Metadata Enrichment of Multi-disciplinary Digital Library: A Semantic-Based Approach

Hussein T. Al-Natsheh[1,2,3(✉)], Lucie Martinet[2,4], Fabrice Muhlenbach[1,5], Fabien Rico[1,6], and Djamel Abdelkader Zighed[1,2]

[1] Université de Lyon, Lyon, France
[2] Lyon 2, ERIC EA 3083, 5 Avenue Pierre Mendès France, 69676 Bron Cedex, France
[3] CNRS, MSH-LSE USR 2005, 14 Avenue Berthelot, 69363 Lyon Cedex 07, France
hussein.al-natsheh@cnrs.fr
[4] CESI EXIA/LINEACT, 19 Avenue Guy de Collongue, 69130 Écully, France
[5] UJM-Saint-Etienne, CNRS, Lab. Hubert Curien UMR 5516, 42023 Saint Etienne, France
[6] Lyon 1, ERIC EA 3083, 5 Avenue Pierre Mendès France, 69676 Bron Cedex, France

Abstract. In the scientific digital libraries, some papers from different research communities can be described by community-dependent keywords even if they share a semantically similar topic. Articles that are not tagged with enough keyword variations are poorly indexed in any information retrieval system which limits potentially fruitful exchanges between scientific disciplines. In this paper, we introduce a novel experimentally designed pipeline for multi-label semantic-based tagging developed for open-access metadata digital libraries. The approach starts by learning from a standard scientific categorization and a sample of topic tagged articles to find semantically relevant articles and enrich its metadata accordingly. Our proposed pipeline aims to enable researchers reaching articles from various disciplines that tend to use different terminologies. It allows retrieving semantically relevant articles given a limited known variation of search terms. In addition to achieving an accuracy that is higher than an expanded query based method using a topic synonym set extracted from a semantic network, our experiments also show a higher computational scalability versus other comparable techniques. We created a new benchmark extracted from the open-access metadata of a scientific digital library and published it along with the experiment code to allow further research in the topic.

Keywords: Semantic tagging · Digital libraries · Topic modeling
Multi-label classification · Metadata enrichment

1 Introduction

The activity of researchers has been disrupted by ever greater access to online scientific libraries –in particular due to the presence of open access digital libraries.

E. Méndez et al. (Eds.): TPDL 2018, LNCS 11057, pp. 32–43, 2018.
https://doi.org/10.1007/978-3-030-00066-0_3

Typically when a researcher enters a query for finding interesting papers into the search engine of such a digital library it is done with a few keywords. The match between the keywords entered and those used to describe the relevant scientific documents in these digital libraries may be limited if the terms used are not the same. Every researcher belongs to a community with whom she or he shares common knowledge and vocabulary. However, when the latter wishes to extend the bibliographic exploration beyond her/his community in order to gather information that leads him/her to new knowledge, it is necessary to remove several scientific and technical obstacles like the size of digital libraries, the heterogeneity of data and the complexity of natural language.

Researchers working in a multi-disciplinary and cross-disciplinary context should have the ability of discovering related interesting articles regardless of the limited keyword variations they know. They are not expected to have a prior knowledge of all vocabulary sets used by all other related scientific disciplines. Most often, semantic networks [6] are a good answer to the problems of linguistic variations in non-thematic digital libraries by finding synonyms or common lexical fields. However, In the scientific research context, using general language semantic network might not be sufficient when it comes to very specific scientific and technical jargons. Such terms also have the challenge of usage evolution over time in which having an updated semantic network counting for new scientific terms would be very expensive to achieve. Another solution could be brought by the word embedding approach [11]. This technique makes it possible to find semantically similar terms. Nevertheless, this approach presents some problems. It is not obvious to determine the number of terms that must be taken into account to be considered semantically close to the initial term. In addition, this technique does not work well when it comes to a concept composed of several terms rather than a single one. Another strategy is to make a manual enrichment of the digital libraries with metadata in order to facilitate the access to the semantic content of the documents. Such metadata can be other keywords, tags, topic names but there is a lack of a standard taxonomy and they are penalized by the subjectivity of the people involved in this manual annotation process [1].

In this paper we present an approach combining two different semantic information sources: the first one is provided by the synonym set of a semantic network and the second one from the semantic representation of a vectorial projection of the research articles of the scientific digital library. The latter takes advantage of learning from already tagged articles to enrich the metadata of other similar articles with relevant predicted tags. Our experiments show that the average F1 measure is increased by 11% in comparison with a baseline approach that only utilizes semantic networks. The paper is organized as follows: the next section (Sect. 2) provides an overview of related work. In Sect. 3 we introduce our pipeline of multi-label semantic-based tagging followed by a detailed evaluation in Sects. 4 and 5. Finally, Sect. 6 concludes the paper and gives an outlook on future work.

2 State of the Art

According to the language, a concept can be described by a single term or by an expression composed of multiple words. Therefore the same concept may have different representations in different natural languages or even in the same language in the case of different disciplines. This causes an information retrieval challenge when the researcher does not know all the term variations of the scientific concept he is interested in. Enriching the metadata of articles with semantically relevant keywords facilitates the access of scientific articles regardless of the search term used in the search engine. Such semantically relevant terms could be extracted thanks to lexical databases (e.g., *WordNet* [12]) or knowledge bases (e.g., *BabelNet* [13], *DBpedia* [8], or *YAGO* [10]). Another solution is to use word embedding techniques [5] for finding semantically similar terminologies. Nevertheless, it is difficult in this approach to identify precisely the closeness of the terms in the projection and then if two terms have still close meanings.

When the set of terms is hierarchically organized, it composes a taxonomy. A *faceted* or *dynamic taxonomy* is a set of taxonomies, each one describing the domain of interest from a different point of view [16]. Recent research in this area has shown that it improves the interrogation of scientific digital libraries to find specific elements, e.g., for finding chemical substances in pharmaceutical digital libraries [18].

The use of *Latent Dirichlet Allocation* (LDA) [3] for assigning documents to topics is an interesting strategy in this problem and it has shown that it helps the search process in scientific digital libraries by integrating the semantics of topic-specific entities [14]. For prediction problems, the unsupervised approach of LDA has been adapted to a supervised one by adding an approximate maximum-likelihood procedure to the process [2]. Using LDA for topic tagging however has a fundamental challenge in mapping the user defined topics with the LDA's latent topics. We can find a few variations of LDA trying to solve this mapping challenge. For example, *Labeled LDA* technique [15] is kind of a supervised version of LDA that utilize the user define topic. Semi-supervised LDA approaches are also interesting solutions for being able to discover new classes in unlabeled data in addition to assigning appropriate unlabeled data instances to existing categories. In particular, we can mention the use of weights of word distribution in *WWDLDA* [19], or an interval semi-supervised approach [4]. However, in the case of a real application to millions of documents, such as a digital library with collections of scientific articles covering many disciplines, over a large number of years, even recent evolutionary approaches of LDA require the use of computationally powerful systems, like the use of a computer cluster [9], which is a complex and costly solution.

3 Model Pipeline

The new model we propose can be resumed following a pipeline of 4 main components as illustrated in Fig. 1. In this section we will describe each of this components.

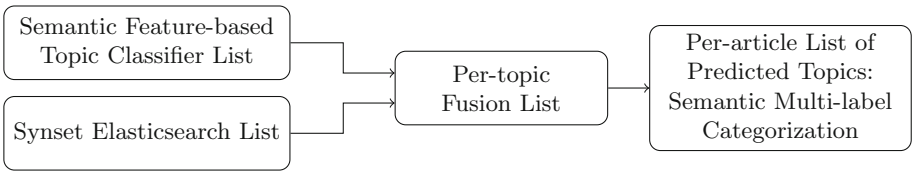


Fig. 1. High-level illustration of the model pipeline. The *Semantic Feature-based Topic Classifier* phase is used to generate *Top N* articles ranked by the probability of topic belonging. Another ranked list is generated by querying the synonym set (synset) of the topic using a text-based search engine which is presented in *Synset Elasticsearch* phase. A *Per-topic Fusion List* is then generated using a special mean rank approach in which only *Top* $a \times N$ are considered where a is experimentally determined. Finally, each article is tagged by a list of topics that was categorized with in the *Fusion list*.

3.1 Semantic Feature-Based Topic Classifier

This is computationally a big component that itself includes a pipeline of data transformation and a multi-label classification steps. The main phases of it are described as the following:

Extract Semantic Features. Starting from a multi-disciplinary scientific digital library with an open-access metadata, we extract a big number of articles, i.e., millions in which researchers want to explore. The retrieved data from the metadata of these articles are mainly the *title* and the *abstract*. These two fields will then be concatenated in order to be considered as the textual representation of the article in addition to a unique *identifier*. These set of articles will be denoted as *Corpus*. A TF–IDF weighted bag-of-word vectorization is then applied to transform the *Corpus* into a sparse vector space. This vectorized representation is then semantically transformed into a dense semantic feature vector space, typically 100–600 vector size. The result of this stage is an $(N \times M)$ matrix, where N is the semantic feature vector size and M is the number of articles. It must be accompanied with a dictionary that maps the article unique identifier of the article to the row index of the matrix.

Topic Classifier. For each topic name, i.e., scientific category name or a key-phrase of a scientific topic, we generate a *dataset* of *positive* and *negative* examples. The *positive* examples are obtained using a text-based search engine, e.g. *Elasticsearch*, which is a widely used search engine web service built on Apache Lucene, as the resulted articles that have *topic name* matches in *title* OR *abstract*. The negative examples, however, are randomly selected articles from the *Corpus* but with no matches with the *topic name* in any of the metadata text fields. Using this *dataset*, we build a kind of *One-vs-All* topic classifier. This classifier must have the ability of providing the predicted probability value of belonging to the topic, i.e. the class.

Probability-Based Multi-label Classification. Each of the obtained *One-vs-All* topic classifiers are then used in a multi-label classification task where each article in *Corpus* will have a probability value of belonging to the topic. This could be thought of as a kind of *fuzzy clustering* or *supervised topic modeling* where the article can be assigned to more than one topic but with a probability of belonging. The result of this stage is a top 100 K ranked list of articles per topic with the probability value as the ranking score.

3.2 Synset Elasticsearch

This component is computationally simple but has a great value in the pipeline. It is a kind of query expansion where the query space is increased by finding synonyms and supersets of query terms. So, it also requires a text-based search engine, e.g., *Elasticsearch*. We first need a semantic network or a lexicon database, e.g., WordNet, that can provide a set of synonyms of a giving concept name. For each topic in the set of topics, we generate a set of topic name synonyms, that is denoted by *Synset* (synonym set). Using *Elasticsearch* we then generate a ranked list of articles that have matches in their metadata with any of the synonyms in the topic *Synset*. So, the output of this component is a ranked list of articles per topic. As in Sect. 3.1, this output could be considered as a multi-label classification output but with ranking information rather than a probability score.

3.3 Fusion and Multi-label Categorization

This final stage constitutes the main contribution part of this experimentally designed pipeline. It uses an introduced ranked list fusion criteria of combining the 2 rankings of an article A which are the rank in the *Synset Elasticseach* list denoted by s_A and the rank in the semantic feature-based topic classifier list, denoted by r_A. If an article is present both in the 2 lists, we use a special version of *Mean Rank* score ($t_A = \frac{s_A + r_A}{2}$). Otherwise, the default score value of the article is given by equation ($t_A = r_A \times |S|$) where $|S|$ is the size of the *Synset Elasticseach* list.

The rank score of the *Fusion List* will be finally used to re-rank the articles to generate a new ranked list with a list size that ranges from the $max(|S|, |R|)$ and $|S| + |R|$ where $|R|$ is the size of the semantic feature-based topic classifier list. However, in our model we define a hyper-parameter a that determines the size of the *Fusion* list as in equation ($|F| = a \times |S|$). The hyper-parameter a will be experimentally determined based on multi-label classification statistics and evaluation that would be presented in Sect. 4.

The output of this component, and also the whole pipeline, is a list of articles with their predicted list of topics, i.e. scientific category names. Such list is obtained by applying a *lists inversion* process that takes as input all the per topic *Fusion* lists and generates a per article list of topics for all articles presented in any of the *Fusion* lists. The obtained list of predicted topics per article are optionally presented with a score value that reflects the ranking of the article in

the *Fusion* list of the topic. That score could be used to set an additional hyper-parameter replacing a which would be a score threshold that determines if the topic would be added to the set of predicted topic tags of the article. However, a simple and efficient version, as would be shown in Sect. 4, would only relay of the ranking information but having in place the design parameter a.

4 Experiments

4.1 Data Description

Scientific Paper Metadata from ISTEX Digital Library. The dataset used for running the experiments is extracted from *ISTEX*[1], a French open-access metadata scientific digital library [17]. This digital library is the result of the *Digital Republic Bill*, a law project of the French Republic discussed from 2014, one of whose aims is a "wider data and knowledge dissemination"[2].

ISTEX digital library contains 21 million documents from 21 scientific literature corpora in all disciplines, more than 9 thousands journals and 300 thousands ebooks published between 1473 and 2015 (in April 2018).

Private publishers (e.g., Wiley, Springer, Elsevier, Emerald...) did not leave access to their entire catalog of publications, that is why the publication access does not cover the most recent publications. In addition, because the contracts were signed with the French Ministry of Higher Education and Research, even if anybody can access to the general information about the publications with ISTEX platform (title, names of the authors and full references of the publication, and also metadata in MODS or JSON format), the global access is limited to the French universities, engineering schools, or public research centers: documents in full text (in PDF, TEI, or plain text format), XML metadata and other enrichments (e.g., bibliographical references in TEI format and other useful tools and criteria for automatic indexing).

For our experiments, we considered only a subpart of ISTEX corpus: the articles must be published during the last twenty years, written in English and related to sufficient metadata, including their title, abstract, keywords and subjects.

Scientific Topic from Web of Science. For each scientific article, we also use a list of tags extracted from the collection of *Web of Science*[3] which contains more than 250 flattened topics. These flattened topics are obtained as follows: when a topic is a sub-topic of another one, we can aggregate to the subcategory terms those of the parent category (e.g., [computer science, artificial intelligence] or [computer science, network]). Some of the topics are composition of topics, like "art and humanities."

[1] Excellence Initiative of Scientific and Technical Information https://www.istex.fr/.
[2] https://www.republique-numerique.fr/pages/in-english.
[3] https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html.

The selected 33 topics are: [Artificial Intelligence; Biomaterials; biophysics; Ceramics; Condensed Matter; Emergency Medicine; Immunology; Infectious Diseases; Information Systems; Literature; Mechanics; Microscopy; Mycology; Neuroimaging; Nursing; Oncology; Ophthalmology; Pathology; Pediatrics; Philosophy; Physiology; Psychiatry; Psychology; Rehabilitation; Religion; Respiratory System; Robotics; Sociology; Substance Abuse; Surgery; Thermodynamics; Toxicology; Transplantation].

In our experiments, to facilitate the analysis of the results without bias due to lexical pretreatment, we work only with topics containing neither punctuation nor linkage words. Moreover, we have kept in our experiences only *Web of Science* topics with enough articles (in ISTEX digital library) for having a significant positive subset of documents not used for the learning part (at least 100 scientific articles). The topics, which can be single words (as "thermodynamic") or a concatenation of words (as "artificial intelligence"), should be known in the semantic network to benefit of a consequent synonyms list. In our work, we present the results obtained with 33 topics, which are English single words or the concatenation of several words.

Synonym Sets from BabelNet. In our experiments, we produce a semantic enrichment by using a list of synonyms for each concept, also known as "synset" (for "synonym set"). To build our *synset* list, we need a semantic network. After some preliminary tests on several semantic networks, we chose *BalbelNet* [13] which gave better results. A sample synset from *BabelNet* for the topic *Mycology* is [Mycology, fungology, History of mycology, Micology, Mycological, Mycologists, Study of fungi].

Supervised LDA. Based on the state-of-the-art review as described in Sect. 2, we started by developing a model based on LDA. We defined a supervised version of the LDA (*sLDA*) where we the number of topics was set to 33 topics. Each topic was guided by boosting the terms of the topic synonym set obtained from *BabelNet* where the boosting values were [1, 10, 20, 30]. The dataset for experimenting this model were extracted from ISTEX scientific corpus by using *Elasticsearch* getting all articles that have at least one match of any of the 33 topics in any of these metadata fields: *Title*, *abstract*, *subjects* or *keywords*. However, the text used to build the *sLDA* were limited to the *title* and the *abstract*. The evaluation of the *sLDA* model will then be performed on a test set that is constructed from the *keywords* and the *subjects* fields.

4.2 Experimental Process

Initially, we defined an accuracy indicator that is based on the count of tagged articles with a list of prediction topics that has at least one label intersection with ground truth. This indicator will be denoted as *At least one common label*

metric. The other measure including label cardinality, Hamming loss and Jaccard index could be found in the literature[4].

In order to build an experiment of our proposed pipeline, we need to experimentally determine some hyper-parameters of it as follows:

Semantic Feature-Based Topic Classifier: We limit our text representation of the article to its title and abstract, which are available metadata. Comparing Paragraph vector [7] and Randomized truncated SVD [7] based on a metric that maximizes the inner cosine similarity of articles from the same topics and minimizes it for a randomly selected articles, we choose SVD decomposition of the TF–IDF weighted bag of words and bi-grams resulting in 150 features for more than 4 millions articles. As for the topic classifier, also by comparative evaluation, we select *Random Forest Classifier*, tuning certain design parameters, and use it to rank the scientific corpus. We consider the top 100 K articles of each topic classifier to be used in the fusion step.

Synonym Set Elasticsearch: Reviewing many available semantic networks, we found that BabelNet was the most comprehensive one combining many other networks [13]. So, we use it to extract a set of synonyms, i.e., a *synset* for each topic. This synset is then used to query the search engine of ISTEX which is built on Elasticsearch server. As would be shown in Sect. 5. This technique will be used as the experiment baseline.

Fusion and Per Multi-label Categorization: The main design parameter of this phase is the size of the ranked list that is achieved by setting it to the double size of the *Synset Elasticsearch* list.

5 Results and Discussion

First, we run an experiment on *sLDA* as described in Sect. 4. The result of this designed experiment was very disappointing based on the evaluation metrics. The best performing *sLDA* model, that was with a boosting value of 30, resulted in the following evaluation: *F1 measure* = 0.02828, *At-least-one-common-label* = 0.0443, *Jaccard index* = 0.0219 and *Hamming loss* = 0.0798. Comparing to using our pipeline with $a = 2$ having *F1 measure* of the 33 topics was 0.6032. So, *sLDA* was obviously not a good candidate to be used as a baseline. However, it was an additional motivation for designing and proposing our pipeline. After dropping *sLDA* from further experiments due to the very low evaluation results, we have added 2 more topics to the set of the 33 topics totaling to 35 topics. The 2 additional topics were [International Relations; Biodiversity Conservation]. We have also added more examples to the test set counting for an additional ISTEX metadata field called *categories:wos* that is actually does not exists in all the

[4] https://en.wikipedia.org/wiki/Multi-label_classification.

Table 1. Evaluation results based on the evaluation metrics *Recall* and *At least one common label* denoted here as the *Common-Match* metric. The table also shows the size of the intersection between the method results and the test set that was used in computing the evaluation metric, denoted here as *Intersection*. The value of *Intersection* might also be a good indicator of the method being able to tag more articles.

Method	Intersection	Common-Match	Recall
Synset	22,192	0.5284	0.5285
Fusion1	22,123	0.5736	0.5735
Fusion2	41,642	0.6375	0.6374
Fusion3	56,114	**0.6470**	**0.6473**
Fusion4	**67,625**	**0.6470**	0.6464

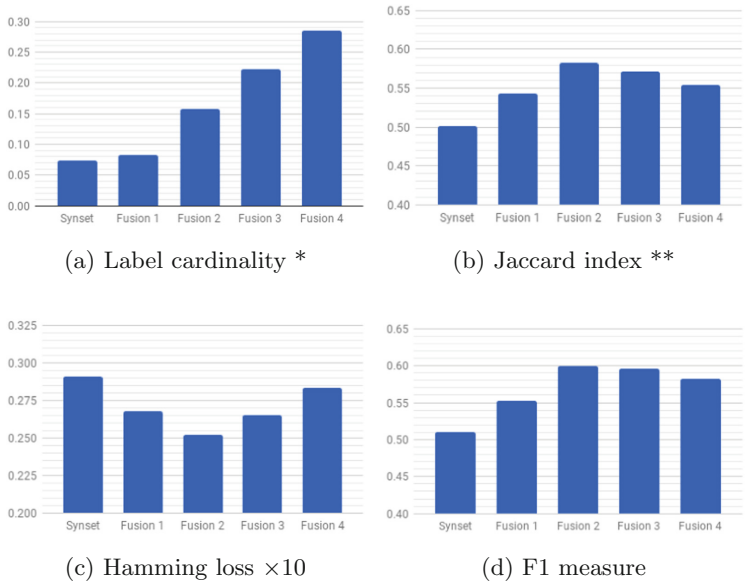
articles but was still considered as a good source for increasing the test examples in our published benchmark.

We define 5 methods for the experiment. One is a method of *Synset Elasticsearch*, denoted here by *Synset* which will be the baseline of benchmark. The other 4 methods are variations of our proposed pipeline but with variant values of the design parameter $a = [1, 2, 3, 4]$. The pipeline methods are then denoted respectively with the value of a as *Fusion1*, *Fusion2*, *Fusion3* and *Fusion4*. The results of the multi-label classification evaluation metrics, described in Sect. 4.2, are shown in Table 1 and Fig. 2.

While the evaluation metric values in Table 1 recommend higher a values, 3 or 4 with no significant value difference, we can see from Fig. 2 that the best value is $a = 2$ based on *Precision*, *F1 measure*, *Jaccard index* and *Hamming loss*. This means that if we increase the size of the fusion ranked list more than the double of the size of the Synset method, we will start loosing accuracy. Another indicator that we should limit the size of the Fusion list is Fig. 2a that shows that if we increase the size of the Fusion list, the difference of the *Label Cardinality* between the predicted results and the compared test set will increase. This difference is a negative effect that should be minimized, otherwise, the model will tend to predict too much labels that would be more probably irrelevant to the article.

Due to the fact that the test set was not generated manually but by filtering on a set of scientific category terms in relevant metadata fields, we believe that it is an incomplete ground truth. However, we think it is very suitable to compare models as a guidance for designing an efficient one because the test labels are correct even incomplete. Accordingly, we tried to perform some error analysis where we found that in most of the cases, the extra suggested category names are either actual correct topic having the article a multi-disciplinary one or topics from very similar and related topic. For example, a medical article from ISTEX[5] is tagged with the category name ['Transplantation'] in the test set. The predicted topics by our method was ['Mycology', 'Transplantation'] resulting into 0.5 precision value. However, when we read the abstract of that article, we find

[5] https://api.istex.fr/document/23A2BC6E23BE8DE9971290A5E869F1FA4A5E49E4.



(a) Label cardinality *

(b) Jaccard index **

(c) Hamming loss ×10

(d) F1 measure

Fig. 2. Results of *label cardinality difference*, *Jaccard index*, *Hamming loss* and *F1 measure* evaluation metrics. While Synset is the method that uses synonyms of the category name as a query in Elasticsearch, Fusion 1, 2, 3 and 4 represent respectively the values of the pipeline design parameters $a = [1, 2, 3, 4]$ that determine the number of annotated articles per topic as an integer multiple of the size of *Synset Elasticsearch* list. *: Difference value with the label cardinality of the compared test set of each of the methods. **: Equivalent to *Precision* in our case of a test set label cardinality = 1.

that it talks about *dematiaceous fungi* which is actually a 'Mycology' topic. So, in many cases where there is at least one common tag, the other tags are actually the aimed discovered knowledge rather than a false prediction. In another example, the model predicted the tags 'Psychology', 'Sociology' in addition to 'Religion' resulting in 0.3333 precision while they are actually relevant predicted tags when we read the abstract of the article[6] that also talks about *social networks*. The complete list of results –where these cases could be verified– are published as well as all the experimental data and reproducibility code[7].

6 Conclusion and Future Work

Governments, public organizations and even the private sector have recently invested in developing multi-disciplinary open-access scientific digital libraries. However, these huge scientific repositories are facing many information retrieval issues. Nevertheless, this opens opportunities for text-mining based solutions

[6] https://api.istex.fr/document/BA63065CCE8B0520F36B7DA90CF26F2DEF6CED7F.
[7] https://github.com/ERICUdL/stst.

that can automate cognitive efforts in data curation. In this paper, we proposed an efficient and practical pipeline that solves the challenge of the community-dependent tags and the issue caused by aggregating articles from heterogeneous scientific topic ontologies and category names used by different publishers. We believe that providing a solution for such a challenging issue would foster trans-disciplinary research and innovation by enhancing the corpus information retrieval systems. We demonstrated that combining two main semantic information sources – the semantic networks and the semantic features of the text of the article metadata – was a successful approach for semantic based multi-label categorization. Our proposed pipeline does not only enable for a better trans-disciplinary research but also supports the process of metadata semantic enrichment with relevant scientific categorization tags.

Other available methods in semantic multi-label categorization, such as LDA, are not suitable in this context for many reasons. For instance, they require powerful computational resources for processing big scientific corpus. Moreover, they need a pre-processing step to detect concepts that are composed of more than one word (e.g., "Artificial Intelligence"). Finally, LDA is originally an unsupervised machine learning model in which it is problematic to define some undetermined parameters like the number of topics. Our proposed pipeline, however, overcomes all of these limitations and provides efficient results. Towards improving the query expansion component of the pipeline (Synset Elasticsearch), we are planning to study the impact of using extra information from *BabelNet* semantic network other than only the synonym sets. In particular, we want to include the neighboring concept names as well as the category names of the concept. We expect that such term semantic expansion will improve the performance of the method.

Acknowledgment. We would like to thank ARC6 program (http://www.arc6-tic.rhonealpes.fr/larc-6/) of the Region Auvergne-Rhône-Alpes that funds the current PhD studies of the first author and thank ISTEX project.

References

1. Abrizah, A., Zainab, A.N., Kiran, K., Raj, R.G.: LIS journals scientific impact and subject categorization: a comparison between web of science and scopus. Scientometrics **94**(2), 721–740 (2013). https://doi.org/10.1007/s11192-012-0813-7
2. Blei, D.M., McAuliffe, J.D.: Supervised topic models. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (eds.) Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 3–6 December 2007, pp. 121–128. Curran Associates, Inc. (2007)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Bodrunova, S., Koltsov, S., Koltsova, O., Nikolenko, S., Shimorina, A.: Interval semi-supervised LDA: classifying needles in a haystack. In: Castro, F., Gelbukh, A., González, M. (eds.) MICAI 2013. LNCS (LNAI), vol. 8265, pp. 265–274. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-45114-0_21

5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. TACL **5**, 135–146 (2017)
6. Borgida, A., Sowa, J.F.: Principles of semantic networks - Explorations in the representation of knowledge. The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, Burlington (1991)
7. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011)
8. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134
9. Liang, F., Yang, Y., Bradley, J.: Large scale topic modeling: improvements to LDA on Apache Spark, September 2015. https://tinyurl.com/y7xfqnze
10. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: a knowledge base from multilingual wikipedias. In: CIDR 2015, Asilomar, CA, USA, 4–7 January 2015 (2015). www.cidrdb.org
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting held, 5–8 December 2013, Lake Tahoe, Nevada, United States, pp. 3111–3119 (2013)
12. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM (CACM) **38**(11), 39–41 (1995). https://doi.org/10.1145/219717.219748
13. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
14. Pinto, J.M.G., Balke, W.: Demystifying the semantics of relevant objects in scholarly collections: a probabilistic approach. In: Proceedings of the 15th ACM/IEEE-CE Joint Conference on Digital Libraries, Knoxville, TN, USA, 21–25 June 2015, pp. 157–164. ACM (2015). https://doi.org/10.1145/2756406.2756923
15. Ramage, D., Hall, D.L.W., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6–7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 248–256. ACL (2009)
16. Sacco, G.M., Tzitzikas, Y. (eds.): Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience. The Information Retrieval Series. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02359-0
17. Scientific and Technical Information Department - CNRS: White Paper - Open Science in a Digital Republic. OpenEdition Press, Marseille (2016). https://doi.org/10.4000/books.oep.1635
18. Wawrzinek, J., Balke, W.-T.: Semantic facettation in pharmaceutical collections using deep learning for active substance contextualization. In: Choemprayong, S., Crestani, F., Cunningham, S.J. (eds.) ICADL 2017. LNCS, vol. 10647, pp. 41–53. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70232-2_4
19. Zhou, P., Wei, J., Qin, Y.: A semi-supervised text clustering algorithm with word distribution weights. In: Proceedings of the 2013 the International Conference on Education Technology and Information System (ICETIS 2013). Advances in Intelligent Systems Research, 21–22 June 2013, Sanya, China, pp. 1024–1028. Atlantis Press (2013). https://doi.org/10.2991/icetis-13.2013.235