Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Automated Ontology Population and Enrichment of Scientific Publications

To cite this article: Maricela Bravo *et al* 2021 *J. Phys.: Conf. Ser.* **1828** 012139

Automated Ontology Population and Enrichment of Scientific Publications

Maricela Bravo[1*], Arantza Aldea[2], and Luis F. Hoyos-Reyes[1]

[1]Universidad Autónoma Metropolitana, Ciudad de México, México
[2]Oxford Brookes University, Oxford, UK

*Email: mcbc@azc.uam.mx

Abstract. Scientific publications are the most important resources available to the research communities. Researchers want their work to be widely recognized and available and also need powerful search engines to identify other publications and researchers working in the same area. Therefore, a good representation and organization of scientific products is crucial for an accurate retrieval of information. This paper describes an approach for automated population and semantic enrichment of an ontology model that represents scientific publications. Specifically, the type of enrichment used in this approach consists of implementing semantic similarity measurements between publications. Several experiments were performed to identify the best similarity measurement, using a statistical approach and the precision of the measurements.

1. Introduction

Universities and research institutions count with highly specialized researchers who are continuously producing data, information, papers, tools, etc. The scientific products and publications generated by researchers are an important resource available to the research communities. Acquiring, organizing, and processing scientific publications is a very important issue for institutions. To carry out the acquisition and representation of publications requires the implementation of novel approaches that facilitate the automated acquisition and management of bibliographic data.

The accuracy of the representation of research publications depends mainly on the quality of data sources and the information processed. Correct data sources and cross-validating publication data is required to acquire good and relevant useful bibliographic data. The validation of bibliographic data acquisition is a difficult task that must deal with: huge amounts of data that need to be stored and managed, heterogeneity in representation formats of the data, the incompleteness in the information retrieved, author disambiguation, among others.

Bibliographic information can be stored in a wide variety of formats, such as: plain text files, XML, relational data bases, among others. In particular, this paper reports an approach to automatically populate and enrich a scientific publication ontology model. The main motivation for this work is to investigate the advantages of using ontologies as a way to represent and manage large volumes of scientific publication data. Once the ontology is populated with bibliographic data, the ultimate goal is to enable the execution of logical inferences to discover possible scientific collaboration and common research topics.

Another important requirement of the scientific publication representation is the continuous actualization of data. Every day new research publications appear, that will need to automatically be captured and incorporated to the representation.

The contribution reported in this paper belongs to the area of ontology learning [1], in particular, the automatic population and enrichment of the ontology is presented. The enrichment method is based on the calculation of semantic similarity measures between publications. For experimentation, a set of semantic similarity measures are calculated, and the results are evaluated by means of a statistical analysis and the calculation of precision, determining which is the measure that offers the better results.

This paper is organized into the following sections: in Section 2, a revision of related approaches reported in literature is presented; Section 3 describes the layers that constitute the architecture of the solution; in Section 4, experimentation results are presented; and finally in Section 5 conclusions are presented.

2. Related Work

Representation and management of scientific publication has been attempted by using a variety of techniques.

Yao, Tang and Li in [2] describe a methodology to capture researchers' profiles by searching their homepages and annotating with labels the characteristics of their profiles. ArnetMiner is used to collect data from the web pages. Tokens such as position, affiliation, email address, etc are heuristically identified and tags are assigned to each token. Profiling extractions are then performed automatically. Liu et al. in [3] present a solution to address expert match based on an input requirement. The proposed system uses a collection of documents from CNKI as input data to KNN model. Two ontologies were developed as part of this prototype: an ontology that represents the core concepts of expertise, covering the concepts related with persons, publications, research projects and research interests; and an ontology to represent research areas and semantic relations for example: broader, narrower and part-of. Thiagarajan et al. in [4] aims to match a user profile with an area of expertise. Their software system extracts topics of expertise based on a set of documents provided by the participants. Authors represent user profiles using weighed topics to define the level of interest and implement various similarity measures to determine the level of user matching. A well-known system that aims to develop an academic social network system is ArnetMiner [5]. ArnetMiner extracts researcher profiles, integrates publication data from bibliographic digital libraries, generates models of the academic network, and provides search services for the academic network. In the work reported in [6] authors present a semantic web-based system with agents that generate researcher profiles and associations between them. The ontology model includes concepts related with the representation of profiles of researchers, conference publications, and research institutions, among others. The system proposed integrates the Dublin Core metadata, the FOAF, and the DBLP digital to execute SPARQL queries. Punnarut and Sriharee in [7] presented an ontology-based approach to identify the expertise and skills of a researcher to find the most suitable person to a task. This work reports a methodology by which a classification of skills is generated, which is implemented as an ontology, the goal is to identify the area of specialty of a researcher by using the ontology, and to find those researchers whose area of research and expertise are common.

Based on the review of related works, we can establish that the use of ontologies for the representation of research profiles is a solution that allows the generation and automated processing of information from scientific publications.

3. Architecture of the Solution

The system depicted in Figure 1 consists of two main layers: the scientific publication data extraction layer, and the scientific publication ontology management layer.

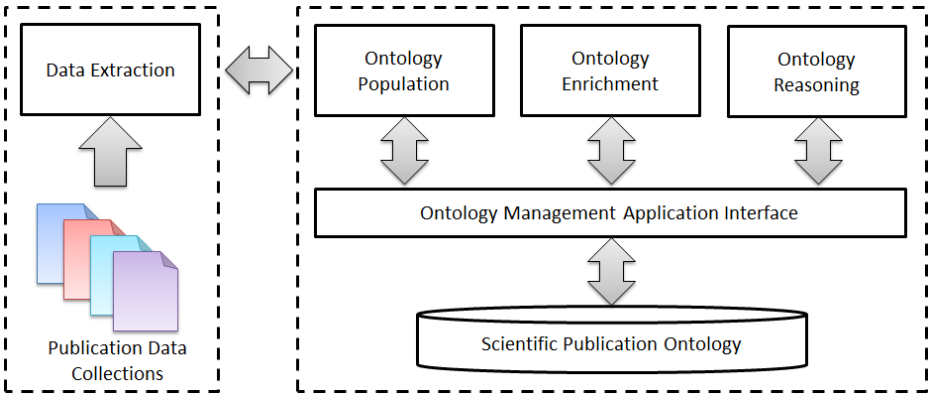

Figure 1. Architecture for the representation of scientific publication data.

3.1. Publication Data Source

Scientific publications data could be extracted from well-known bibliographic databases, such as: DBLP, CiteSeer, Google Scholar, Scopus, and ArnetMiner.

DBLP is an on-line computer science bibliography data source which provides free access to scientific publication data and references to on-line editions of publications. DBLP records more than 4 million of indexed publications.

CiteSeer is a digital library that provides a search engine for scientific publications of computer and information science. CiteSeer archives and indexes documents from publicly available websites.

Google Scholar is a search engine that indexes complete documents or metadata of publications from multiple scientific disciplines in various formats. Google Scholar facilitates access to abstracts of publications that have cited the article. Google Scholar provides the indexing of the citations. Google Scholar does not provide an API, and they block crawling attempts.

Microsoft Academic Search provides a search engine that indexes 220,607,278 papers, 240,779,188 authors, and 48,753 journals.

Scopus is a database of publications provided by Elsevier, its records citations and bibliographic abstracts with linked data about papers published in refereed journals, proceedings of conferences, and scientific book chapters. Scopus counts with a RESTful API that allows to obtain publication data. This service is not free of charge; only for institutions which count with a Scopus license.

ArnetMiner is an infrastructure that allows searching and data mining of scientific publications published on the Internet. ArnetMiner implements methods that exploit the structure of social network to discover academic and collaborative relationships. Using the ArnetMiner Infrastructure, you can perform expert searches, advisor recommendation, association discovery, course search, academic performance evaluation, and topic modeling.

After a careful consideration of the different data sources available, ArnetMiner was selected as the source for the data extraction due to the structured text files that the tool creates with the information about authors, publications, and citations. In particular, AMiner-Author.zip and AMiner-Paper.rar files were used as a case study for the protype described in this paper.

3.2. Scientific Publication Ontology Model

To manage the scientific publication an ontology model must be developed (see Figure 2) This ontology model aims at representing the most important elements of a publication such as: author, the publications itself and its research topics. The ontology model represents at this stage only the terminological box (T-Box) and will be populated in a later stage.

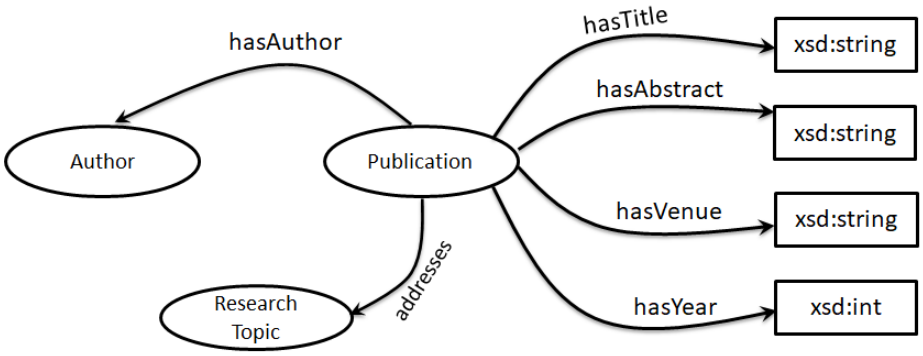

Figure 2. Scientific publication ontology model.

The Ontology population module shown in Figure 1 parsers the data extracted from the bibliographic data sources and stores the information into the ontology model. Figure 3 shows the class diagram designed for this purpose.

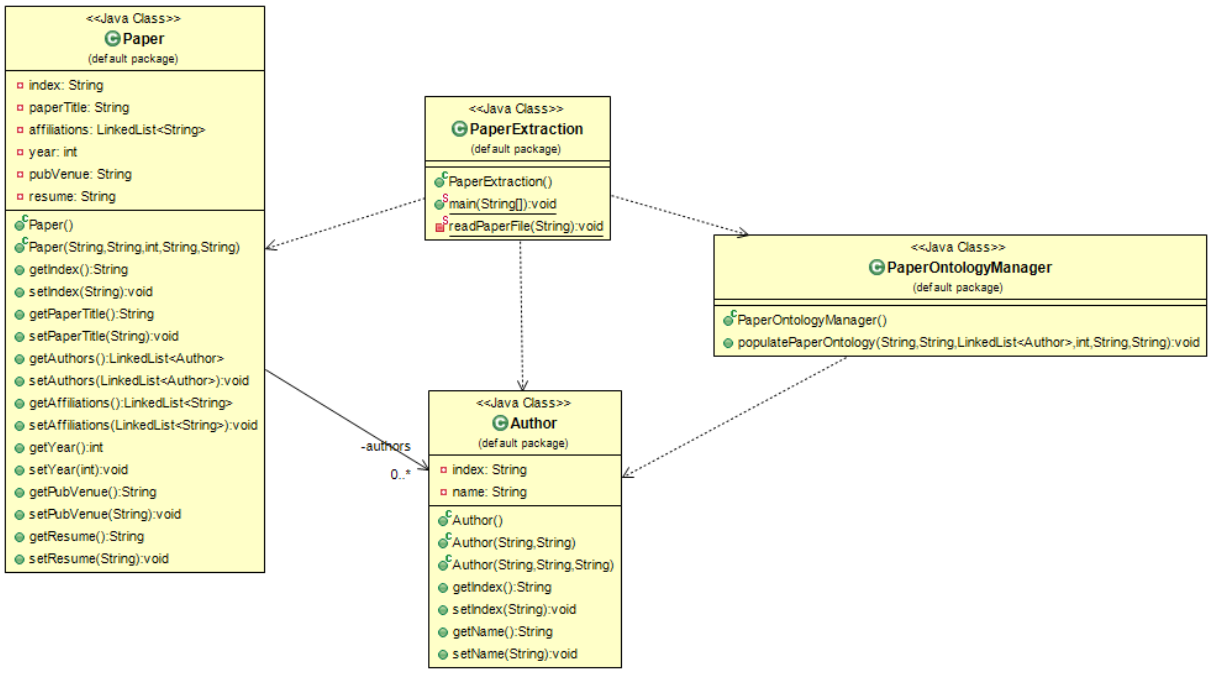

Figure 3. Design of the ontology population module.

3.3. Ontology Enrichment

Ontological enrichment is the automatic process by which semantic relationships are discovered between the concepts and individuals in an ontology. In the case of scientific publication, the ontology enrichment process aims to discover common areas of possible collaborations between different research groups by analyzing their publications. There are two main approaches to enrich an ontology model: *Enrichment within the ontology* by reasoning using logic axions and inference engines. *Enrichment outside the ontology*, using machine learning algorithms (supervised and unsupervised), extracting the data from the ontology, analyze the data, and discover new entities that are fed back into the ontology.

The ontology enrichment approach reported in this paper is the external enrichment. In order to discover the scientific collaborations and similar research interests, the ACM Computer Science taxonomy was incorporated to our ontology model incorporating a method to semantically correlate publications with the ACM concepts were developed.

3.4. Semantic Similarity Measures

In order to establish a correlation between the title of the article and the classification of the ACM it is necessary to calculate similarities. To calculate semantic similarity between titles and ACM classifications six WordNet-based measures were evaluated.

- In [8] authors presented a similarity measure that is calculated as follows: given two concepts (c1 and c2) find the closest common superclass c3, as well as the root node, based on the path distances between c1 and c3, c2 and c3, c3 and the root, determine the average distance of the path to the root node.
- Resnik [9] proposed a semantic approach to measure similarity between two concepts, this approach uses concept information, calculated as the occurrence frequency. Resnik states that the similarity between a pair of concepts is the measure of information in common or shared.
- Jiang and Conrath [10] formulated an information content-based approach to determine the semantic distance between concepts; for this measure the information content of a concept is defined as the probability of occurrence of the concept. This measure considers the probability of occurrence of the concepts, as well as the closest superior concept.
- Leacock and Chodorow [11] defined a semantic measure that indicates how similar the meanings of two words w1 and w2 are; this measure considers the shortest path connecting the meanings and the maximum depth in the taxonomy in which the meanings occur.
- PATH [11] semantic measure is calculated by means of the node count (path). The calculation of similarity between two concepts is the inverse value of the number of nodes considering the shortest distance measured in contiguous nodes between sets of synonyms. The smallest distance occurs when both sets of synonyms are equal, in this case the distance returned is 1, the maximum possible value.
- In [12] Lin proposed to measure the similarity between two concepts A and B as the ratio of the amount of information needed to establish the common concepts of A and B and the information needed to describe concepts A and B completely.

4. Analysis of Similarity Measures

To conduct an assessment of similarity measurement approaches revised in section 3.4, two ontologies were developed: The Publication ontology that was created by using the ontology population algorithm over the ArnetMiner data source file, this publication ontology contains 2081 publications; and the ACM classification ontology which contains 2500 research topics.

Once the two ontologies were completed, the next step is to calculate these semantic similarities between titles and research topics. Given the size of the two ontologies, the number of similarity calculations is more than 30 million (2081 titles times 2500 research topics times 6 similarity measures). To make the computation of the similarities more efficient and produce better results, the most suitable similarity algorithms must be found.

In order to evaluate semantic similarity models, two approaches were utilized: a statistical analysis and the evaluation of their performance by using precision, recall and F1 measure. Figure 4 shows the process of generating 10 sample files with similarity calculations. The process starts by randomly selecting a sample of 100 ACM topics from the ontology, and 100 publication titles from the publication ontology. Then the six similarity measurements are calculated between all pairs of topics and titles. To filter representative results, the mean of all measurements is used to select those similarities that are higher than 0.4.

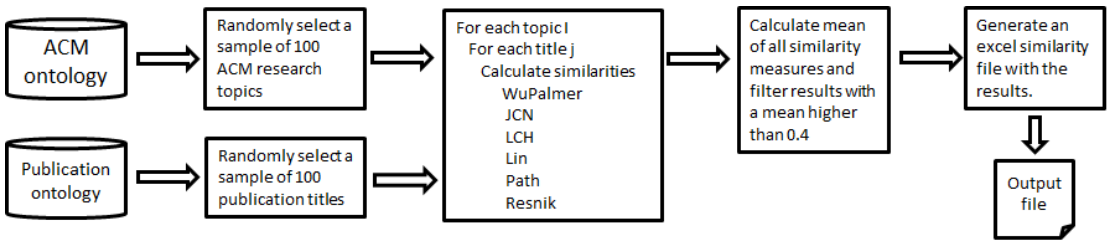

Figure 4. Process to randomly generate similarity sample files.

4.1. Exploratory Statistical Analysis of Similarity Measures

This statistical analysis aims to determine the stability of the similarity approach under a variance criterion (see Table 1).

Table 1. Exploratory Statistical analysis of similarity measures.

Similarities samples	Wu Palmer	JCN	LCH	Lin	Path	Resnik
Sample 1	0.51619336	0.17166925	0.86420904	0.25045354	0.21087887	0.65576238
Sample 2	0.52538175	0.17523732	0.82150527	0.26374271	0.2214339	0.72224
Sample 3	0.51461753	0.2018436	0.84657772	0.2750127	0.24099431	0.65534076
Sample 4	0.4525342	0.26407365	0.70802012	0.3126669	0.27430433	0.63055835
Sample 5	0.47479922	0.25382032	0.73659554	0.34159844	0.27176552	0.6755739
Sample 6	0.50657947	0.13513243	0.92798221	0.23316118	0.17964823	0.76025746
Sample 7	0.52282548	0.15522012	0.85554766	0.26629569	0.19318727	0.78180292
Sample 8	0.51194417	0.19981332	0.80169707	0.27086563	0.24399806	0.64674605
Sample 9	0.49068779	0.26529114	0.75884988	0.30640834	0.29640042	0.60478231
Sample 10	0.50109492	0.15704171	0.82687845	0.24158274	0.19112075	0.71912867
Mean	0.50166579	0.19791429	0.8147863	0.27617879	0.23237317	0.68521928
Variance	0.00053059	0.00229991	0.00432583	0.00116845	0.00157976	0.00336345
Standard Deviation	0.0230346	0.04795739	0.06577106	0.0341826	0.03974617	0.05799529
Interval	0.01603248	0.03337917	0.0457778	0.02379168	0.02766402	0.04036573
Relative error	**2.661513197**	14.04561751	**4.679005884**	7.174261546	9.914520739	4.905978605

The results summarized in Table 1 show that the e most stable similarity measure is the Wu Palmer similarity, followed by LCH.

4.2. Precision and Recall of Similarity Measures

Precision and recall calculations of each semantic similarity measure are shown in Table 2 Lin and Path measures show better results than the others. As the results of similarity calculations showed that there is a large number of dissimilarities between ACM concepts and titles, a balance between precision and recall is necessary (F1). F1 is the average of the values obtained by precision and recall. F1 penalizes classifiers with unbalanced precision and recall scores. Table 2 show that the measures with the best F1 scores are Lin and Path.

Table 2. Precision and Recall of similarity measures.

	Wu Palmer	JCN	LCH	Lin	Path	Resnik
Precision	0.1069	0.1500	0.0875	0.3000	0.3000	0.0916
Recall	0.5833	0.1500	0.9889	0.1833	0.1833	0.9445
F1 Measure	0.1807	0.1500	0.1608	**0.2276**	**0.2276**	0.1669

To determine which is the best measurement, a calculation is performed that combines both evaluation approaches. However, because the statistical analysis results are based on reducing the error, the resulting value that is selected is the smallest. While in the measurement precision evaluation approach, the returned value that is selected is the highest, since the higher the precision the measurement performs better. To obtain a harmonized average between both results, the following formula is used.

Overall evaluation = (0.6 * F1 measure) + (0.4 / Relative error)

Results shown in Table 3, correspond to the final evaluation of the two approaches, the Wu Palmer similarity measure has better values. As a conclusion, the Wu Palmer measure will be applied between all publications and ACM concepts to calculate similarity and determine if a semantic relatedness can be stablished into the ontology.

Table 3. Evaluation of the two approaches.

	Wu Palmer	JCN	LCH	Lin	Path	Resnik
Exploratory statistical	2.661513197	14.04561751	4.679005884	7.174261546	9.914520739	4.905978605
F1 Measure	0.1807	0.1500	0.1608	0.2276	0.2276	0.1669
Overall evaluation	0.258710444	0.118478634	0.181968245	0.192314867	0.176904865	0.181673173

5. Conclusions

This paper reports an approach to automatically populate and enrich an ontology of scientific publications. In particular, a set of semantic similarity measurements are evaluated to determine which is the most appropriate measurement that will offer the best results, considering two approaches: a statistical analysis and an analysis of the precision of the measurement. To reduce the number of calculations and computations required, a sampling was performed to determine a level of confidence about the use of the measurement.

References

[1] Maedche A, Staab S 2004 Ontology Learning. In: *Handbook on Ontologies*
[2] Yao L, Tang J, and Li J 2007 A unified approach to researcher profiling. *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (2007, November) IEEE Computer Society pp 359-366
[3] Liu P, Liu K, and Liu J 2007 Ontology-based expertise matching system within academia. *In 2007 International Conference on Wireless Communications, Networking and Mobile Computing WiCom 2007* (2007, September) IEEE pp 5431-5434
[4] Thiagarajan R, Manjunath G, and Stumptner M 2008 *Finding Experts by Semantic Matching of User Profiles* (Doctoral dissertation, CEUR-WS)
[5] Tang J, Zhang J, Yao L, Li J, Zhang L, and Su Z 2008 Arnetminer: extrac-tion and mining of academic social networks. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008, August, ACM) pp 990-998
[6] Adnan S, Tahir A, Basharat A, and de Cesare S 2009 Semantic agent oriented architecture for researcher profiling and association (semora) *In IEEE/WIC/ACM International Joint Conferences on Web Intelligence & Intelligent Agent Technologies* (2009, September, IEEE) **3** 559-562
[7] Punnarut R and Sriharee G 2010 A researcher expertise search system using ontology-based data mining *In Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling* **110** pp 71-78 (Australian Computer Society, Inc)
[8] Wu Z and Martha P 1994 Verbs semantics and lexical selection *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (Association for Computational Linguistics) pp 133-138

[9] Philip R 1995 Using information content to evaluate semantic similarity *In Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Montreal, Canada) pp 448–453
[10] Jiang J J and Conrath D W 1997 Semantic similarity based on corpus statistics and lexical taxonomy *In Proceedings of International Conference on Research in Computational Linguistics*, Taiwan
[11] Leacock C and Chodorow M 1998 Combining local context and WordNet similarity for word sense identification *Wordnet: An Electronic Lexical Database* **49**(2) 265-283
[12] Lin D 1998 An information-theoretic definition of similarity *in Proceedings of the 15th International Conf. on Machine Learning* (Morgan Kaufmann, San Francisco, CA) pp 296– 304