A service based on Linked Data to classify Web resources using a Knowledge Organisation System

A implementation to classify Open Educational Resources

Janneth Chicaiza, Nelson Piedra and Jorge López
Universidad Técnica particular de Loja
Departamento de Ciencias de la Computación
Loja, Ecuador
{jachicaiza, nopiedra, jalopez2}@utpl.edu.ec

Edmundo Tovar Caro
Universidad Politécnica de Madrid
Dpto. Lenguajes y Sistemas Informáticos e Ingeniería
Software, Madrid, Spain
edmundo.tovar@upm.es

***Abstract*—**One of the reasons why Web resources could stay hidden and therefore to be underutilized is that each person or producer of this kind of resources, labels them using tags or informal and heterogeneous knowledge schemes. The lack of semantics in the relations between areas and subjects make difficult to find associations between topics. It is possible to support semi-automatic classification of resources, taking advantage from Linked Data available in the Web through systems made by people who can converge to a formal knowledge fields classification system. The representation of areas or subjects and their relationships through semantic technologies will help the discovery of such kind of resources for user at worldwide. This paper presents the design of a service, which takes advantage from the potential of Linked Data available on the Web for the classification of Open Educational Resources according to subject areas and to improve the discovery of courses by users.

Keywords— Linked Data, classification, knowledge organisation system, Web of Data, thesaurus, social sources

I. INTRODUCTION

In repositories or collections of high amounts of information objects, such as libraries, different classification systems have been used in order to facilitate the location or recovery of academic materials. To perform this task, a closed schema of knowledge organization is generally chosen -proposed by an organization or by experts-, and a librarian is responsible for determining the most appropriate topic for a given resource.

In the Web, new ways and means to create content have emerged thanks to the ease of use of the social services. Wikipedia is one of the most successful cases, it is the encyclopedia created by people who is constantly editing it.

One of the facilities offered by the current Web 2.0 services is the human cataloging of content, i.e., people add tags and metadata (tags) to Web resources in an ascending order (bottom-up) without expert cataloguers. This open classification system is called folksonomy, the "voice of the people" [1]

In addition to the annotations created by people, today is possible to find on the Web, large amounts of descriptions on resources and entities such as subjects, people, licenses, and locations, in dozens of open repositories of structured data. The Linking Open Data cloud currently provides access to hundreds of datasets in various areas such as Media, Geography, Publications, Government, and Life Sciences. As a consequence of Linking Open Data community project, datasets in a wide range of domains are now semantically described and connected to each other. The language recommended by W3C to describe Web resources is RDF[1] (Resource Description Framework) and SPARQL[2] (SPARQL Protocol and RDF Query Language) that is the language to query RDF data from different repositories across the network.

Thanks to the emergence of the Semantic Web technologies some approaches have been proposed to take advantage of the linked data and to face to the problems and complexity of locating and classifying open resources. Firstly, this paper shows a method for the categorization of resources, which takes advantage of the capacity of organization of the controlled vocabularies and the capacity of enrichment that the open vocabularies and repositories created socially can provide.

Next, the section II put forward the problem addressed in this paper and also the fundamentals of the approaches used to solve it. In section III the framework based on Linked Data to characterize Web resources is presented. A proof of concept, it is explained in section IV. Finally, in section V the conclusions and future works are presented.

II. KNOWLEDGE ORGANISATION SYSTEMS TO CLASSIFY WEB RESOURCES

The classification systems of knowledge, such as thesauri, have traditionally been used to improve the organization and retrieval of documents. Make a better categorization of resources according to the subject areas could be a mechanism of guidance to help people find the documents that they need.

Therefore, there are a clear need to use thesauri and semantic models to provide a controlled source of terms or concepts, this is an essential pre-condition to guarantee quality

[1] http://www.w3.org/RDF/

[2] http://www.w3.org/TR/rdf-sparql-query/

in document indexing and retrieval [2]. Some thesauri or controlled vocabularies used in libraries to classify documents are now available in the formal languages of Semantic Web.

Semantic Web technologies and Linked Data [3] are changing the way information is stored, described and exploited. The Linked Data Design Issues, outlined by Tim Berners-Lee back in 2006, provide guidelines on how to use standardized Web technologies to set data-level links between data from different sources [4]. Through the life cycle of Linked Data can be enrich, disambiguate, connect and retrieve data from heterogeneous domains, repositories or systems.

A. Knowledge organization systems

In the academic field, there are some thesauri to classify knowledge. The decimal systems DDC[3] and UDC[4] are the most widely used in libraries, however, they are proprietary schemas. JACS[5] is one of the most complete thesauri and updated, though it is available only in English and its adoption has been restricted to the United Kingdom. Finally, EuroVoc[6] and UNESCO thesaurus[7] are multilingual and multidisciplinary systems, however, the first one covers the activities of the European Parliament and the second one is more popular. The decision of which system is better depends on the context and uses.

In Semantic Web, Simple Knowledge Organization System (SKOS) is used for representing mapping relationships among systems of classification of knowledge. In 2009, SKOS[8] reached Recommendation status at W3C because it provides a standard way to represent knowledge organisation systems using RDF to describe concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies and other types of controlled vocabulary, thus guaranteeing interoperability among applications[9].

In SKOS, the elements of a thesaurus are represented by means of concepts among which there are established hierarchic relations. The properties `skos:broader` and `skos:narrower` are used to assert a direct hierarchical link between two SKOS concepts.

B. Using sources collaboratively created to classify resources

One of the aspects that is questionable in a traditional thesaurus is the lack of update. Therefore, a formal classification approach can be enriched with the categories that people spontaneously shared on social sources.

In reference [5] some features are mentioned for the use of social data sources for the organization of knowledge: i) constitute some of the largest repositories built in a collaborative manner, ii) provide an up-to-date channel of information and knowledge on a large number of topics. Thus, one should expect a high coverage of subjects that are emergent [6].

If in the Social Web, the Wikipedia is the major case of success, in the Semantic Web, DBPedia[10] is the most popular structured Web data sources.

The DBPedia ontology enables a broad coverage of entities in the world and allows entities to bear multiple overlapping types; it includes RDF data derived from Wikipedia; each resource is harvested from a Wikipedia article (which content is maintained by thousands of editors and it broad and multilingual). [5] In addition, DBPedia resources are linked to other linked data sources and ontologies such as Geonames, YAGO, OpenCyc, and WordNet, providing more semantic information in the form of relations such as typeOf and sameAs. [7]

In [5, 6, 7, 8, 9] the use of DBPedia is addressed to annotate, to enrich and to classify content. According to [7] DBpedia resources are a good starting point to entity recognition due to the fact that a huge part of the knowledge base is related to classes in the DBpedia Ontology.

C. Approaches for subject classification

Social content, such as posts or micro-posts, has been classified according to categories created by people through social services.

In order to classify objects in topics or subjects, different techniques of learning machine have been used. In [7] an unsupervised method in combination with NLP (Natural Language Processing) and semantic-based techniques are used to identify the main topic in posts. In [6] a subset of blog posts were classified by topics using supervised learning machine techniques. In these cases, a considerable effort can be required to obtain training and test data and to construct the underlying classification model. To avoid this problem, in [9] a method to categorize blogs using a domain dictionary is proposed, it uses Wikipedia categories and links between articles to predict concepts common to a set of documents.

In summary, approaches like [6, 7], except [10], do not take advantage of the hierarchic relations between concepts of a knowledge organisation system. Unlike the traditional methods of classification, in this work we have put forward the use of an algorithm to exploitation of graph structures and semantic relations between categories.

III. DESCRIPTION OF THE PROCESS

So that resources can be discoverable, they should be characterized and classified according to main topics enriched with a variety of related tags.

Next, the general process for the topical classification of a resource is presented which try to harnesses the graph structure of the Data Web and knowledge created collaboratively.

A. Data Processing

[3] Dewey Decimal Classification: https://www.oclc.org/dewey.en.html

[4] Universal Decimal Classification: http://www.udcc.org/

[5] Joint Academic Coding System: http://www.hesa.ac.uk/content/view/1787/281/

[6] European Training Thesaurus: http://eurovoc.europa.eu/

[7] http://databases.unesco.org/thesaurus/

[8] http://www.w3.org/2004/02/skos/

[9] http://www.w3.org/TR/skos-reference/

[10] http://dbpedia.org/About

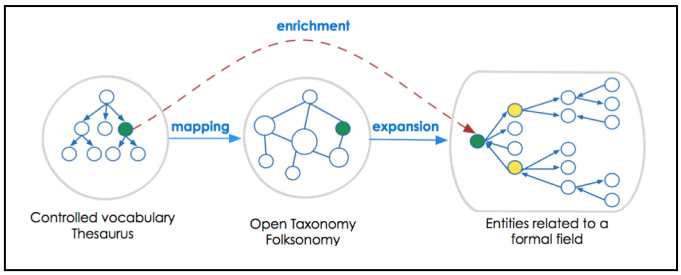

Fig. 1. Flow to data enrichment

Goal: Enriching the subjects of a thesaurus (controlled vocabulary) from semantic resources found on open sources of RDF data. The result will be a catalog of entities (people, locations, concepts, etc.) related to a specific topic.

Tasks:

- Mapping between controlled and formal classification system and open or collaborative organization system.
- Enriching main concepts through topics existing in social data sources. In this research we proposed the use of the spreading activation algorithm [10] for traversing across the sources of linked data. The main idea of spreading activation is that it is possible to retrieve relevant resources if they are associated with other resources by means of some type of connection [8].

B. Document Processing

Goal: Recognizing key entities into each resource of corpus. Entities are DBPedia resources collected in previous stage.

Tasks:

- Performing a semantic annotation of information objects.
- Making a semantic derivation to generate new mappings from the semantic relations defined by SKOS.

C. Classification Layer

Goal: Finding a list of the main topics or subjects, which can fit a particular resource.

Task:

- Accessing to the local knowledge base by mean of grouping queries, in order to obtain the concepts associated with the entities recognized in the content of a resource.
- Creating an ordered list of topics according to their frequency and weight within a field or sub-domain specific.

In addition, in this layer a service of rating of topics could be incorporated so the system can learn based on the decision of the people.

IV. Implementation of the Process for the Classification of Open Educational Resources

In the last years, the amount of Open Educational Resources (OER) on the Web has increased dramatically. In 2001, came up the first educational resources in open, the MIT OpenCourseWare of the Massachusetts Institute of Technology. Currently, dozens of institutions have adopted this philosophy and have made available to the world, thousands of those courses.

In this paper, the authors propose to characterize an OER according to the fields and disciplines of a controlled thesaurus, in order to improve the discoverability and exploration of these resources by the search tools and users.

Table I enlists the sources selected to perform the method proposed in this work.

TABLE I. Corpus Used to Classify Resources

Source	Description	Access point
Corpus: Serendipity LOCWD	Serendipity Linked Open Data includes data about OpenCourseWare and Open Educational Resources.	http://datahub.io/dataset/serendipity
Controlled vocabulary: UNESCO Thesaurus	We chose the system of 6-digit of the nomenclature UNESCO. Since 2013, there is an implementation of this thesaurus according to SKOS. As case of implementation of the framework, Computer Science have been chosen field	http://skos.um.es/sparql/
Open vocabulary: DBPedia	Currently DBpedia has 995.911 instances of SKOS concept and 4'125.606 resources related to these concepts by dcterms:subject predicate	http://dbpedia.org/sparql

A. Data Processing

1) Data Collection

A script implemented in python has allowed recovering information of every discipline and sub-discipline of Computer Science field of UNESCO thesaurus (source dataset). The SPARQL query showed in Fig. 2 was executed to collect data about UNESCO concepts.

```
from SPARQLWrapper import SPARQLWrapper, JSON
import sys
sys.path.append('..')
from generic.bdr import *

predicateList = [
'http://www.w3.org/2004/02/skos/core#prefLabel',
'http://www.w3.org/2004/02/skos/core#narrower',
'http://www.w3.org/2004/02/skos/core#notation',
'http://www.w3.org/2004/02/skos/core#inScheme',
'http://www.w3.org/2004/02/skos/core#broader'
]

b = BDdatos()
con = b.conectar()

sparql = SPARQLWrapper("http://skos.um.es/sparql/")

for predicate in predicateList:
    sparql.setQuery("""
        PREFIX u: <http://www.w3.org/2004/02/skos/core#>
            SELECT DISTINCT ?s ?label WHERE {
            GRAPH <http://skos.um.es/unesco6>
            { ?s a u:Concept.
            ?s <%s> ?label}
        }""" %(predicate))
    sparql.setReturnFormat(JSON)
    results = sparql.query().convert()
```

Fig. 2. Extract of the python script to collect data

From each title of topics, the service Sem4tags[11] was used to choose the DBPedia semantic entity that better defines each concept. Applying this method in Table II, the first five associations found are showed. In summary, of the 27 sub-disciplines of the Computer Science field, 16 were found based on similarity of titles between the two sources, 10 were mapped manually.

TABLE II. EXTRACT OF MAPPINGS BETWEEN UNESCO AND DBPEDIA RESOURCES

UNESCO resource	UNESCO concept	DBPedia resource
unesco:12	Mathematics	dbpedia:Category:Mathematics
unesco:1203	Computer Science	dbpedia:Category:Computer_science
unesco:120301	Accounting	dbpedia:Category:Accounting
unesco:120303	Analog computing	dbpedia:Category:Analog_computers
unesco:120304	Artificial Intelligence	dbpedia:Category:Artificial_intelligence

2) Entity Enrichment

To find entities (topics, tags, people) related to each of the sub-disciplines of Computer Science, an iterative querying process was performed to traverse DBPedia according to the spreading activation method.

Applying this method, the categories with more related resources were: Artificial Intelligence (1215), Data Banks (1416) and Hybrid Computing (4205). By contrast, the categories for which fewer resources were found are: Accounting (20), Automated manufacturing systems (36) and Computer and software (39).

B. Document processing

In this implementation, the Ontotext semantic platform[12], KIM, was used to index and annotate each OER from corpus.

A customization of KIM was needed in order to the system can recognized entities related to each discipline of Computer Science field. In the Fig. 3, an example of key features found on the content of an OER is depicted.

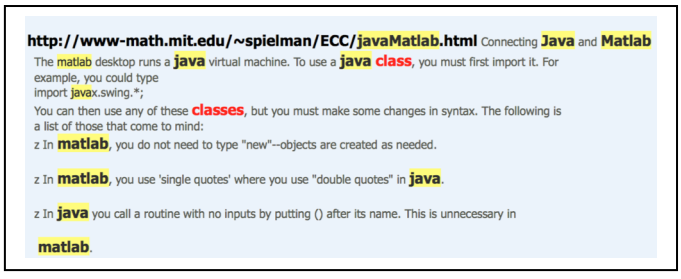

Fig. 3. Semantic annotations found in OER about programming languages

C. Categorization according a sub-discipline

To illustrate the principle of categorization using the spreading activation algorithm, we expose a case of a resource entitled “Connecting Java and Matlab”.

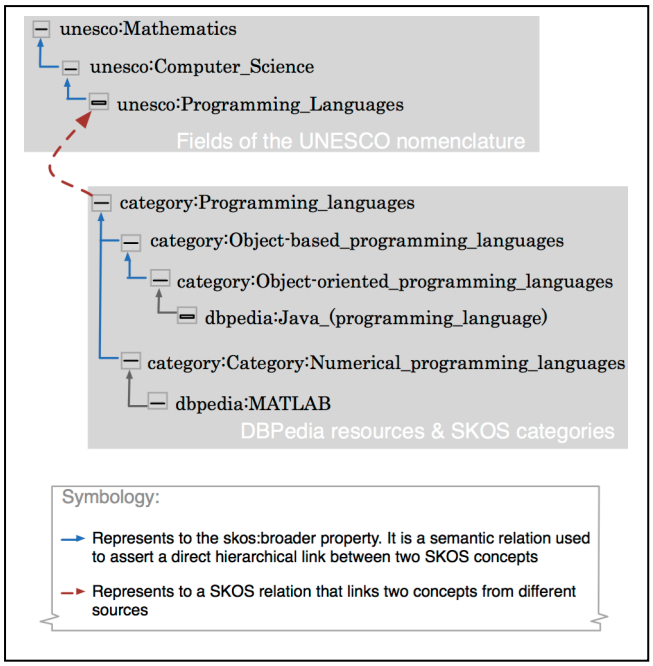

Fig. 4. Hierarchically related concepts found in the “Connecting Java and Matlab” resource

Through the process of annotation, were found as main entities: Java[13] and MatLab[14]. From these two resources recognized in the content of the resource starts the traversing through the hierarchical relationships that link resources. The iterative process should end when it is found a DBPedia category equivalent to a UNESCO sub-discipline. Fig. 4 shows the path between annotated entities and the goal category "Programming Language".

V. CONCLUSIONS

In this work we tried to demonstrate that it is possible to support semi-automatic classification of Web resources, taking advantage from linked data available in the Web through systems made by people, which can be used to enrichment a formal knowledge fields classification system. Different systems or applications could make inference on subject demanded by a user and could display recommendations to get more relevant resources to support several tasks.

Through a concrete case, we tried to demonstrate that is possible to find semantic relationships between two different schemas of knowledge organization: thesaurus-folksonomy. In the specific case of this study, it has been shown that it is possible to classify a resource according to the thematic areas defined by the UNESCO Thesaurus, based on the entities found in the content of the resource and whose descriptions are obtained from DBPedia.

[11] http://grafias.dia.fi.upm.es/Sem4Tags/

[12] http://www.ontotext.com/products/ontotext-semantic-platform/

[13] http://dbpedia.org/resource/Java_(programming_language)

[14] http://dbpedia.org/resource/MATLAB

In order to release this service, we follow testing the results. In a second phase we will test the effectiveness of the system against the cataloging made by experts.

ACKNOWLEDGMENT

The work has been partially funded by scholarship provided by the "Secretaría Nacional de Educación Superior, Ciencia y Tecnología e Innovación" of Ecuador (SENESCYT).

REFERENCES

[1] E. Kroski, "The hive mind: Folksonomies and user-based tagging" (Tech. Rep.). InfoTangle Blog, 2005

[2] E. Francesconi, S. Faro, E. Marinai and G. Perugi, "A Methodological Framework for Thesaurus Semantic Interoperability," Proceeding of the Fifth European Semantic Web Conference, 2008, pp. 76-87.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," International Journal on Semantic Web and Information Systems, vol. 5, iss. 3, 2009, pp. 1–22.

[4] C. Bizer, "The Emerging Web of Linked Data," Intelligent Systems, IEEE, 24(5), 2009, pp. 87 –92.

[5] A. E. Cano, A. Varga, M. Rowe, F. Ciravegna and Y. He, "Harnessing Linked Knowledge Sources for Topic Classification in Social Media," Proceedings of the 24th ACM Conference on Hypertext and Social Media , 2013, pp. 41 - 50.

[6] S. D. Husby and D. Barbosa, "Topic Classification of Blog Posts Using Distant Supervision," Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, págs. 28 - 36

[7] O. Muñoz-García, A. García-Silva, O. Corcho, M. de la Higuera and C. Navarro, "Identifying Topics in Social Media Posts using DBpedia," Proceedings of the Networked and Electronic Media Summit (NEM summit 2011), 2011, Torino, Italia.

[8] Z. Syed, T. Finin, and A. Joshi, "Wikipedia as an ontology for describing documents," In Proc. of the Second Int. Conference on Weblogs and Social Media. AAAI Press, 2008.

[9] C. Hashimoto, and S. Kurohashi, "Blog Categorization Exploiting Domain Dictionary and Dynamically Estimated Domains of Unknown Words," Proceedings of ACL-08: HLT, 2008, pp. 69-72.

[10] A. Troussov, D. Parra, and P. Brusilovsky, "Spreading Activation Approach to Tag-aware Recommenders: Modeling Similarity on Multidimensional Networks," In Proceedings of Workshop on Recommender Systems and the Social Web at the 2009 ACM conference on Recommender systems, RecSys '09, New York, NY, October 25, 2009.