Identifying Common Research Areas: A Study Case

Víctor Saquicela*, Jorge Bermeo*, Mauricio Espinoza*, Kenneth Palacio-Baus†, Boris Villazón-Terrazas‡
*Department of Computer Science, University of Cuenca, Ecuador
{victor.saquicela, jorge.bermeo, mauricio.espinoza}@ucuenca.edu.ec
†Department of Electrical&Electronic Engineering and Telecommunications, University of Cuenca, Ecuador
kenneth.palacio@ucuenca.edu.ec
‡Intelligent Software Components, iSOCO, Madrid, España
bvillazon@isoco.com

***Abstract*—Currently, there is an increasing presence of researchers datasets (services) in the Internet. In this paper, we present an approach for extracting publications made by different authors and identifying common research areas among them. This work makes use of semantic technologies in order to describe authors and their publications through keywords clustering techniques involving data mining algorithms.**

***Keywords*—Clustering, Semantic Web, Research Areas, BIBO**

I. Introduction

The rapidly changing academic environment present in Universities is characterized by a growing number of researchers or postgraduate students creating new publications and constantly facing the diversification of their research topics. Identifying similar knowledge research areas has become a prerequisite in promoting collaboration between researchers interested in proposing new projects in a given field. In particular, for private companies and public organizations identifying common interests among their researchers constitute a major concern. One example of this is the case of the Ecuadorian Government, which in recent years has promoted a massive overseas education program aimed to improve the national human talent. In the domain of research, several types of datasets or services have been published and accessed through APIs in order to make researchers information available.

Keeping this in mind, our aim is to create a system that makes it possible to identify common research areas among a list of authors as input. To achieve this goal, we have defined a procedure that follows these steps: (1) the extraction of authors by the automatic invocation of services using the OAI-PMH[1] standard, (2) the extraction of publications of authors, (3) the semantic description of authors and their publications with respect to an ontology, and (4) the application of data mining techniques (clustering) to detect common areas through the keywords present in publications. We combine the ideas of several initiatives, and propose a new system focused on the identification of common research areas.

The remainder of this paper is structured as follows: First, we present the background and related work done in the domain of identifying common research areas. Then, we describe an scenario that shows the problems currently found in this context. Next, we present the architecture of the system. Finally, we present some conclusions and identify future lines of work.

[1] www.openarchives.org/pmh/

II. Background and Related Work

This section provides a brief introduction to the DSpace[2] repository and related tools used in searching for publications. Moreover, the existing approaches related to the identification of common research areas are described.

A. DSpace Repository of Dissertation Authors

DSpace is an OpenSource platform that allows the management and distribution of digital contents on the Web, using a workflow based on publication requests and a series of programmable filters. DSpace acts as a repository for digital research and educational material produced by a particular organization or institution [7]. From a technical point of view, DSpace is implemented in Java and uses PostgreSQL as its database.

B. Searching for Publications

Currently, several tools found in the Internet allow people to find scientific publications from specific authors. In general, these search tools rely on the use of keywords, however, in recent years many of these tools have started using semantic technologies to describe the authors' publications. As an initial approach, this work proposes the use of tools based on keywords searching only. Next, we describe the tools that have been used in this work to find the publications about a specific author.

There is a wide variety of information sources related to academic articles. A complete list of these sources can be found at *http://libguides.mit.edu/apis*. A disadvantage commonly found among them is that they do not have an API that allows access for information retrieval, and therefore, the need of further interpretation of the obtained search results. Next, three of the most popular academic platforms are analyzed:

- Google Scholar[3] does not have an API that allows automatic publications searching, however, there is an unofficial API that allows searching by title, author or keyword in order to automatically extract the following fields: title, URL, number of citations, number of versions, links to citations, and links to versions.
- Microsoft Academic[4] provides a REST API for publications searching. The results obtained from a query

[2] http://www.dspace.org
[3] http://www.icir.org/christian/scholar.html
[4] http://academic.research.microsoft.com

are in JSON format containing the fields: title, abstract, keywords, authors, number of citations, year, and URL.

- IEEE Explore[5] is a search service for publications. This service has an API aimed to automatically perform searches based on different fields (author, title, keywords). The obtained results include the fields: title, abstract, keywords, authors, type of document, year and URL.

C. Related Work

Karimzadehgan et al. [1] proposed an algorithm to solve the problem of committee review assignment by modeling the multi-aspect expertise matching as an integer linear programming problem which can can accommodate any probabilistic or deterministic method.

Dimou et al. [2] present RML, a generic mapping language, based on R2RML[6], that provides a uniform way used to map data present in any format to RDF[7]. Authors made use of RML to extract and map data of workshop proceedings published in HTML to an RDF model, that represents the research topics of the papers.

Atanassova et al. [3] present an Information Retrieval (IR) system for scientific publications. It provides the possibility of filtering results according to semantic facets. Semantic annotations are obtained using a rule-based method that identifies specific linguistic clues organized in a linguistic ontology.

Osborne et. al [8] proposed the Rexplore system, aimed to support the exploration and visualization of research trends. We use a similar ideas for data sources managing and publications enriching, however, we will dynamically add new data sources to improve author's information.

After having analyzed the related work of approaches that deal with identifying research topics, we can state that the existing works do not automatically enrich the research topics obtained by accessing third party research paper repositories, such as GoogleScholar or the IEEE repository. Furthermore, we propose the use data mining algorithms (clustering) differently from the aforementioned works.

III. Scenario

To depict an application scenario we introduce the case of the Ecuadorian Government, which states through Article 350 of the Country Constitution that: "The higher education system has the purpose of academic and professional training of scientific and humanistic vision, scientific and technological research, innovation, promotion, development and dissemination of knowledge and cultures, the construction of solutions to the problems of the country, in relation to the objectives of the arrangement of development" [5]. To follow these purposes, Ecuadorian universities have been investing important amounts of their resources and efforts in order to improve their infrastructure and human capital. Particularly, in the case of scientific and technical research. Initiatives like the "Rules of Selection and Award Programs and/or Projects of Scientific Research and Development Funded or Co-financed by the National Secretary of Higher Education, Science, Technology and Innovation (SENESCYT)", have been proposed to promote both: academic improvement and the creation of new knowledge through research. These norms regulate how the selection and adjudication process of programs and/or projects of scientific research and technological development are established, so that, public and private organizations can access to the funds managed by SENESCYT [4]. This new regulation system governs the Ecuadorian Higher Education System and leads the transcendental changes experienced in the country in at least the last 7 years. In this context, there are two elements involved in obtaining the best results pointed out by this challenge:

First, the responsibility of a researcher who is part of a University in Ecuador and that is working within a specific area, is to publish his/her results and findings. Unfortunately, this task has not been fully performed yet, mainly because traditionally it has not been considered by researchers as a high priority issue and only started being supported and funded in recent years. The problem is exacerbated by the lack of knowledge among people and the lack of tools used for this purpose. When a researcher identified in a given area of action needs to know about the progress of his/her topic at the local, national and international level, the common procedure evolves around literature review. However, we believe that contacting people who are involved in the same research areas of interest could highly benefit society.

Second, SENESCYT, as the government agency in charge of research management, may require for instance, a list of the researchers working on a specific area among all universities in the country to start developing a new research project related to the country's needs.

In order to solve these problems, we propose a searching process that retrieves potential research works in a specific area, specifically, applied to the digital repository of the University of Cuenca.

IV. Architecture

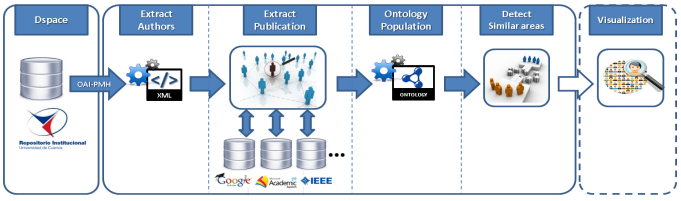

Fig. 1. System Architecture

Figure 1 shows the process of automating the identification of research areas. Our system consists of five main components: i) the authors extraction, which retrieves a list of dissertation authors, ii) publication extraction, which retrieves a list of publications belonging to the authors, iii) ontology population, which stores instances, iv) similar research areas, which makes use of data mining algorithms to detect similar areas, and v) visualization, which shows the result. Next, we briefly explain these components by illustrating the description with some examples.

[5]http://ieeexplore.ieee.org/gateway/
[6]http://www.w3.org/TR/r2rml/
[7]www.w3.org/RDF/

A. Authors Extraction

Authors data is normally located in the DSpace servers of the organization that holds them. Authors having dissertations have consequently registered publications in the institutional repository. However, these registers might not be necessarily up to date and could contain incomplete information. There are different ways to access DSpace. Some of them are:

- Database, access through connectors.
- OAI-PMH, access through a specific protocol.

After an exhaustive analysis, we opted for the access through the OAI-PMH protocol. Thus, the first step in our approach is to take the URL of a OAI-PMH service as an input and extract a list of authors from it. An example of OAI-PMH invocation is:

http://dspace.ucuenca.edu.ec/oai/request?verb=ListRecords &metadataPrefix=xoai

This service retrieves information related to authors. More specifically, it returns information about the following parameters: contributor, advisor, language, identifier, URI. The results obtained after the OAI-PMH are shown in the following listing:

Invocation Results Listing

```
<metadata>
    <element name="dc"> <element name="contributor"> <element name="advisor">
    <element name="es_ES">
    <field name="value">
        Saquicela Galarza, Victor Hugo
        ...
        <element name="author">
    <element name="es_ES">
    <field name="value">
    Haro Valle, Valeria Alexandra
    </field>
    <field name="value">
    Pérez Rocano, Wilson Rodrigo
    ...
    <element name="subject">
    <element name="es_ES">
    <field name="value">
    CENTRO DE DOCUMENTACION JUAN BAUTISTA
    VAZQUEZ
    </field>
    <field name="value">
    DATAWAREHOUSE
    </field>
    <field name="value">
    BIBLIOMINING
    </field>
    ...
    <element name="title">
    <element name="es_ES">
    <field name="value">
    Data warehouse para el Centro de
    Documentaci\'on Regional "Juan Bautista
    V\'azquez"
    </field>
    </element>
    </element>
    ...
</metadata>
```

The following authors taken from the results, are two representative examples of authors data, that will serve from now on, as an illustration for our findings:

- Víctor Saquicela. This author has a dissertation and three dissertations as director.
- Mauricio Espinoza. This author has a dissertation and nine dissertations as director.

B. Publication Extraction

Once a relevant author has been discovered, we can extract its main characteristics such as: publication, keywords, co-authors, etc. The list of authors is the one used to invoke the different services (mentioned in section II-B) in order to obtain another list containing the publications associated to the authors (if this list is not available, our system cannot continue without further human intervention). Then the system analyzes the response to obtain a basic description structure of the publications. This process is performed by the proposed algorithm 1:

Algorithm 1 Publication Extraction Algorithm

Require: author
  $publications \leftarrow null$;
  $publicationsScholar \leftarrow googleScholarSearch(author)$;
  $publicationsMicrosoft \leftarrow microsoftAcademicSearch(author)$;
  $publicationsIEEE \leftarrow IEEESearch(author)$;
  for all $publication \in publicationScholar$ **do**
    if $exist(publication)$ **then**
      // Publications matching, attributes updating and aggregation
      enrichPublications(publication)
    else
      $publications.add(publication)$
    end if
  end for
  for all $publication \in publicationMicrosoft$ **do**
    if $exist(publication)$ **then**
      // Publications matching, attributes updating and aggregation
      enrichPublications(publication)
    else
      $publications.add(publication)$
    end if
  end for
  for all $publication \in publicationIEEE$ **do**
    if $exist(publication)$ **then**
      // Publications matching, attributes updating and aggregation
      enrichPublications(publication)
    else
      $publications.add(publication)$
    end if
  end for
  return publications

The services invocation may or may not return a value. Here we show how we handle responses, which are represented in a structured manner that can be easily consumed by different technologies. The result of an invocation of our samples authors are shown in Figure 2.

Author	Title	Keywords
Victor Saquicela	Semantic Annotation of RESTful Services Using External Resources	[Domain Ontology, Semantic Annotation, Semantic Description, Semantic Web, Semantic Web Technology]
	Lightweight Semantic Annotation of Geospatial RESTful Services	[Domain Ontology, Semantic Annotation]
	...	...
Mauricio Espinoza	Querying the web: a multiontology disambiguation method	[Natural Language Processing, Search and Retrieval, Search Engine, semantic relatedness, Semantic Web]
	Discovering the Semantics of Keywords: An Ontology-based Approach	[Data Semantics, Emergent Semantics, Ontology Matching, Semantic Interoperability, Semantic Web, Web Pages]
	Discovering the Semantics of User Keywords	[Context Information, Digital Media, Formal Language, Information Need, Search Engine, Semantic Interoperability, Semantic Web, Web Search Engine, information re trieval]
	...	...

IEEE	Institute of Electrical and Electronics Engineers
MSA	Microsoft Academic Search
GS	Google Scholar

Fig. 2. Result of invocation

The process consists of using the above detailed APIs to obtain the publications related to the authors extracted in the previous step. To do this, we have created a model of objects that represents a scientific article to their respective authors, keywords and data source. Each data source is processed according to the heterogeneity of the data they contain.

C. Ontology Population

The result of the list of publications with their corresponding authors, is used to populate an ontology. We store this

result into a triple store using as reference the BIBO[8] ontology. We selected this storage tool in order to increase the discovery relations between authors, publication and sources. Moreover, the results established between different authors through both: keywords and the ontologies, are registered and stored in the repository, so that they can be used later. We use BIBO to describe publications and FOAF[9] to describe the authors. For URIs we have defined http://www.ucuenca.edu.ec/publication/ and http://www.ucuenca.edu.ec/author/. For conversion, we use the library JENA and the following properties of the ontology will be considered to map the results. Figure 3 depicts how the model presented in this approach is mapped to the BIBO ontology in order to represent author data. The result of this relation allows the generation and publication of dissertation data as linked data. Furthermore, it can be noted that we rely on triple store, which allows access via SPARQL.

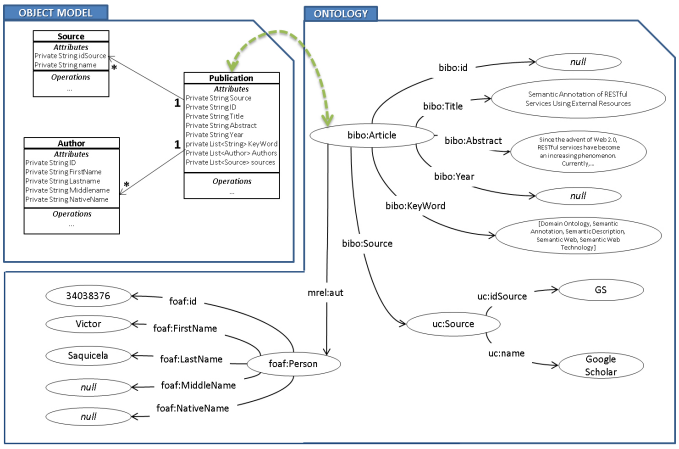

Fig. 3. Mapping between BIBO ontology and our model

D. Discovering Similar Areas

To automatically discover similarities, we describe how clustering algorithms can be used to discover similar research areas. For clustering algorithm execution, our system uses the WEKA[10] library. WEKA is a collection of machine learning algorithms used in data mining tasks. It contains tools that can be employed in: data pre-processing, classification, regression, clustering, association rules and visualization.

In the discovering process, the keywords of each author's publications are extracted, forming a kind of document containing just keywords associated to authors. Before running the algorithm, we pre-process, normalize and transform the data using different WEKA filters. In particular, clustering algorithms for documents. Clustering is the task of uncovering unanticipated trends by segmenting no predefined clusters. This approach is used in situations where a training set of pre-classified records is unavailable [6].

In this matter, we want to cluster keywords related to a similar area by looking at word weights. We use WEKA Simple-KMeans clustering. This algorithm is based on the Euclidean distance measurement to compute distances between instances and clusters. Data, once the algorithm is applied, will tend to cluster around certain keywords groups, allowing the user to quickly determine patterns in the data. The results of applying clustering algorithms associated with keywords publications, show that the authors found during the process described in section IV-A have indeed common research areas.

[8]http://bibliontology.com/bibo/

[9]http://xmlns.com/foaf/

[10]www.cs.waikato.ac.nz/ml/weka/

V. Conclusions and Future Works

In this paper, we have presented an approach for identifying common research areas. The goal of this study case is to analyze existing technologies used to search for publications, ontologies employed to represent publications, and data mining algorithms aimed to discover patterns based in keywords. Additionally, this approach includes the use of different tools that allow publications search. We describe the process we followed to demonstrate the potential of this proposal through an example.

Future work will focus on the addition of new search tools that could improve the obtained results. Also, we want to carry out the evaluation of the clustering results obtained at this stage of development. Furthermore, we pretend to show the results of linked data and clustering execution in a visual way. Finally, we plan to create a platform able to integrate data from other universities. Thus, we aim to discover similar areas of research between universities.

Acknowledgment

This work has been supported by the project " Plataforma de integración, publicación y consulta integrada de recursos bibliográficos en la Web Semántica", funded by CEDIA[11].

References

[1] Karimzadehgan Maryam and Zhai ChengXiang, "Constrained multi-aspect expertise matching for committee review assignment", In Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09), ACM, New York, NY, USA, pp. 1697-1700, 2009.

[2] Dimou A., Vander Sande M., Colpaert P., De Vocht L., Verborgh R., Mannen E and Van de Walle R., "Extraction and Semantic Annotation of Workshop Proceedings in HTML using RML", In Proceedings of The Semantic Publishing Challenge of the 11th Extended Semantic Web Conference, 2014.

[3] Atanassova Iana and Bertin Marc, "Faceted Semantic Search for Scientific Publications", 11th ESWC 2014 (ESWC2014), 2014.

[4] SENESCYT, "Reglamento de Selección y Adjudicación de Programas y/o Proyectos de Investigación Científica y Desarrollo Tecnológico (I+D) Financiados o Cofinanciados por la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación", Acuerdo 2012-009, 2012.

[5] Asamblea Constituyente, "Constitución de la República del Ecuador", 2008.

[6] Chen, Sherry Y and Liu Xiaohui, "The Contribution of Data Mining to Information Science", Journal of Information Science, no. 6, vol. 30, pp. 550-558, 2004.

[7] Tansley, Robert and Bass, Mick and Stuve, David and Branschofsky, Margret and Chudnov, Daniel and McClellan, Greg and Smith, MacKenzie, "The DSpace institutional digital repository system: current functionality", Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, IEEE Computer Society, pp. 87-97, 2003.

[8] Osborne, Francesco and Motta, Enrico, "Exploring Research Trends with Rexplore", D-Lib Magazine, Corporation for National Research Initiatives, vol. 19, no. 9, 2013.

[11]www.cedia.org