Towards a Semantic Search Engine for Scientific Articles

Bastien Latard[1,2(✉)], Jonathan Weber[1], Germain Forestier[1], and Michel Hassenforder[1]

[1] MIPS, University of Haute-Alsace, Mulhouse, France
bastien.latrad@uha.fr
[2] MDPI AG, Basel, Switzerland

Abstract. Because of the data deluge in scientific publication, finding relevant information is getting harder and harder for researchers and readers. Building an enhanced scientific search engine by taking semantic relations into account poses a great challenge. As a starting point, semantic relations between keywords from scientific articles could be extracted in order to classify articles. This might help later in the process of browsing and searching for content in a meaningful scientific way. Indeed, by connecting keywords, the context of the article can be extracted. This paper aims to provide ideas to build such a smart search engine and describes the initial contributions towards achieving such an ambitious goal.

1 Introduction

Keeping up-to-date in a specific research field is a tedious and complex task. This is mandatory as it allows researchers to increase their knowledge on a domain and acquire latest ideas. Hence, choosing the correct approach is the first step of any research work. Despite—*or because of*—the data deluge in scientific publication, researchers spend a significant amount of time searching for articles related to their scientific interests.

An editorial from *Nature* [1] 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 growth of scientific literature is estimated to be 3 million new articles per year from journals and conferences over the last 4 years, with 3.3 million articles produced in 2016 (http://www.scilit.net). This massive amount of data is published by more than 6000 publishers in around 47,000 scientific journals. These de-centralised and separated platforms further complicate the research process because scientists are unable to go through them all in order to search for relevant articles. Thus, they have to rely on big databases or indexing companies which provide either an incomplete corpus due to selection criteria or only display articles from their own platforms. Moreover, their search engines often offer very limited search functionalities, and this is the problem we want to tackle.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 608–611, 2017.
DOI: 10.1007/978-3-319-67008-9_54

To tackle this problem, our approach consists in using semantic relations between keywords to extract the main categories of the articles. This approach simultaneously validates both the context of the article and the context of the word, thus providing the correct category. Effendy and Yap [2] discussed the potential of using semantic mining tools to extract the best category of a conference. This is exactly what our framework aims to do.

2 Method

Our approach uses BabelNet [3] 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). A query for a term through BabelNet returns "dictionary entries", synonyms, categories or domains. Each synset S contains the relative categories C, domains D and synonyms syn within the specific concept:

$$S = \{C, D, syn\} \tag{1}$$

Assuming that synonyms of keywords might be an interesting way to connect several articles, BabelNet is the knowledge database on which our framework will rely. However, BabelNet lacks specificity and searching for one word can return synsets from various different contexts. For example, "flight" returns 36 synsets, from a South Korean movie to the verb 'to fly'. Consequently, a method to filter out unrelated synsets is mandatory.

Because synonyms are too specific, and domains are too general, categories have been naturally chosen in order to identify overlapping between synsets from different keywords. Indeed, if several keywords share the same category, then this is potentially the correct category in regards to the article context. In addition, the greater the number of keywords sharing the same category, the higher the confidence. Thus, connecting the returned synsets based on their categories is an interesting way to naturally filter out all of the unrelated synsets.

This approach does filter some content, but still returns "living people; English-language films; celestial mechanics; American films" as the main categories for keywords "nonlocal gravity; celestial mechanics; dark matter". Constant noise (**_singer*, **_album*, etc.), meaningless in our scientific context, has been identified. A parameter can now be set in order to force the automatic filtering of identified noise. Most of the remaining noise is finally naturally filtered out, and "celestial mechanics" is finally returned as the main category.

Our final goal is to apply this valuable added knowledge to all articles from the scientific literature database, Scilit (http://www.scilit.net), developed by MDPI (http://www.mdpi.com). To validate our approach, a manual analysis on a subset of 595 articles from seven journals (six about Physical Science and one about Pediatrics) has been conducted. We evaluated the correctness of the categories based on the connection of keywords by their synsets. This approach provides good precision—from 96% to 100%—depending on the threshold which identify the data as correct not. Indeed, strictly selecting only categories shared

by three different keywords or more leads to a high degree of confidence (100% precision), but a recall of 9%. By being more tolerant and considering all categories shared by at least two keywords, precision slightly decreased (96%) but we significantly gain in recall (47%). Moreover, similar proportions are observed for *Children*, the journal about Pediatrics (from 100% to 92%). This validates that our approach may be used in several domains.

The main drawback of our approach is that correct categories have been identified for only 22% of the articles within the subset. Figure 1 illustrates the reason for the law recall and coverage of our approach.

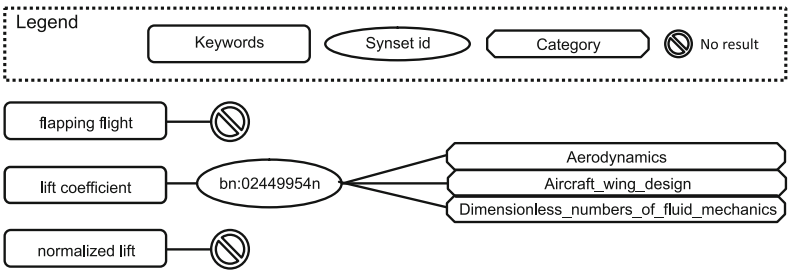

Fig. 1. Limits of the exact search: only one keyword from three return data.

One of the reasons for this law coverage is that BabelNet often returns no result for composed keywords (multi-word keywords), as shown in Fig. 1, where no data is returned for two of three keywords. In our approach—proposing only categories shared by at least two keywords—the degree of confidence is not high enough to return the categories. We will investigate further a way to propose some categories from these composed keywords in our future work. In doing so, we aim to significantly gain in recall, and cover many more articles.

3 Perspectives

Making scientific recommender systems smarter is crucial in order to help scientists in their mandatory and tedious bibliographical research phase. The approach proposed could be the first step in building such a smart system. Indeed, analysing the correctness of the main category based on the overlapping of the keywords category confirms the logic of our approach. In the future, we plan to extend the search in order to extract categories from composed keywords. Splitting on spaces would provide some data for sub-keywords. Then, applying the same logic as described in our approach (i.e., connect by common category) will filter out unrelated items, and categories from connected items might be used for the global category connection. By taking the example from Fig. 1, splitting "flapping flight" on spaces will return 3 and 25 synsets, respectively for "flapping" and "flight" (Fig. 2):

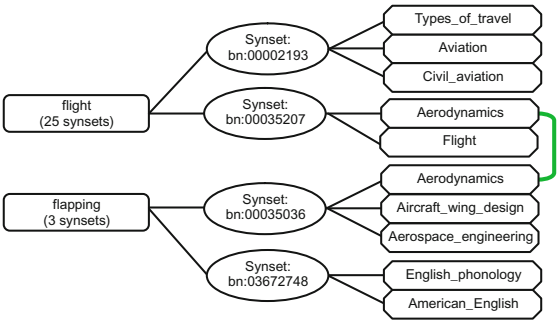

Fig. 2. The category "Aerodynamics" is returned as the main category of "flapping flight". Other categories are filtered out by this connection.

This further search will successfully identify "Aerodynamics" as the main category of "flapping flight". Thus, our approach would connect "Aerodynamics" based on both keywords. Extracting the part-of-speech (with a syntactical analyser like SyntaxNet [4] or CoreNLP [5]) from long keywords could be an interesting extra source of information for refining requests on BabelNet. Finally, Fig. 3 shows the main logic of our next contribution: to process in a smarter way the keywords that do not return any satisfactory results. Later, we might also generate a graph inherited from the BabelNet's synsets as in [6].

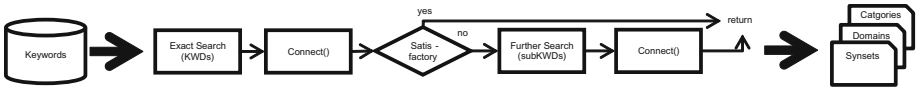

Fig. 3. Illustration of the general logic of our approach in a future work

References

1. (Editorial), N.: Gold in the text? Nature **483**(7388), 124 (2012)
2. Effendy, S., Yap, R.H.C.: The problem of categorizing conferences in computer science. In: Fuhr, N., Kovács, L., Risse, T., Nejdl, W. (eds.) TPDL 2016. LNCS, vol. 9819, pp. 447–450. Springer, Cham (2016). doi:10.1007/978-3-319-43997-6_41
3. 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)
4. Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., Collins, M.: Globally normalized transition-based neural networks. In: ACL (2016)
5. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: ACL, pp. 55–60 (2014)
6. Franco-Salvador, M., Cruz, F.L., Troyano, J.A., Rosso, P.: Cross-domain polarity classification using a knowledge-enhanced meta-classifier. Knowl. Based Syst. **86**, 46–56 (2015)

Development of an RDF-Enabled Cataloguing Tool

Lucy McKenna[1(✉)], Marta Bustillo[2], Tim Keefe[3], Christophe Debruyne[1], and Declan O'Sullivan[1]

[1] ADAPT Centre, Trinity College Dublin, Dublin, Ireland
lucy.mckenna@adaptcentre.ie
[2] UCD Library, University College Dublin, Dublin, Ireland
[3] Digital Resources and Imaging Services, Trinity College Dublin, Dublin, Ireland

Abstract. By generating bibliographic records in RDF, libraries can publish and interlink their metadata on the Semantic Web. However, there are currently many barriers which prevent libraries from doing this. This paper describes the process of developing an RDF-enabled cataloguing tool for a university library in an attempt to overcome some of these obstacles.

Keywords: Semantic web · Linked Data · MODS · RDF · Library · Interface design · Usability testing

1 Introduction

The Digital Resources and Imaging Services (DRIS) department of the Library of Trinity College Dublin (TCD) hosts the Digital Collections Repository of the university. This repository provides open access to TCD's collection of digitised cultural heritage materials which includes manuscripts, letters, books, images, and other archival materials. DRIS aims to publish the bibliographic data of its collections as RDF in order for these materials to be discoverable on the SW, increasing the visibility and use of the library's resources. Additionally, RDF metadata published by DRIS could be interlinked with Linked Data (LD) emerging from other institutions, facilitating library users to access a web of related data from a single information search [1].

2 Libraries and Linked Data

Although not yet widely used, libraries are publishing bibliographic metadata as RDF in increasing quantities [1,2]. However, librarians have reported a number barriers in using LD to its full potential including that LD software is not tailored to the specific needs and expertise of librarians but rather technical experts. Other reported challenges included a lack of authority control on the SW, difficulties establishing interlinks, and few examples of useful applications of LD

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 612–615, 2017.
DOI: 10.1007/978-3-319-67008-9_55

in the library domain that would justify the allocation of time and resources to its generation [3,4]. These challenges were experienced by DRIS and prevented the library from publishing its metadata to the SW. As such a bespoke RDF-enabled cataloguing interface was developed for DRIS. The aim of the interface was to explore whether such a tool could be used by DRIS to successfully generate MODS-RDF records for a small sample of records thus demonstrating the potential for LD software specifically designed for library use.

3 MODS and MADS

The Metadata Object Description Schema (MODS) is an XML schema for a bibliographic element set that can be used for the purpose of cataloging digital resources [5]. The full schema consists of 20 top-level elements, for example TitleInfo and Name, which are used to provide information on the title and creator of a work. The majority of MODS elements contain subelements, such as title, subtitle, and namePart, as well as attributes which describe the metadata itself, for example, the authority source from which a title or name was taken, or the language used when cataloguing.

MODS was selected as the output schema for the tool as it was sufficiently detailed for DRIS's cataloguing purposes and a MODS-RDF ontology was already available [6]. Additionally, a set of MODS implementation guidelines was developed by the Digital Library Federation's (DLF) Aquifer Initiative thus allowing for the standardisation of MODS records [7].

The Metadata Authority Description Schema (MADS) [8] can serve as a companion to MODS to provide metadata regarding the authority sources used in a record when describing names, organisations, genres, or subjects for example. Like MODS, a MADS-RDF ontology already exists [9]. Both MODS and MADS share a number of subelements, such as those in TitleInfo, Name and Subject. The schemas also share all attributes. Interestingly the MODS-RDF ontology excludes all elements it has in common with MADS. As such, in order to generate a full MODS record in RDF, both ontologies must be used.

4 Interface Design and Testing

A semi-structured interview was carried out with the DRIS metadata cataloguer in order to establish a set of tool requirements, and a mock-up of the cataloguing interface was subsequently developed. User requirements included:

- Facilitating cataloguing efficiency by automating input where possible.
- Publishing MODS records that meet DLF-Aquifer requirements by forcing data entry for certain fields and constraining data entry options for others.
- Further constraining data entry options as per the specific needs of DRIS.
- Providing additional administrative data entry fields.

The completed interface was programmed to initially constrain data entry options to only those elements and subelements which were identified as required fields by the DLF. This was done to ensure that the minimal data requirements for each record were met prior to the addition of supplementary metadata. Once these fields were complete, data entry options expanded to include recommended and optional fields.

Data entry fields and dropdown menu options were programmed to dynamically alter based on prior selections made during the cataloguing process. This ensured that data entry options were restricted to DLF recommendations. For example, in the Name element, DLF require that the resource creator's name should be taken from the Name and Title Authority Source Codes maintained by the Library of Congress (LOC). Thus the list of options in the authority menu was constrained to these sources, this was then further constrained to display only the sources used by DRIS. Data entry fields also self-populated based on prior selections allowing for a more efficient cataloguing process. For example, again in Name, after selecting an authority source the Authority-URI field self-populated. This also highlights how the tool was capable of accepting URIs to other LD datasets - a first step in the LD interlinking process.

The interface was tested by observing the DRIS metadata cataloguer using the tool to create a bibliographic record. Although results indicated some issues with the interface layout, the librarian felt that the tool would be useful for creating more authoritative RDF datasets and that it could facilitate increased LD generation by librarians rather than technical experts alone.

5 Record Generation

Data from the interface was stored in a relational database. In order to uplift this data to RDF an R2RML mapping was developed based on the MODS and MADS RDF ontologies. R2RML is a W3C Recommendation for declaring mappings from relational databases to RDF datasets [10]. In the process of adding MADS to the mappings it was noted that, unlike MODS-RDF where properties are represented individually, some MADS-RDF properties were grouped in collections including the subelements in TitleInfo and Name. Collections are a special RDF construct used to represent lists. This grouping allows for labels, such as title and subtitle, or first and last names, to be reconstructed with all elements in the correct order. However, at the time of the project, R2RML did not support the mapping of RDF collections, thus some metadata, such as subtitle, and more than one namePart were omitted. Despite this setback, semi-complete RDF records were generated for a small sample of DRIS's materials. A number of SPARQL (RDF query language) queries were successfully run over the RDF dataset including typical searches by author, date, and genre, as well as more interesting and detailed searches by ISO Language and Country Codes, authority sources, controlled vocabulary terms, and URIs.

This issue inspired a separate project in which an R2RML expansion supporting the mapping of RDF Collections (and Containers) was developed [11].

This expansion facilitated the uplift of all metadata in the database to RDF, allowing for the publication of complete MODS records.

6 Conclusions and Future Directions

Providing librarians with bespoke LD tools would allow for increased publication of rich LD datasets. It is likely that LD generated by librarians would be treated with increased credibility and thus used more frequently as libraries are viewed as trustworthy and authoritative sources of information. LD created by librarians will follow specific and standardised bibliographic schemas, and use long established authorities and controlled vocabularies to describe resources. This would increase the level of authority control on the SW, allowing for similar entities to be identified consistently across the SW leading to richer search results.

Future research will explore how to engage librarians in the process of interlinking with LD datasets published by other libraries and related institutions rather than just large scale authorities (LOC) and LD datasets (DBpedia). This would allow library users to access larger amounts of related data from single information search.

Acknowledgments. This study is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre for Digital Content Platform Research (http://www.adaptcentre.ie/) at Trinity College Dublin.

References

1. Hastings, R.: Linked data in libraries: status and future direction. Comput. Libr. **35**, 12–16 (2015)
2. Mitchell, E.T.: Library linked data: early activity and development. Libr. Technol. Rep. **52**, 5–33 (2016)
3. Hallo, M., Lujan Mora, S., Trujillo Mondejar, J.C.: Transforming library catalogs into Linked Data. In: ICERI (2013)
4. OCLC: Linked Data Survey (2017). http://www.oclc.org/research/themes/data-science/linkeddata.html
5. Library of Congress: MODS (2017). http://www.loc.gov/standards/mods/
6. Library of Congress: MODS-RDF (2012). http://www.loc.gov/standards/mods/modsrdf/v1/modsrdf.owl
7. Digital Library Federation: MODS Implementation Guidelines (2009). https://wiki.dlib.indiana.edu/download/attachments/24288/DLFMODS_ImplementationGuidelines.pdf
8. Library of Congress: MADS (2017). http://www.loc.gov/standards/mads/
9. Library of Congress: MADS-RDF (2017). http://www.loc.gov/standards/mads/rdf/mads-ontology-20101119.owl
10. W3C: R2RML (2012). https://www.w3.org/TR/r2rml/
11. Debruyne, C., McKenna, L., O'Sullivan, D.: Extending R2RML with support for RDF collections and containers to generate MADS-RDF datasets. In: TPDL (2017)