Using Semantic Relations between Keywords to Categorize Articles from Scientific Literature

Bastien Latard*†, Jonathan Weber*, Germain Forestier* and Michel Hassenforder*
*MIPS, University of Haute-Alsace, Mulhouse, France
†MDPI AG, Basel, Switzerland
Email: firstname.lastname@uha.fr

Abstract—**The amount of digital data is growing exponentially, and it is time consuming for researchers and readers to locate relevant information. Hence, being up-to-date in a specific research field (or topic) is a tedious and complex task. Our final goal is to create an intelligent scientific search engine by taking semantic relations into account. Our approach described in this paper is the starting point of such a smart system. Semantic relations between keywords are extracted from scientific articles in order to later help in the process of browsing and searching for content in a meaningful scientific way. By computing the most correlated categories and domains inherited from the keywords, we are able to extract the correct meaning of these keywords in relation to the article's concept. Our approach achieves a precision of 0.92 for both categories and domains extraction and a recall of 0.89 and 0.96, respectively.**

1. Introduction

Studying related work within scientific literature is a crucial step in order to have a global overview about a research topic. Indeed, the work from other researchers might save precious time in choosing the correct approach, or in keeping up-to-date with researchers. Nevertheless, finding a collection of related papers is still a tedious and complex task both in narrow and broad fields. Researchers spend a significant amount of time searching for work related to their scientific interests. This task is mandatory in any project to confront new ideas to existing solutions or to gain more knowledge about a scientific field. In this context, improving bibliographic search can have a great impact on scientific publishing.

An editorial from *Nature* [3] clearly expressed the continued frustration of the scientific community concerning the incredible potential that text mining of scientific literature represents. However, text miners often face the barrier of publishers' legal restrictions (i.e.: closed access). The average yearly growth of scientific literature is estimated to be 3 million new articles from journals and conferences over the last 4 years (http://www.scilit.net). Manually collecting and analysing this massive amount of data disseminated in 47,000 scientific journals is time consuming. Because of these de-centralized and isolated platforms, scientists have to rely on large databases which either provide an incomplete corpus or only display articles from their own platforms, often within very limited search functionalities. A solution to this limitation could be to make recommender systems smarter by integrating recent advances of text mining and semantic analysis. This has recently aroused considerable interest and attention (see Section 2).

In this paper, a new method is proposed to extract connections between categories and domains of scientific articles' keywords. This tackles the limitations of our naive approach inherited from the exact search highlighted in [8]. Indeed, semantically connecting keywords can be useful to validate the article field. Zhang et al. [24] suggested that using semantic relations between articles' keywords might have improved the precision and recall of their keyword extractor. After studying several NLP approaches, Soudani et al. [22] planned to integrate semantic similarities between words of a query in order to build their semantic information retrieval system. Hence, adding semantic relations is the first step towards our final research goal which is to make scientific search engines smarter. Indeed, depending on the number of results returned (and the matching categories/domains), a more refined/broader query could be proposed to the user.

2. Related Work

Due to the exponential growth of digital data, categorization, classification and more generally clustering or text mining have been widely studied in scientific literature. In this paper, we distinguish keywords (article's keywords) and key terms (main subjects of an article). Menaka and Radha [12] used Term Frequency Inverse Document Frequency (tf-idf) together with WordNet [13] as a knowledge base to extract key terms from scientific articles and then apply machine learning — *k-Nearest Neighbor (kNN), Decision Trees (DT), Naive Bayes (NB)* — to classify them. Nam and Quoc [14] combined Term Frequency (tf) with cluster-based approach in order to classify documents. These methods illustrate classical ways of classification (tf-idf, extraction of key terms) within scientific literature. Complete surveys about classification methods have been written in [4] and [2].
Analysis of scientific literature is not a simple task and has aroused particular attention over the last decades. A recent and comprehensive survey from 200 research articles about

recommender systems in scientific literature was performed by Beel et al. [1]. This highlights that approaches using citations data have been widely developed. Hamedani et al. [18]. applied an approach using citations to compute similarity between objects to scientific literature. Gonzalez-Pereira et al. [6] developed an approach to evaluate journals' prestige (SJR indicator). They concatenated citations with the prestige of citing journals, computed by a modified version of the PageRank algorithm. Citations can also be really general or even off-topic (e.g., philosophical citations in exact science). Moreover, this approach is not reliable in our context because only a portion of our data has citations information (see Section 4).

To counter the drawbacks of these co-occurrence recommendation systems, some approaches use collaborative filtering in diverse applications [9], [10]. The main goal of these recommender systems is to propose content in the same area as what the readers (or similar readers) read or like, by analysing users' interactions. This method is of particular interest because no text analysis is needed. Unfortunately, collaborative filtering approaches faced the cold-start problem and the motivation to participate is often very low.

Because of the limitations of both other approaches (collaborative filtering and co-occurrence), Content-Based Filtering is the most suitable approach in our context. This is also the most used in scientific recommender systems, according to Beel. Hence, using content to propose the most related articles or to compute similarity between articles seems to be a natural and the best choice. Jiang et al. [7] proposed a method that extracts problems/solutions information from abstracts, and which computes similarity models with td-idf and topic/concept models. Nasciemento et al. [15] implemented a real content-based approach which takes an article as the input and extracts related articles by querying three scientific libraries. Potential candidates are then ranked by computing the linear combination of cosine similarity matrices from abstracts and titles.

3. Materials and Methods

3.1. Proposed Approach

Our approach uses BabelNet [16] which is a multilingual lexicographic and encyclopaedic database based on the smart superposition of semantic lexicons (WordNet, VerbNet) together with other collaborative databases (Wikipedia and other Wiki data). BabelNet has already been widely used for text and data mining. Extraction of the most meaningful data from scientific articles [20], summarizing of documents [17] and word sense disambiguation (WSD) for multilingual document classification [19] are some of the tasks where BabelNet proved the suitability of its data.

This valuable added knowledge is used to search for all the keywords from Scilit (http://www.scilit.net), the scientific literature database developed by MDPI (http://www.mdpi.com). Even though Scilit contains metadata for more than 94 million articles as of today, our approach is evaluated with only a subset of 595 articles in order to be able to annotate and analyse the results. From their research work, Shah et al. [21] highlighted that this is legitimate to focus on the abstract to extract scientific articles' key terms (in biology), and concluded that abstracts contain the best ratio of key terms per total of words. Scilit keywords are either from the authors or generated by a topic extractor (MAUI [11]) from abstracts and titles. Thus, we assume that they are legitimate.

The concept of synset inherited from BabelNet is defined as a set of words sharing the same meaning in [16]. In other words, a synset (S) returned by BabelNet can be seen as a dictionary entry — *or a word within a specific concept* — from where one can obtain its corresponding categories (C), domains (D), synonyms (syn), and other interesting data. We define $S = \{C, D, syn\}$. Our function $F(K)$ returns, for each keyword K, its corresponding synsets ($\{S_1, \ldots, S_n\}$). There are 34 general domains (e.g., 'health and medicine' or 'physics and astronomy') and a lot of specific categories (e.g., 'peripheral nervous system disorders' or 'exact solutions in general relativity') mostly inherited from Wikipedia in BabelNet. In order to identify and select the correct dictionary entry regarding the context of the article, the intersections between the returned categories and/or domains are computed. $\mathcal{A}_K$ is the set of n keywords ($\{K_1, \ldots, K_n\}$) from an article $\mathcal{A}$. Our method aims to extract the connections between categories (and domains) of the different synsets resulting from all of the article's keywords. By doing this, we filter the noise coming from false friends from other disciplines. The best category $\mathcal{C}_{\mathcal{A}}$ of the article $\mathcal{A}$ is computed by the following equation:

$$\begin{aligned} C_A \in A_C \text{ such as } \forall\ c' \in A_C,\ count_{C_A} \geq count_{c'} \\ \text{where } count_c = \sum_{i=1}^{n} 1_{A_{S_i}}(c) \\ \text{and } 1_{A_{S_i}}(c) := \begin{cases} 1 \text{ if } c \in\ A_{S_i} \\ 0 \text{ if } c \notin\ A_{S_i} \end{cases} \end{aligned} \quad (1)$$

A_C and A_S are respectively all the categories and synsets of the article A. The same approach on domains provides more connections because they are fewer in number (34) and more general than categories, mostly inherited from Wikipedia. Figure 1 illustrates the main logic of our framework.

3.1.1. Exact Search. Exact search (ES) is the naive approach of our framework which takes keywords without any pre-formatting and tries to make an exact search on BabelNet. Then, connections from the different keywords' results are extracted, as described in Eq. 1. Figure 2 shows the limits of this approach. Indeed, no result is returned for composed keywords (keywords composed of several words) "flapping flight" and "normalized lift". This is problematic as a lot of keywords are actually composed. Furthermore, the more words a keyword contains, the less chance there is of obtaining an answer from BabelNet. In this example, "lift coefficient" only returns one synset, but there may be a lot of potential synsets for a single word. This approach provides a good precision (from 0.95 to 1 — Table 1) but covers only between 4% and 22% of the total data, depending on the value of the threshold parameter α (described in Section 4).

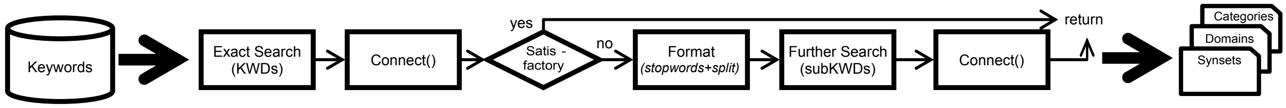

Figure 1: Illustration of the general logic of our approach

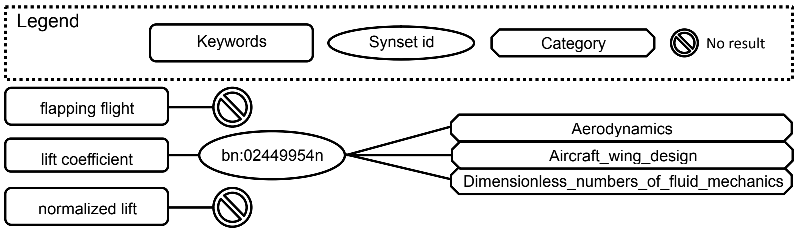

Figure 2: Limits of the exact search – one keyword from three returns data.

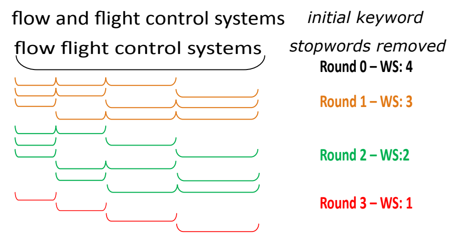

Figure 3: Split – WS: Window Size.

3.1.2. Further Search. In order to improve the recall for composed keywords and to cover more entries, stopwords are removed (because they are meaningless in category/domain detection), and keywords are exploded when there is no result from the exact search. This step is called further search in the rest of this article. By starting from the largest possible window (selection of sub-keywords), all of the linear combinations are tried. Window size (i.e. number of sub-keywords selected) is then reduced until data is found. Figure 3 illustrates the splitting logic. Within our dataset, this splitting logic provides better precision than subsequent searches (i.e. a sliding window) because it found more data. For a keyword "A B C", our framework searches for "A B", "B C", and also "A C". However, elements are not permuted ("C A", "C B", "B A"), as this includes more noise. Indeed, splitting on spaces provides more chances to obtain results, but also the risk to obtain synsets in the wrong context (the less words a keyword has, the less specific it becomes). To reduce this risk, only synsets sharing at least two categories/domains are returned when window size equals one, because they are less specific (and thus return more noise). The only exception is for two-word keywords. For those, even unconnected synsets are returned because they are almost as meaningful as the initial keyword. Synsets from a search containing several words are nevertheless returned even if there is no category/domain connection because they are precise enough. With this rule, we lose in precision (around 6%) but considerably gain in recall (around 11%). More results are given in Section 4.

3.1.3. Improvement. By looking at our initial example from the exact search approach, no category is returned for the composed keyword "flapping flight" (Figure 2). By splitting on spaces, our method is able to extract "Aerodynamics" as the main category of "flapping flight" from the connection between synsets from both words. The keyword "normalized lift" did not return any result from the exact search neither. However, the 43 synsets from "lift" are returned as the main ones for "normalized lift" because this is a two-word keyword. As a result, our approach is able to provide "Aerodynamics" as the main category of "lift coefficient; normalized lift; flapping flight". Figure 4 is a complete summary of our framework. Indeed, "lift coefficient" is a result from the exact search and a category is returned for "flapping flight" thanks to the further search. Moreover, all the 43 synsets from "lift" are kept as potential match for "normalized lift" (in further search) because of the two-word keywords rule. Finally, unrelated synsets from "lift" are naturally filtered out and the main category of the article is successfully extracted.

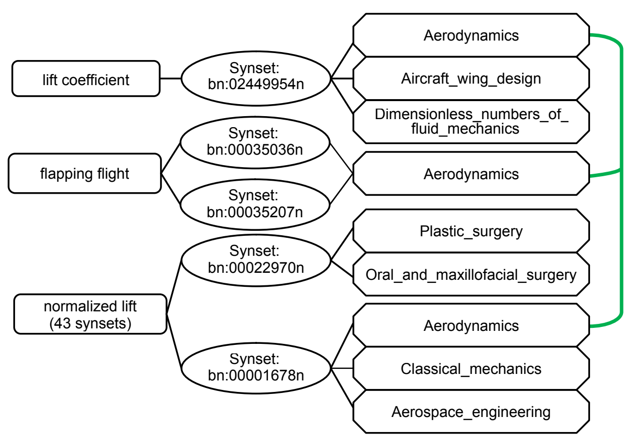

Figure 4: Further search successfully extracts the category shared by all of the three keywords.

3.1.4. Filter. As described in Eq. 1, our method filters noise by removing the unconnected synsets as soon as synsets from other keywords are connected (from their domains/categories). This returns the category "aerodynamics" for "flapping" and "flight" which respectively initially returned three and 25 synsets. However, it still returns "Living people; English-language films; Celestial mechanics; American films" as the main categories for keywords "nonlocal gravity; celestial mechanics; dark matter" after this filter. We identified constant noise (**_singer*, **_album*, etc.) which is meaningless in our scientific context, and set a parameter to force the automatic filtering of identified noise. Most of the remaining noise is finally naturally filtered out, and "Celestial mechanics" is returned as the main category.

Domains extraction provides more results because they are much more general than categories — *only 34 domains in BabelNet*. Thus, synsets without category connection are filtered out when there is a domain connection, and the other way around. This significantly decreases the number of false positives. Therefore, the precision has been improved, but the recall decreased because of our policy to never mark unconnected data as valid (see Section 4 for more details).

4. Results and Analyse

In order to build and evaluate our framework, a dataset of 595 articles from seven journals from two publishers was created in December 2016. All journals but one are in the same field (Physical Science). The exception (*Children*) validates that our approach can be applied in other fields (i.e., Pediatrics) by reaching the global average precision of 92% for categories. Finally, our dataset is made of all the articles from four journals (*Galaxies*, *Aerospace*, *Universe*, *Children*) from MDPI, all the articles from *Preprints* (http://www.preprints.org/) in the field of Physical Sciences and one hundred articles from two journals (*Classical and Quantum Gravity*, *The Astrophysical Journal*) from IOP Publishing (http://ioppublishing.org). Articles metadata, results from different methods, and correct/wrong data for each article (usable for automatic evaluation) are available at this link[1] . The downloadable zip contains detailed results (i.e.: by journal, by article and by mode).

Connecting categories/domains from the different synsets can be done more or less restrictively by changing the value of the minimum selection criteria (threshold), also called α parameter. Its value goes from 1 (more restrictive) to 4 (more flexible), as shown in the following list:

- α = 1: minimum of three keywords share the item
- α = 2: two keywords share only one item
- α = 3: two keywords share 1 $\rightarrow$ 3 categories and domain validated
- α = 4: minimum of two keywords share one, two or three items

Please note that 1. item can be categories or domains, 2. a lower α constraint is integrated into the higher values (e.g., 2 is effective in 2$\rightarrow$4)

In order to evaluate the correctness of our approach, the proposed categories/domains have been manually annotated as correct or incorrect. Precision (P) and Recall (R) are shown in their respective Table 1 and Table 2. $R0$ (Table 3) represents the recall for markable domains or categories. An article is markable when there is at least one connection among its keywords. Our approach does not propose any result for unconnected categories because we cannot define any confidence degree for those. Therefore, it makes sense to differentiate the general recall from all of the results returned and the markable (i.e., usable) ones. Table 4 shows a metric called coverage (C) which indicates the proportion of correct categories proposed in regards to all the articles in entry. We computed $F1$ (Table 5), a unique indicator taking P and R into account, defined as $F1 = 2 * (Precision * Recall)/(Precision + Recall)$. Its representation is interesting in our context to identify the α parameter which provides the best P/R combination. F1($R0$) (Table 6) is a variant using $R0$ instead of R.

Based on all of these tables, we can choose the α value depending on the metrics that matter the most. The best compromise in regards to all four metrics (P, R, C and $F1$) is obtained with $\alpha = 4$. Indeed, it provides a good precision (0.92), an acceptable recall for markable elements (0.89) and correctly covers 38% of the 595 articles. If precision matters more, α can be decreased, but recall and coverage also significantly decrease. If acceptable, it may be more interesting to slightly decrease the precision in order to significantly gain in recall and coverage. Throughout this paper, we focused on the results for categories in order to be consistent. Here are some statistics about the domains extraction. For domains, the best results are obtained with $\alpha = 3$, because precision decreases too much when $\alpha = 4$ (0.88). Globally, the same tendency was observed for domains extraction which also provides a good precision ($P = 0.93$), but with a much higher recall ($R0 = 0.96$, $R = 0.92$) and coverage ($C = 0.74$). The ratio precision / recall / coverage is better for domains than for categories extraction. This is not surprising as domains are more general and consequently more often overlap.

5. Discussion

By validating the main categories obtained from the keywords dictionary entries, the meaning of these keywords sharing the best category is also verified. Thus, using the words in relation with connected keywords (inherited from their synsets) is one of our ongoing contributions. However, an idea to improve our framework would be to extract the part-of-speech (POS) of composed keywords from a syntactical analyser or to lemmatize the keywords which do not provide any result. Then a more precise search on BabelNet might be executed.

Later, we could use some standard datasets in order to compare our results against results from other approaches, as Spanakis et al. [23] did in order to evaluate their metrics for semantic relatedness between words. Applying competitors' methods to our dataset would also be a good evaluation. Unfortunately there is, as far as we know, no comparable work in open source program as of today. The fact that we share the dataset used in this paper will hopefully help other researchers in the field to apply their approach to our data.

Nevertheless, the obtained results are a first step towards our main goal which is to integrate semantic relations into a scientific search engine. In a future work, we might use the output of our method to build a connected graph, which would represent a scientific wisdom tree. Articles would then be connected from their best categories and their corresponding BabelNet data. Hence, the similarity between articles keywords (and their synonyms within the correct context) inherited from the categories/domains connections might be the next step. Such graph inherited from BabelNet's synsets has already been developed by Franco-Salvador et al. [5], and seems to be an interesting idea to explore for our research. BabelNet data might also be

1. http://img.mdpi.org/data/latard_ictai2017.zip

TABLE 1: Categories – Precision.

Search	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exact	**1.00**	0.96	0.96	0.96
Further	0.88	**0.93**	0.92	0.92

TABLE 2: Categories – Recall.

Search	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exact	0.09	0.38	0.44	**0.47**
Further	0.17	0.55	0.72	**0.77**

TABLE 3: Categories – Recall ($R0$).

Search	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exact	0.18	0.76	0.89	**0.93**
Further	0.20	0.63	0.83	**0.89**

TABLE 4: Categories – Coverage.

Search	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exact	0.04	0.18	0.21	**0.22**
Further	0.08	0.27	0.35	**0.38**

TABLE 5: Categories – F1.

Search	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exact	0.16	0.54	0.61	**0.63**
Further	0.29	0.69	0.81	**0.84**

TABLE 6: Categories – F1 ($R0$).

Search	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exact	0.30	0.85	0.92	**0.94**
Further	0.32	0.75	0.87	**0.90**

used to create clusters of scientific articles, and dynamically assign the new papers to the correct clusters. This may be interesting for information retrieval, in order to propose related articles from the same topic.

6. Conclusion

Making scientific recommender systems smarter is crucial in order to help scientists in their mandatory and tedious bibliographical research phase. Our approach is the first step in making such a smart system. Indeed, the results (Section 4) confirm that using semantic relations between keywords provides a good way to classify scientific articles. It actually provides a good precision (around 0.92 for both categories and domains) and a good recall $R0$ (0.89 for categories, 0.96 for domains). Finally, correct categories are found for 38% of the articles, and 74% of articles obtained correct domains. Our work opens many perspectives for further research, some are presented in Section 5.

References

[1] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, November 2016.

[2] Michael W. Berry and Malu Castellanos. *Survey of text mining II*. Springer, 2008.

[3] Nature (Editorial). Gold in the text? *Nature*, 483(7388):124–124, March 2012.

[4] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[5] Marc Franco-Salvador, Fermín L. Cruz, José A. Troyano, and Paolo Rosso. Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowledge-Based Systems*, 86:46–56, September 2015.

[6] Borja González-Pereira, Vicente P. Guerrero-Bote, and Félix Moya-Anegón. A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3):379–391, July 2010.

[7] Yichen Jiang, Aixia Jia, Yansong Feng, and Dongyan Zhao. Recommending academic papers via users' reading purposes. In *RecSys*, pages 241–244. ACM, 2012.

[8] Bastien Latard, Jonathan Weber, Germain Forestier, and Michel Hassenforder. Towards a Semantic Search Engine for Scientific Articles. In *TPDL*. Springer, September 2017.

[9] Nikolaos F. Matsatsinis, Kleanthi Lakiotaki, and Pavlos Delias. A system based on multiple criteria analysis for scientific paper recommendation. In *PCI*, pages 135–149, 2007.

[10] Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In *CSCW*, pages 116–125. ACM, 2002.

[11] Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, 2009.

[12] S. Menaka and N. Radha. Text classification using keyword extraction technique. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(12), 2013.

[13] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[14] Le Nguyen Hoai Nam and Ho Bao Quoc. A Combined Approach for Filter Feature Selection in Document Classification. In *ICTAI*, pages 317–324. IEEE, November 2015.

[15] Cristiano Nascimento, Alberto H.F. Laender, Altigran S. da Silva, and Marcos André Gonçalves. A source independent framework for research paper recommendation. In *JCDL*, page 297. ACM Press, 2011.

[16] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, December 2012.

[17] Haniyeh Rashidghalam, Mina Taherkhani, and Fariborz Mahmoudi. Text summarization using concept graph and BabelNet knowledge base. In *AIR*, pages 115–119. IEEE, 2016.

[18] Masoud Reyhani Hamedani, Sang-Wook Kim, and Dong-Jin Kim. SimCC: A novel method to consider both content and citations for computing similarity of scientific papers. *Information Sciences*, 334-335:273–292, March 2016.

[19] Salvatore Romeo, Dino Ienco, and Andrea Tagarelli. Knowledge-based representation for transductive multilingual document classification. In *ECIR*, pages 92–103. Springer, 2015.

[20] Francesco Ronzano and Horacio Saggion. Knowledge extraction and modeling from scientific publications. In *SAVE-SD*, 2016.

[21] Parantu K. Shah, Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. Information extraction from full text scientific articles: Where are the keywords? *BMC bioinformatics*, 4(1):20, 2003.

[22] Nadia Soudani, Ibrahim Bounhas, and Yahya Slimani. Semantic Information Retrieval: A Comparative Experimental Study of NLP Tools and Language Resources for Arabic. In *ICTAI*, pages 879–887. IEEE, 2016.

[23] Gerasimos Spanakis, Georgios Siolas, and Andreas Stafylopatis. A Hybrid Web-Based Measure for Computing Semantic Relatedness Between Words. In *ICTAI*, pages 441–448. IEEE, November 2009.

[24] Chengzhi Zhang. Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3):1169–1180, 2008.