Author–Topic Classification Based on Semantic Knowledge

Chapter · May 2019

DOI: 10.1007/978-3-030-21395-4_5

6 authors, including:

José Luis Segarra
University of Cuenca
7 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Xavier Sumba
University of Cuenca
6 PUBLICATIONS **7** CITATIONS
SEE PROFILE

José Ortíz-Segarra
University of Cuenca
24 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Ronald Gualán
University of Cuenca
8 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Author-Topic Classification Based on Semantic Knowledge

José Segarra(✉), Xavier Sumba, José Ortiz, Ronald Gualán, Mauricio Espinoza-Mejia, and Víctor Saquicela

Department of Computer Science, University of Cuenca, Cuenca, Ecuador
{jose.segarra,xavier.sumba93,jose.ortizv,ronald.gualan,mauricio.espinoza, victor.saquicela}@ucuenca.edu.ec

Abstract. We propose a novel unsupervised two-phased classification model leveraging from semantic web technologies for discovering common research fields between researchers based on information available from a bibliographic repository and external resources. The first phase performs coarse-grained classification by knowledge disciplines using as reference the disciplines defined in the UNESCO thesaurus. The second phase provides a fine-grained classification by means of a clustering approach combined with external resources. The methodology was applied to the REDI (Semantic Repository of Ecuadorian researchers) project, with remarkable results and thus proving a valuable tool to one of the main REDI's goals: discover Ecuadorian authors sharing research interests to foster collaborative research efforts.

Keywords: Author-topic classification · Knowledge base · Data mining · Semantic web · Linked data · Data integration · Query languages

1 Introduction

In today's WWW (World Wide Web), the massive amount of bibliographic resources available through a number of digital repositories hinders data discovery and causes that many publications to go unnoticed due to the lack of interpretability of their databases. To overcome this limitation and take advantage of all kind of text resources available on the web, the scientific community has devised text processing technologies specialized in analysis and identification of bibliographic resources content. NLP (Natural Language Processing) and clustering are two well known of such technologies. However, most technologies perform syntactic analysis only and ignore the semantic analysis. This incomplete approach lead to poor results unable to fulfill users' expectations. Semantic web technologies have the potential to fill this missing gap, by preserving the meaning of language elements and make them processable and understandable for people and machines. Likewise, following the aforementioned principle, semantic knowledge bases such as DBpedia have emerged to try and preserve complete

B. Villazón-Terrazas and Y. Hidalgo-Delgado (Eds.): KGSWC 2019, CCIS 1029, pp. 56–71, 2019.
https://doi.org/10.1007/978-3-030-21395-4_5

knowledge by means of structures aimed to maintain not only meanings but also relationships between elements.

In this paper, we focus on the author-topic classification problem, i.e. modeling authors and their respective research fields based on their publications, by means of semantic web technologies. These models are useful to support interactive and exploratory queries over bibliographic resources, including analysis of topic trends, finding authors who are most likely to write on a given topic, discovering potential collaborative groups, among others [9]. Our proposed approach consists of a two-phase classification model: the first phase tries to associate authors to classes obtained from a thesaurus using semantic metrics to assess matching quality; while the second phase leverages clustering techniques to identify research fields associated to authors using information extracted from knowledge bases. The results from this phases are used as inputs to classify authors of scientific publications within automatically generated research areas.

The remainder of the paper is organized as follows. Section 2 reports the related work. Section 3 describes the proposed methodology. Next, Sect. 4 discusses the results obtained after applying the proposed methodology inside the REDI project. Finally, Sect. 5 presents the conclusions and future work.

2 Related Work

To the best of our knowledge, no unsupervised methods for author-topic classification exploiting semantic knowledge have been found. On the other hand, there is plenty of research about document classification, which might serve as a previous step to author classification systems [6]. Text classification methods have become very popular nowadays because of the increasing amount of documents published in digital format and the need to properly exploit them. For this reason, these techniques are very popular in tasks such as text mining, knowledge recovery, information retrieval, among others. There is an extense variety of text classification models; most of them mainly belonging to clustering models and machine learning algorithms, such as SVM, Naive Bayes, k-nearest neighbor, Neural Networks, Boosting strategies, etc. [8]. Leveraging knowledge bases has also been considered as an alternative to the traditional text classification models, and is mainly used to strengthen the process of document classification and clustering. In [12], for example, ontological models are used to improve the distance calculation of fuzzy classification techniques. Other proposals such as [1,2] harnessed popular knowledge bases such as WordNet[1] as the basis to identify the structure of sentences (e.g. nouns, verbs, adjectives) and extend their meaning during the classification process. In [5], the authors also use Wordnet plus a domain ontology to demonstrate how these knowledge bases help to overcome gaps associated with the syntactic representation of words and obtain better results in the task of documents clustering.

[1] https://wordnet.princeton.edu/.

In recent years, Wikipedia[2] has also supported a number of proposals related to text processing, information enrichment, and semantic classification as can be seen in [3,7,10]. This has been possible thanks to the huge amount of information available in Wikipedia and thanks to its growing community support. Thus, recognizing the relevance of Wikipedia in the semantic web domain, the project DBpedia emerged[3] as a semantic knowledge base harnessing most of Wikipedia's information but with an emphasis on using appropriate representation structures designed to facilitate querying and processing by both people and machines. Since its appearance, DBpedia has supported a considerable number of proposals for a variety of applications particularly in the information retrieval field. For instance [4] leverages DBpedia and page-rank techniques to semantically enrich meaning and structure (associated nodes) of data, to offer document categorization methods. Most of the proposals in the same line than [4], focus on document classification, mainly using categorization and clustering techniques according to the scope of the problem to be solved. However, most of those models present a great limitation in their practical application: they require large volumes of data for training. The need for pre-classified text is not trivial, especially when the classification problem does not consider predefined classes. For this reason, we propose an approach to author-topic classification based on the application of heuristics with the help of knowledge bases to offer a two-phased methodology. The proposed method was successfully used to classify authors based on their publications taking as use case the REDI project [11].

3 Author-Topic Classification

The approach presented in this section aims to classify an author to his corresponding research field, by means of a two-phase classification process. The proposed approach is depicted in Fig. 1. Two input parameters feed the process: the first one is a value that allows the unique identification of the author, while the second parameter is a set of keywords of publications in which he has worked. These data are common and easy to obtain for research scenarios, where all publications have associated keywords. For scenarios where there are not keywords, an alternative is to perform a pre-processing step of keyword extraction, which can usually be found in NLP frameworks.

As can be seen in Fig. 1, the result is a two-phased or two-level classification. The first phase is a classification by knowledge disciplines, while the second phase is more specific and includes a classification by research areas. These phases are further described next.

3.1 Phase 1: Classification Based on an External Taxonomy

The first phase of the proposed approach for author-topic classification begins with a general classification with respect to an external taxonomy such as the

[2] https://www.wikipedia.org/.
[3] https://wiki.dbpedia.org/.

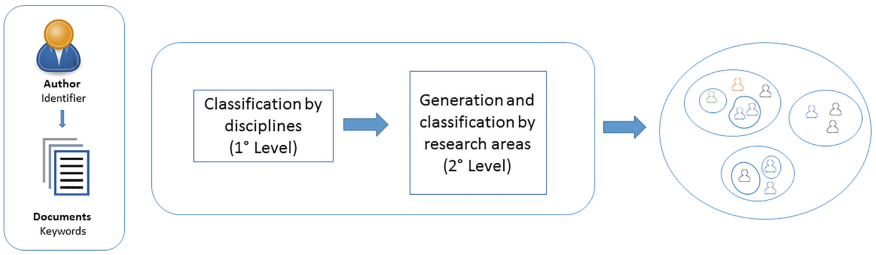

Fig. 1. Overall author-topic classification approach

UNESCO thesaurus[4]. Since the classification is done with respect to a controlled vocabulary (thesaurus), this first phase aims to reduce the number of possible groups or categories in a classification process. We have chosen the UNESCO thesaurus because it is worldwide known and is oriented to the classification of knowledge mainly related to research projects, thus covering several areas of knowledge. The UNESCO taxonomy contains a hierarchical three-level categorization:

- Fields: They refer to the most general sections and comprise several disciplines. They are encoded in two digits.
- Disciplines: They assume a general description of groups of specialties, and are encoded with four digits. Despite being different between themselves or disciplines with cross-references, it is assumed they have common features.
- Subdisciplines: These are the most specific entries in the nomenclature and represent the activities that are carried out within a discipline. They are encoded with six digits.

Initially, the classification process envisaged the use of a single-phase classification process, using an association with the subdisciplines of the UNESCO thesaurus. However, since the subdisciplines were outdated, many recent research areas might be left out. For this reason, it was decided that the UNESCO classification process would be the first classification phase. Additionally, instead of using the most specific subdisciplines, we decided to take as reference the disciplines (second level in the hierarchy), which are more general and remain valid for the intended use.

There are several ways to access the UNESCO thesaurus; however, in this work the SPARQL access point[5] was used. To classify the authors according to UNESCO's second level categories (disciplines), a relatively simple strategy has been chosen: comparing authors, represented by their keywords (from their publications), with each of the disciplines and subdisciplines of the UNESCO thesaurus. This is exemplified in Fig. 2. In this way, the disciplines with a higher

[4] http://skos.um.es/unescothes.
[5] http://skos.um.es/sparql/.

level of correspondence will be the one that best identifies the author's work. To perform this operation the authors are therefore represented as follows:

$$a_i = \{d_1, d_2, d_3, ..., d_m\}$$

Each author a_i is represented by a set of associated documents d_i about which he had participation. And, d_i are represented as a set of keywords s_i

$$d_i = \{s_1, s_2, s_3, ..., s_n\}$$

Finally, this implies that the authors a_i can be represented as combined set of the keywords from all their documents in the following way.

$$aS_i = \{s_1, s_2, s_3, ..., s_k\}$$

So that the results do not depend solely on the syntactic representation of keywords, the comparison is also made using semantic metrics (*SemSim*). This semantic metric is provided by the service of cortical.io[6]. Cortical.io, a company focused on *machine learning* and *big data*, has proposed a new data processing and representation methodology known as *Semantic Folding*. Within this representation, concepts are expressed by *semantic fingerprints* able to preserve multiple meanings and contexts and able to be used in several tasks including concept comparison. Cortical has started supporting multiple languages; however, in order to maintain homogeneity in the data and achieve better results, these are translated into English prior to comparison. The calculation of the score for each author with respect to a UNESCO's discipline is represented as follows:

$$ScoreAutorDisc(aS_i, Unesco_j) = \sum_{l=0}^{k} SemSim(s_l, Disc_j)$$

Where $Disc_j$ are the UNESCO's disciplines represented by the concept of the discipline j and the set of subdisciplines that underlies it.

$$Disc_j = \{disc, sub_1, sub_2, sub_3, ..., sub_p\}$$

Once each author has been compared to each of the UNESCO's disciplines and a score has been obtained, the disciplines are ordered such that the highest scored disciplines are associated with the author.

$$Clasf(aS_i) = \{(Unesco_j, ScoreAutorDisc(aS_i, Unesco_j)) \mid Unesco_j \in UNESCO\}$$

For example, according to the first phase of the classification approach, the author exemplified at Fig. 2 is correctly associated with the highest scored disciplines, namely computer science and computer technologies.

[6] https://www.cortical.io/.

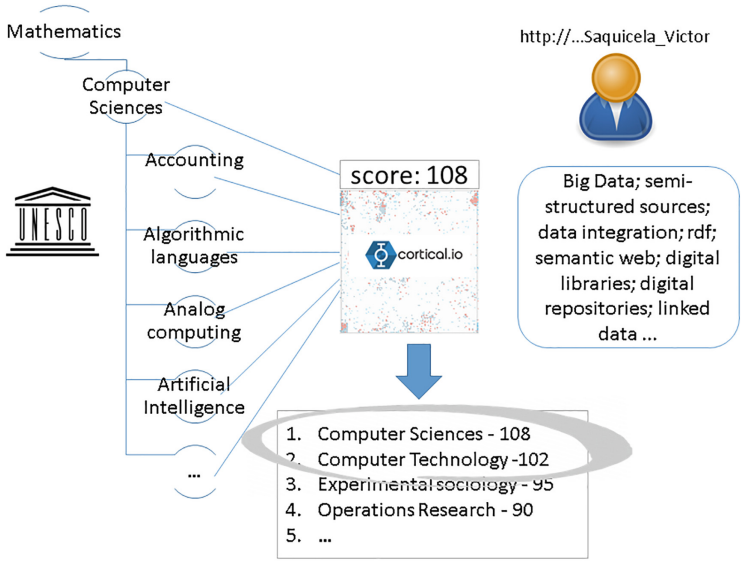

Fig. 2. Example of a semantic association of UNESCO's disciplines with an author

3.2 Phase 2: Classification to a Research Area

The second-level classification is meant to improve the first-level classification presented in the previous section, by providing a more specific approach. As can be seen in Fig. 3, this additional phase aims to group the authors according to their research areas based on their publications. To achieve this goal and given that the clusters (classification groups) are unknown, we intend to use the publications' keywords available to the authors, through a selection and filtering process to identify those that are suitable to be converted into valid classification groups.

The publications' keywords are quite suitable alternatives to be identified as research areas because they are generally relevant words placed by the author to try and reflect the scope of research covered by his work. However, this strategy was not considered as the only classification method because the set of keywords compiled by all the authors and even for the same author is extensive, which makes it difficult to identify valid groups in the face of a large number of possibilities.

To address this problem, firstly the clusters previously formed are taken into account to reduce the universe of keywords to analyze for generating second-level clusters. Taking the keywords from the first-phase clusters (previously identified classification groups) considerably increases the possibilities of finding common or closely related second-phase clusters. The second strategy is to identify the most relevant keywords within the incoming clusters to provide more representative research areas. For this, the most frequent keywords of each first-level

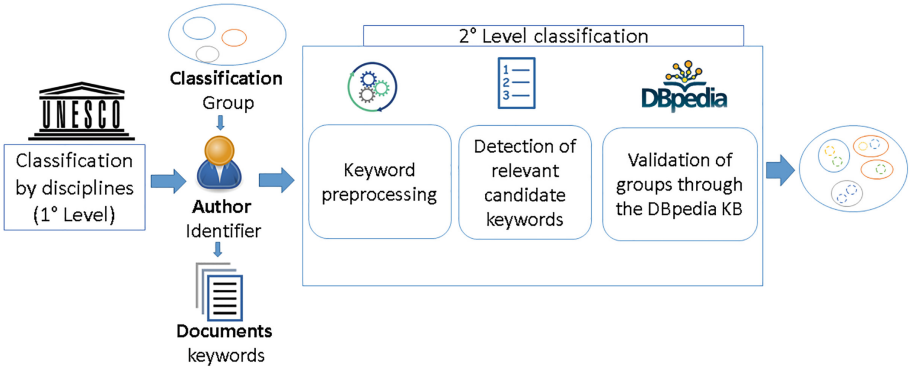

Fig. 3. A general scheme of the second phase classification

cluster are taken into consideration. This strategy alone provides good candidates to form new clusters; however, it is susceptible to the next problems:

- High-frequency keywords which do not reflect a research area: Sometimes high-frequency keywords reflect trivial data. For example, the location "Ecuador" or a word associated with the field "scientific article" may have high occurrence.
- Repeated words with different representation: It may be the case that keywords representing the same concept are repeated several times with different forms or languages. For example "Linked Data" with "Datos enlazados" or "Digital TV" with "Digital Television".

To address the aforementioned problems and select the most suitable keywords as valid research areas in the author classification, we propose using a semantic knowledge base for its validation. Specifically for this task, it is recommended to use DBpedia, which is suitable for the intended purpose because it contains a large amount of information of general nature. Additionally, DBpedia offers services such as *DBpedia Spotlight* to allow associating set of words with DBpedia concepts. Through the use of these tools, we intend to filter those keywords that are of interest for the classification, excluding those that represent very specific entities such as locations, people, among others. Furthermore, the structure containing the knowledge base can be used to refine keywords and detect those that are structurally close or represent the same concept with another representation. To accomplish the second-phase classification it is necessary to execute the following steps:

Data Extraction. The second phase classification begins by collecting keywords from the documents of all the authors listed in the first-phase clusters (UNESCO's disciplines). For example, all the authors belonging to the *computer science* cluster will be queried, and the corresponding keywords will be extracted. Thus obtaining a bag of words or set of words for this cluster.

$$BoW(Cluster_g) = \{aS_1, aS_2, aS_3, ..., aS_i\}$$

In this case, the classification process will not be done by author as in the previous phase, but by cluster. The results of this process will be assigned to the authors.

Keyword Preprocessing. The keywords extracted in the previous step are passed through a series of transformations aimed to correct some problems and help improve the results of this classification. The transformations applied are as follow:

- foreign characters removal (i.e. removing quotation marks, curly brackets, etc.),
- to-lowercase transformation,
- translation to English, and
- duplicated keyword removal (for each author).

Some of these preprocessing steps, such as translation and lowercase transformation, are mainly oriented to improve detection by the DBpedia Spotlight service.

Relevant Keyword Detection. To obtain the most relevant keywords from the words collected above, each one is counted and scored based on its frequency of appearance. From the most often words, the first 50 words are extracted and passed for the next step. If the number of words is less than 50, all of them pass. The defined number is arbitrary and was chosen to reduce the number of possible clusters that must be processed. The idea behind this step is that the keywords selected as research areas are the most relevant within each discipline; and therefore, include the largest number of authors.

Validation of Candidate Clusters Through a Knowledge Base. From the obtained set of keywords, a filtering and refining process is carried out to identify which keywords are suitable as research areas in the classification. To carry out this refinement process, the following tools are used: DBpedia Spotlight service for the detection of DBpedia entities, and DBpedia SPARQL endpoint to expand the information provided by the entities. The inputs feeding this step are the most relevant keywords. DBpedia Spotlight recognizes the input keywords and associates them with DBpedia entities that represent them. This service has the advantage of detecting common entities independent of their syntactic representation. Additionally, when there are multiple possible concepts for a given word, DBpedia Spotlight returns the closest one according to the context. For instance, if it finds *Apple* for computer science keywords, it will return the concept associated to the computer company's brand, instead that of the fruit.

Once the DBpedia concepts associated with the keywords are obtained, this link is used to carry out some additional validation and generalization processes.

Validation consists of recognizing only the keywords that can represent valid research areas. On the other hand, generalization aims to enforce that specific concepts are grouped together into a more general one and therefore the research area envelops the largest number of authors. For example, although concepts such as 'linked data' and 'semantic web' are different, through the structure of the knowledge base it can be discovered that they are close and that one of them encompasses the other. With this strategy, therefore, it is intended to prioritize the most general clusters, i.e. those that would contain the greatest number of elements. For the validation and generalization process the following strategies are followed:

Entity Filtering. Entities identified as persons or locations are ignored as candidate keywords. This is achieved through the type relationship (rdf:type) available to each entity.

Detection of Academic Type Relationships. In DBpedia there is a relationship between two entities known as *academicDiscipline*[7] that is used to identify an academic discipline or field of study and associate it with a scientific journal that contains it. By checking the existence of this relationship, it can be verified that the concept associated with the keyword analyzed can represent a valid area of research and thus obtain more congruent clusters. To detect this type of relationship, several strategies are performed as described below:

1. Direct verification: Checks if the entity has academicDiscipline relationships. If available, it will be marked as a valid research area.
2. Enrichment with parent categories: The parent entities are extracted from the current entity, and then it is checked if they have an academicDiscipline relation. If the entity has only one parent entity, the keyword is automatically identified as a research area.
3. Enrichment with sibling categories: sibling entities having the academicDiscipline relation are extracted from the current entity. If it has a single sibling entity, it is identified as research areas.
4. Direct classification: When there are no other possibilities, it is checked whether the concept is represented as a category independently whether or not it has an academicDiscipline relationship.

The strategies mentioned above are depicted in Fig. 4 and make the best effort to find the relationship of academic discipline both directly and through the knowledge structure. As a secondary result of this step, the possibility of finding common general concepts that encompass other concepts is also increased.

[7] http://dbpedia.org/ontology/academicDiscipline.

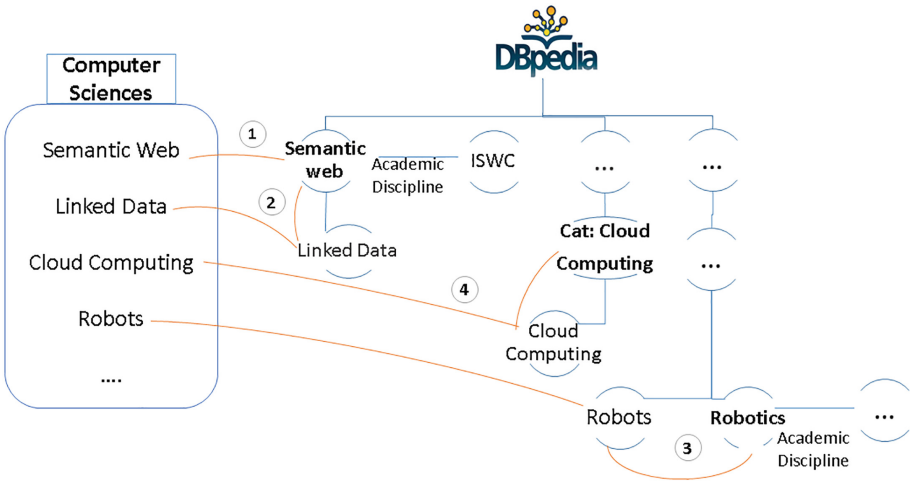

Fig. 4. Validation through the DBpedia structure.

If, after processing all the possibilities, an entity or set of entities that represent the processed keyword has not yet been found, a second pass is made where the following is done:

1. The entities that successfully passed the previous process are stored and identified as research areas.
2. Entities that have not been recognized are processed again, this time taking the parents and siblings of the previous process as a means of comparison. If they coincide with any research area previously obtained, this concept is linked to the matching areas.

An example of the process aforementioned is the cluster of *computer science* (See Fig. 5). In this case, it can be seen that although there is no direct relationship of the concept *mobile robots* with any research area, an indirect relationship through the structure of concepts (*skos:broader*) can be found. The previous proposal aims to maximize the possibility of relating concepts to common research areas among the authors, both directly and indirectly, using the knowledge base structure. The research areas finally obtained are associated with the authors of the keywords which resulted in those research areas. To achieve this, the history of changes applied to the author's keyword until it relates to a valid research area is stored. An example that represents this process is presented in Fig. 6.

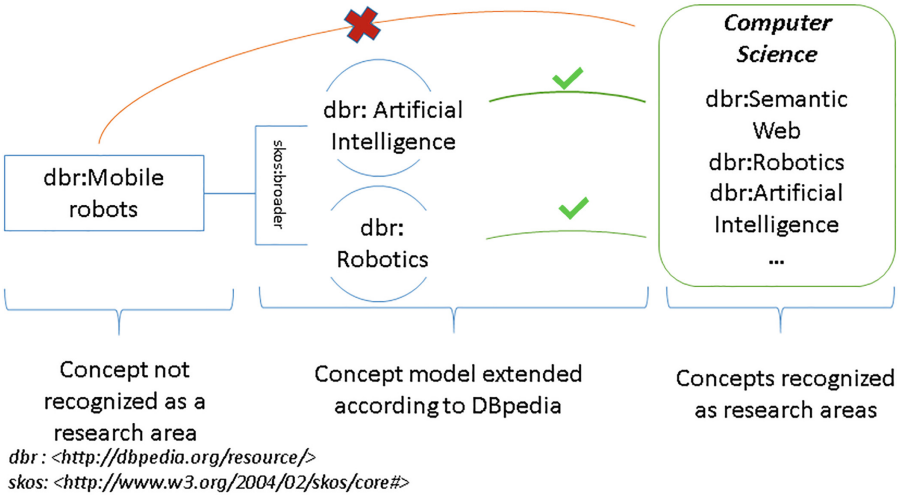

Fig. 5. Example of indirect association (Second pass)

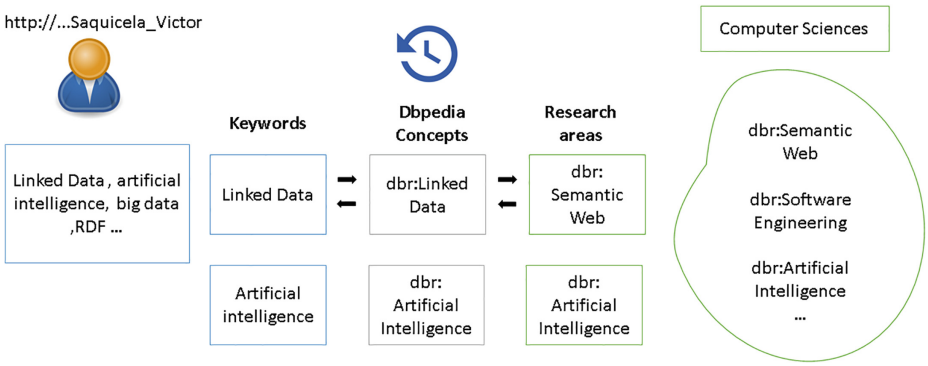

Fig. 6. Example of second-phase cluster assignment to authors

4 Discussion

The described method has been implemented within the REDI project (Semantic Repository of Ecuadorian Researchers), which compiles information related to Ecuador's scientific production. The information encompasses authors, publications, journals, etc. In this case, the main objective of the classification has been to recognize the most relevant research areas of the repository on which the researchers have focused their efforts. Through the execution of the process described in Sect. 3, up to two levels of classification have been achieved. The first based on UNESCO's disciplines, which was identified within the project as a knowledge area. The second-level classification was obtained from the same data with the support of DBpedia and has been recognized as the research area.

Most relevant resulting knowledge areas with their respective research areas are shown in Table 1.

Table 1. Main knowledge areas with their research areas obtained from REDI database

Knowledge areas	Research areas
Computer Sciences	Algorithm, Applied Mathematics, Artificial Intelligence, automation, Big data, Biomedical Engineering, Cloud Computing, Computer Network, Computer Simulation, Computer Vision, Control System, Control theory, Data mining, Data transmission, Decision Theory, Human-Computer Interaction, Image Processing, Information Technology, Machine learning, Mathematical Optimization, Pattern recognition, Risk Management, Robotics, Semantic web, Signal Processing, Social Science, Software Engineering, Systems Engineering, technology, Telecommunication, Theoretical Computer Science, World Wide Web
International Economics	Capitalism, Economic development, economic liberalism, economic policy, foreign direct investment, governance, human rights, international relations, microeconomics, monetary economics, public policy, social justice, social policy, sociology, unemployment
Policy Sciences	Education, Social science, Technology, Agriculture, Ecology, International development, Ethnology, Culture, Economic development, Health, Governance, Social policy, Public policy, Academic publishing, Capitalism, Politics, Environmental policy, Sociology, Youth, Globalization, Poverty, Environmental sociology, Cooperation

Based on the results shown in Table 1, it can be concluded that most of the classifications obtained are acceptable considering that no human intervention was necessary for the identification of the groups and their labeling. However, it can not be omitted that some results are not very intuitive and seem to be incorrect. Analyzing the latter highlights some Research areas not much related to the Knowledge areas. For instance: Biomedical Engineering, Social Science to Computer Sciences; Human Rights, Sociology to International Economics; Education, Technology, Agriculture to Policy Sciences. A further review of these cases reveals that they emerged due to very relevant authors on multidisciplinary works, whereby their keywords relate to several research areas. Another factor

Table 2. Example of authors belonging to the Computer Vision research area

Author	Main keywords
PONGUILLO INTRIAGO RONALD ALBERTO	Kalman filter, robotics, fpga, inertial navigation system, fuzzy logic, unmanned aerial vehicle, computer vision
CHILUIZA GARCIA KATHERINE MALENA	Learning analytics, multimodal learning analytics, human computer interaction, gamification, computer visión, user-centered design, educational data mining
CHANG TORTOLERO OSCAR GUILLERMO	Deep learning, artificial intelligence, image processing, artificial neural networks, machine visión, robotic visión, artificial vision

Table 3. Example of possible authors belonging to Computer Vision excluded by the algorithm

Author	Main keywords
RUEDA AYALA VICTOR PATRICIO	Fuzzy logic, image analysis, mapping, selectivity, site-specific harrowing, machine learning, remote sensing
OCHOA DONOSO DANIEL ERICK	Segmentation, feature extraction, recognition, hyperspectral imaging, image analysis, gene expression, fuzzy logic, unmanned aerial vehicle, social media, tracking
BENALCAZAR PALACIOS MARCO	Pattern recognition, machine learning, hand gesture recognition, image processing, mathematical morphology, neural networks

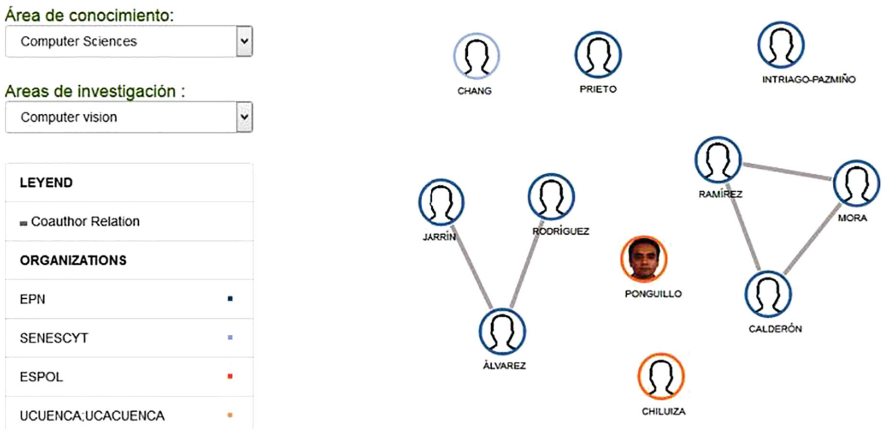

Fig. 7. Example of authors clustering

producing this kind of miss-classification is the lack of precision in the classification of first level authors because the semantic comparison algorithm is still in development and is not mature enough to cover all the topics with enough accuracy. This could be improved by placing minimum requirements on both the quality of the keywords representing an author during pre-processing, and the application of a threshold for the first-level classification process, which filters out ambiguous or low-trust authors.

Also, some sub-disciplines have been found too general and therefore can be placed at the same level of disciplines. Examples of such cases are: Mathematics, Energy, Electronics, Physics, Theory within Computer Sciences; Politics, Sociology within Policy Sciences. This happens because there is no model able to inform and limit the level of granularity in the process of grouping and labeling at the second-level phase (sub-discipline). This could be avoided by mapping the results of the sub-disciplines with respect to the UNESCO model or another standard model which provides a reference for the granularity. These problems will be further analyzed in a future work.

Regarding the conformation of the clusters, it has been noted that most of the authors belonging to these clusters do contain keywords associated with the tag of the generated cluster. This indicates that there is a high probability of membership between authors and the inferred sub-discipline. An example of some authors associated with the Knowledge area of *Computer Sciences* and the research area of *Computer Vision* are presented in the Table 2 keywords has been included as part of the author's name to provide context. On the other hand, it has also been noted that there is a chance the proposed method might leave authors out of their appropriate cluster due to author's keywords do not linked to cluster's relevant keywords. This can be due to multiple factors such as limitations on the knowledge bases or a high level gap between the keyword and a research area that does not allow knowing that they are related. An example of these cases can be seen in Table 3 which lists authors who were not clustered in the *Computer Vision* cluster. A more extensive and rigorous review of the quality of the groups obtained will be carried out in future work.

Finally, the result of clustering authors in the REDI web tool is presented. In this case, it can be seen how the authors are related with the knowledge area of *Computer science* and the *Computer Vision* research area (See Fig. 7). In addition, a complete list of clusters and authors can be found on the official website of the REDI project (https://redi.cedia.edu.ec/#/group/area).

5 Conclusions and Future Work

A two-phased author classification method is proposed. The approach defines 2 levels for generating clusters based on research areas around publications. The method showed remarkable results on the use case of the REDI project, contributing to the main objective of the project: discover authors sharing research interests. Although so far the proposed method has been tested in the scientific field only, with minor adjustments it has the potential to model author classifications in other applications, e.g. bibliography or institutional classifications.

Nevertheless, a notable limitation of the proposed method is that it can not guarantee successful classifications for all the cases. It can not be guaranteed that all authors belong at least to a second level group. As future work, we plan to evaluate the results against other proposals or using a gold standard which provides more clues about the quality of the obtained results. Additionally, it is planned to integrate state-of-art text classification methods and taxonomies to further improve the proposed strategy.

Acknowledgement. This manuscript was funded by the project "Repositorio Ecuatoriano de Investigadores" of the "Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia" (https://www.cedia.edu.ec/) (CEDIA, Spanish Acronym).

References

1. Bawakid, A., Oussalah, M.: A semantic-based text classification system. In: 2010 IEEE 9th International Conference on Cyberntic Intelligent Systems, pp. 1–6, September 2010
2. Celik, K., Güngör, T.: A comprehensive analysis of using semantic information in text categorization. In: 2013 IEEE INISTA, pp. 1–5, June 2013
3. Ciesielski, K., Borkowski, P., Kłopotek, M.A., Trojanowski, K., Wysocki, K.: Wikipedia-based document categorization. In: Bouvry, P., Kłopotek, M.A., Leprévost, F., Marciniak, M., Mykowiecka, A., Rybiński, H. (eds.) SIIS 2011. LNCS, vol. 7053, pp. 265–278. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-25261-7_21
4. Dostal, M., Nykl, M., Ježek, K.: Exploration of document classification with linked data and pagerank. In: Zavoral, F., Jung, J., Badica, C. (eds.) Intelligent Distributed Computing VII, pp. 37–43. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-01571-2_6
5. Hotho, A., Staab, S., Stumme, G.: Ontologies improve text document clustering. In: Third IEEE International Conference on Data Mining, pp. 541–544, November 2003
6. Korde, V.: Text classification and classifiers: a survey. Int. J. Artif. Intell. Appl. **3**, 85–99 (2012)
7. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links (2008)
8. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**(1), 1–47 (2002)
9. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 306–315. ACM (2004)
10. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using Wikipedia. In: Proceedings of the National Conference on Artificial Intelligence, vol. 2, 01 2006