semantics.

[25] Agerri, R., San-Sebastián, D., Bermudez, J., Rigau, G., Sebastián, D. S. (2014). Multilingual, Efficient and Easy NLP Processing with IXA Pipeline. EACL 2014, 5.

[26] Mascardi, V., Locoro, A., Rosso, P. (2008). Exploiting DOLCE, SUMO-OWL, and OpenCyc to boost the ontology matching process. Technical Report DISI-TR-08-08, University of Genoa, 2008. http://www. disi. unige. it/person/MascardiV/Software/OntologyMatchingViaUpperOntology. html.

[27] Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries, 3(2), 115-130.

[28] Gaio, M., Sallaberry, C., Tien, V. (2013). Typage de noms toponymiques à des fins d'indexation géographique. Traitement Automatique des Langues, 53(2), 143-176.

[29] De Melo, G., Weikum, G. (2009, November). Towards a universal wordnet by learning from combined evidence. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 513-522). ACM.

[30] Bond, F., Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In ACL (1) (pp. 1352-1362).

[31] McCrae, J., Spohr, D., Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In The Semantic Web: Research and Applications (pp. 245-259). Springer Berlin Heidelberg.

[32]Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006). Lexical markup framework (LMF). In International Conference on Language Resources and Evaluation-LREC 2006.

[33] Li, Y., Wang, Y., Huang, X. (2007). A relation-based search engine in semantic web. Knowledge and Data Engineering, IEEE Transactions on, 19(2), 273-282.