Unsupervised Mining of Knowledge Gaps in Scientific Literature

Silvia Fernandez[1], Pierre Jourlin[2], Eric SanJuan[2]

[1] Calle 17 N° 297 X 44 y 46 San Damian, 97218 Mérida, Yucatán, México.

silvia.fernandez@lookingforalab.org

[2] LIA-CERI, BP1228, 84911 Avignon, France.

{pierre.jourlin, eric.sanjuan}@univ-avignon.fr

Abstract

Literature Based Discovery (LBD) relies on the identification of gaps in the scientific literature. Most of the existing methods are supervised and rely on the use of specific large knowledge domain databases like MedLine for medical study. We present here a tractable approach based on Natural Language Processing techniques with few linguistic resources and Formal Concept Lattice exploration. Entities are automatically extracted from full text scientific papers based on their acronym forms. An unsupervised classification is build using syntax and WordNet relations. Resulting classes are clustered into multiple formal concepts and the knowledge gaps are identified in the resulting Galois Lattice. The feasibility and the relevance of the outcome is analyzed on a large corpus of full-text journal articles dealing with nuclear energy research.

Résumé

La découverte au travers de la littérature (Literature Based Discovery ou LBD) repose sur l'identification des lacunes dans la littérature scientifique. La plupart des méthodes existantes sont supervisées et s'appuient sur l'utilisation de larges bases de connaissances spécifiques telles que MEDLINE dans le domaine de la médecine. Dans cet article, nous présentons une approche fondée sur des techniques de Traitement de la Langue Naturelle (TALN), d'exploration de treillis de concepts formels et sur une utilisation minimale de ressources linguistiques. Les entités sont extraites automatiquement à partir du texte intégral des articles scientifiques en fonction de leurs acronymes. Une classification non supervisée est construite en utilisant la syntaxe et les relations de WordNet. Les classes résultantes sont regroupées en plusieurs concepts formels et les lacunes de connaissances sont définies dans le treillis de Galois induit. La faisabilité et la pertinence des résultats sont analysées sur un large corpus textuel d'articles de revues portant sur la recherche en énergie nucléaire.

Keywords: Literature Based Discovery, Natural Language Processing, Formal Concept Lattices

1. Introduction

1.1 Literature-based discovery

Literature-based discovery (LBD) consists in mining unpublished relationships between domain entities, in the aim of finding potentially interesting new hypothesis. Most of the existing tools adopt a supervised approach. In the framework of *closed discovery*, the user provides a pair of entities (A and C) never published in a single paper, and the system looks for indirect relationships between those entities. In the framework of *open discovery*, the user

provides only one entity (A) and the system provides all indirect relationships between A and any entity C for which A and C remain unpublished in a single paper.

According to Swanson D.R., who is considered as the inventor of LBD : "Any LBD search that does not begin by clearly specifying a problem can be doubly open-ended, having, like the universe, neither a beginning nor an end" (Swanson D.R. 2008).

1.2 Our approach

Given a large corpus of full scientific papers, how to detect scientific gaps among its topics, concepts or results?

To answer this question, we have no choice but dealing with this *doubly open-ended* subtask of LBD. In addition, most of LBD experiments are carried out on domains for which it exists knowledge databases like MEDLINE for medicine. Unfortunately, we have no choice but to experiment on non-linguistically, non-semantically annotated corpora.

We seek to solve this problem without restricting ourselves to a specific domain and without requiring special linguistic or semantic resources. Therefore to identify topics we rely on simple Natural Language Processing (NLP) of texts. One possible approach that we experiment here is to retrieve Multi Word Terms (MWT) based on their acronym form following (Okazaki N., Ananiadou S. 2006). We also need a formal model that allows to define and compute some kind of gap between concepts that can rely on non frequent entities. In this paper we choose Formal Concept Analysis (Wille R. 2005) as model and define the notion of gap based on underlying Galois lattices (Barbut M., Monjardet B. 1970). This choice limits the search to strong and direct relationships as opposed to the larger set of indirect relationships explored by open-discovery works on MEDLINE.

This paper is organized as follows. Section 2 describes the extraction of entities using acronyms and different ways of clustering them. Section 3 models the notion of scientific gap based on the Galois Lattice associated with the relation between extracted entities and documents. Section 4 shows the results of this approach on a corpus dealing with nuclear energy and physics. Finally, section 5 provides a discussion on previous results and open perspectives.

2. Entity mining

The overall process of text units extraction relies on simple unsupervised NLP techniques that can be applied on large scale documents collections.

2.1 Multi Word Term extraction based on acronym forms

Multi Word Terms (MWT) - noun phrases referring to a domain concept or entity - are known to play an essential role in scientific literature (Jacquemin, C. 2001). However, in an unsupervised learning process, they are difficult to distinguish from current noun phrases since they have the same form and similar frequency distributions (Jacquemin C. et al. 2002). It is thus necessary to find some linguistic markers in the text that point out the presence of a domain MWT. On large text corpora, we show here that acronyms can be used for that purpose following (Okazaki N., Ananiadou S. 2006).

2.1.1 Acronym definition extraction algorithm

The extraction algorithm can process articles in *HTML* or simple *ASCII* format directly. It starts by looking for sequences of 2 up to 10 words, that are immediately followed by at least two capitals between brackets. *HTML* tags are eliminated from the extract, as well as indexes and exponents that possibly occur after the acronym. Once this is done, the problem amounts to finding out the location of the first word of the acronym in the sequence of 2 up to 10 words that precede the sequence of capitals between brackets. The alignment of definition's words and acronym's capitals is achieved from right to left. The algorithm starts by searching the word most to the right which starts by the acronym's capital most to the right. When successful, the search starts again, with the capital at the immediate left of the last processed capital and from the location of the word most to the right which is not already associated with an acronym's capital. When unsuccessful, the current capital is ignored and the search starts again with the capital at the immediate left of the current capital and from the current location in the definition. This process is repeated until it reaches the acronym's capital most to the left.

This algorithm ensures a perfect extraction of simple acronym's definitions, wherein each word of location *L* in the definition starts with the capital of location *L* in the acronym. The extraction may be imperfect in several other contexts, for instance when *stop words* occur in the definition whilst they are ignored as capitals in the acronym or when the acronym and its definition are expressed in two distinct languages.

However, the algorithm seems to be rather robust : for instance, its captures : *du carbone organique dissous (DOC)* for *Dissolved Organic Carbon*, *elektronenmikroskopie (TEM)* and *mittels transmissionselektronenmikroskopie* (TEM) for *Transmission Electron Microscop/y/e/es/ic*, etc. A slightly more-in-depth evaluation is reported in section 4.2.1.

We made this program publicly available under a GNU Public License. It can be downloaded here : http://hypolite.termwatch.es/IMG/tgz/AcroDefScanner.tgz

2.1.2 From acronyms to MWTs

Each possible definition of an acronym gives rise to a set of MWTs. In English, most of the MWTs under acronyms are simple noun phrases without preposition, made of nouns and adjectives. In that case, the head of the MWT is the last word. The two last words (the MWT head and the closest modifier) often corresponds to a generic term (Jacquemin C. 2002). Generally, the MWT associated with an acronym can be represented as a head and an ordered sequence of modifiers as in (SanJuan E., Ibekwe-SanJuan F. 2006).

2.2 Unsupervised classification of MWTs

The detection of gaps in the literature requires to identify topics that never appear together in the publications. When topics are represented by acronyms, this requires to identify which are subsequences of others. But acronyms can only be interpreted in a local context. They can be ambiguous across the literature and refer to different concepts. On the contrary, MWTs are highly precise and detecting such subsequences is easier, taking into account that a single acronym can be represented by several variant MWTs. Generally, topic representation cannot rely on a single MWT but requires at least a cluster of them.

2.2.1 Clustering by lexical variations

The most obvious way of grouping MWTs that have been extracted based on their acronyms is to merge together those that produce the same acronym and only differ on small lexical variations. We shall refer to these clusters as **lexical clusters**.

2.2.2 Clustering by syntactic variations

In most MWTs that we extracted from acronyms, the head (last word) and the closest modifier (second last word) can be considered as a MWT root. Syntactic relations that preserve these roots tend to form coherent clusters (SanJuan E., Ibekwe-SanJuan F. 2006). We consider four kind of variations:

1. Left expansion (L-exp) that consists in appending modifiers on the left.
2. Insertion (Ins) of some modifiers at a single place.
3. Substitution of a word by another in a common WordNet synset (Sub-wn).
4. Substitution of a modifier in a MWT of length 3 or more (Sub-mod-3)

These four variations define a non-directed simple graph on terms. The MWTs are the vertex and we insert one edge between two vertex whenever there exists at least one variation between them. We then take as clusters the connected components of the graph. We shall refer to this clusters as **syntactic clusters**.

2.2.2 Clustering by association

We shall also group together MWTs that appear almost always on the same documents. Terms *x, y* such that the conditional probabilities of finding one knowing the presence of the other are closed to 1: $P(x/y) \times P(y/x) \simeq 1$. This is known as the equivalence index (Courtial, J-P 1989) and it is a variant of the mutual information index. However, this index is unable to relate synonyms of rare terms that tend to appear in disjoint documents.

We do not consider associations alone. Given a fixed threshold (80%), we add the resulting associations as new edges to the graph of term variants and we consider likewise the resulting connected components.

3. Mining potential scientific gaps

We consider the relation between the classes of terms and the documents. A class considered as related to a document when at least one of the class' MWTs appears in the document. All terms in the class are considered as referring to the same topic. A formal concept is then identified by a closed set of classes. By closed set we mean that for any other class not in the cluster there is at least one document that is not related to it but to all classes in the cluster (Carpineto C., Romano G. 2005). The set of classes is called the intension and the set of document the extension (Wille R. 2005). Formal concepts can be ordered by set-inclusion on the intensions. The structure of the resulting partially ordered set is a inf-sublattice of the complete lattice of all sets of classes. It is in fact isomorphic to the Galois lattice of the relation between classes and documents (Barbut M., Monjardet B. 1970). The idea is that gaps in the literature should correspond to gaps between the Galois lattice and the complete lattice.

3.1. Formal definition of gaps

There are two patterns of gaps that are usually considered in LBD (Yetisgen-Yildiz M., Pratt W. 2009). The first one is a missing transitive edge. The second one that we shall consider

here rely on the existence of disjoint sub-concepts. The first pattern covers a much larger set of hypothesis, from simple term-to-term transitions to complex paths. The downside of the approach is the presence of weaker links, necessitating a much more complex formalism. As a first attempt to automatically mining these gaps, we therefore decided to opt for the second pattern.

3.1.1. Galois lattice

Let us recall the main properties of a Galois lattice derived from a finite relation. Let D be a set of documents, A a set of document descriptors (acronyms or clusters of MWTs). Let R be the relationship between D and A (R is a subset of $D \times A$). Given a set X of documents, we shall denote A(X) the set of common descriptors: $\{a \in A | (\forall d \in X) (d,a) \in R\}$. Likewise, given a set Y of descriptors we shall denote by D(Y) the set of documents that have all these descriptors: $\{d \in D | (\forall a \in Y) (d,a) \in R\}$. A formal concept is a pair (X,Y) such that A(X)=Y and D(Y)=X. X is called the extension of the formal concept and Y is called the intension. The operator $D \circ A$ is a closure on the set of documents. Likewise, $A \circ D$ is a closure on descriptors. These two closure operators define two isomorphic lattices and the set of formal concepts can be structured itself as a lattice such that the extension of the lower bound of two formal concepts (X_1,Y_1) and (X_2,Y_2) is $X_1 \cap X_2$ and the intension of the upper bound is $Y_1 \cap Y_2$.

3.1.2. Lattice gap

Let $S=(X_1,Y_1)$ and $T=(X_2,Y_2)$ be two formal concepts. We shall consider that there is a gap between S and T if we have all of the following:

1. $X_1 \cap X_2$ is empty
2. $Y_1 \cap Y_2$ is not empty
3. $D \circ A(Y_1 - Y_2) \cap D \circ A(Y_2 - Y_1) = Y_1 \cap Y_2$

The first condition states that the two formal concepts do not appear together in the papers. It is an unpublished relationship. Meanwhile the second condition guarantees that there is some relationship between them since they share common descriptors. The last condition reinforces the strength of the relationship since it requires that in each formal concept, the subset of common descriptors is the closure of the remaining descriptors: $Y_1 \cap Y_2 \subseteq D \circ A(Y_i - Y_j)$ for distinct i,j in {1,2}. In relational database language we shall say that they are keys of their respective intensions.

3.2. Model adaptations

Previous model cannot be directly applied for combinatorial reasons. Some adaptations are required. Here are the ones that we chose with their justifications.

3.2.1. sub-semi lattice of frequent closed item sets

The formal concept lattice size is exponential to the number of descriptors and documents. But in practice, the combinatorial explosion happens for formal concepts with small extensions. The cardinal of the extension is usually referred as the support. The set of formal concepts with support under some fixed threshold inherits part of the lattice structure since it is a sub join-semilattice. Therefore, by setting a small threshold on the support, we only consider the frequent item sets, i.e. sets of descriptors that occur frequently. By this mean, we

managed to keep combinatorial explosion to a tractable level. Moreover the formal concepts that are avoided are also the most hazardous. In practice we fix a threshold between 2 and 5.

3.2.2. Gap amplitude

The lattice model does not allow to define a gap amplitude. Meanwhile the set of possible gaps remains exponential despite the threshold based lattice reduction. It is thus necessary to implement some way of ranking gaps according to some criteria. Only the top ranked shall be submitted to users. In this experiment we simply consider the euclidean distance between formal concept extensions.

3.2.3. Implementation

The frequent item sets are computed using *apriori* algorithm. Potential gaps are computed by first looking for formal concepts with overlapping intensions. For those pairs of concepts we then keep those which have disjoint extensions. Then we check the third hypothesis in 3.1.2. All these set operations can be very efficiently implemented using fast grouped binary comparisons, specially on clusters of 64-bit processors. Finally, the most time consuming is the computing of the euclidean distance between extensions since it involves non binary operations. However, the size of the document collection we processed did not lead to problematic computing time : from a few minutes to a few days depending on term frequency thresholds. We also believe that maximizing computing performance is still possible by machine-dependent or machine-independent code optimizations.

4. Results

Our approach is intended to work on any *HTML* corpus of scientific documents downloaded from journal websites. Indeed, today most of the researchers access directly from their computer desktop to the electronic version of papers via their research center intranet. They also tend to read these papers on the screen. This has impulsed the development of interactive *HTML* versions not meant to be printed out but read online.

The journal websites offer complete end efficient Information Retrieval interfaces. Domain mapping tools are also integrating these websites. Moreover, external engines can also be used to query these websites. Therefore, the researcher usually uses his favorite engine to carry out an interactive search on the journals on which he relies. If the information is there, the researcher will find it by this way. Our purpose here is to help the researcher in finding what is missing between known topics and highlight surprising scientific gaps in his domain on which he could work.

We are experimenting our approach in two completely disjoin domains: forest ecology with the INRA (National Institute for Agronomic Research) and nuclear research with one of the libraries of the CEA (National Atomic Energy Center). Both results are available online on our website project: http://hypolite.termwatch.es . In this paper, we shall comment the current results on atomic research, while focusing on these two questions :

1. what is the topic coverage of our shallow entity extractor based on acronyms?
2. is the set of Formal Concept Lattice gaps as defined in 3.1.2 tractable?

4.1. Corpus

We have processed all articles published since 1990 from four majors journals in atomic energy research as listed below:

Journal	Number of articles
Journal of Nuclear Materials	9886
Nuclear Engineering and Design	3584
Annals of Nuclear Energy	2514
Progress in Nuclear Energy	1365

Tab 1: number of articles published since 1990 per journal

4.2 Extracted Entities

8 477 different acronyms were found in the corpus corresponding to 18 428 possible definitions that are MWTs.

4.2.1. Acronym evaluation

We have evaluated the 300 acronyms that appear most frequently, each of them being associated with most frequently related MWT. This results in evaluating the alignment between a set of acronyms that appear at least 32 times in the corpus with the most frequent definition found by the algorithm. Table 2 summarizes the results:

	Forms	Occurrences
Total number	22172	76424
Number of evaluated	300	26542
Number of correct	286	25499
% evaluated	1,35%	34,73%
% correct	95,33%	96,07%

Tab 2: evaluation of 300 most frequent acronyms

In the nuclear physics domain, the acronyms are mostly used for naming technical entities: spectrometry techniques, methods of experimentation, companies generating electricity by nuclear, experimental nuclear facilities or laboratories, identified accidents due to the utilization of certain materials, protocols and test security, among others. On a somewhat more theoretical entities, we find simulation methods. Some examples of acronyms detected: GFR (Gas-cooled Fast Reactor), LOCA (Loss of Coolant Accident), JAERI (Japan Atomic Energy Research Institute), EDF (Electricité de France), ROCOM (Rossendorf Coolant Mixing Model).

4.2.2. MWTs analysis

It results from previous evaluation that frequent acronyms are unambiguous. However, this is not the case for infrequent acronyms. To deal with them it is necessary to handle their MWT form. A total of 18 428 multi word noun phrases were extracted as being some acronym definition. On a pool of 400 noun phrases of any frequency, 15 appeared not to be domain MWTs. All of them appeared to be noun phrases like: *additional set, matching section, thermally aged, work input*. The rest of the pool corresponded to real MWTs. Therefore it appears that acronyms are a very efficient completely automatic way of extracting real domain MWTs without requiring any extra knowledge resource.

Among these extracted MWTs, there are:

1. 387 lexical variations i.e. pairs of terms that only differ in the use of a "-" and use the same lemmas but in different forms: plural, gerundive etc.
2. 329 MWTs use synonyms.
3. 4 302 are left expansions one of each other.

Therefore, while lexical and synonym variations are marginal, left expansions are not. This phenomena has to be taken into account when considering that two terms never occur together. The non-occurrence could be the simple consequence of one term involving the other.

4.3 Evaluation of the detected gaps

The user is presented with a list of *unpublished* relations between a class A and a class C of acronyms or MWTs.

Referring to section 3.1.1 :

- A is the set of terms obtained by subtracting Y to X (A=X-Y)
- B is the set of terms obtained by intersecting X and Y (B=X∩Y)
- C is the set of terms obtained by subtracting X to Y (C=Y-X)

These classes are the intension of some formal concept. Although A∪C occur in no document, when all elements of class A **or** class C occur together in a document, a third class B occur as well. Our hypothesis is that A and C are related in some way via this class B, even though the relation that links them together is unpublished in our document collection. These hypothesis are ranked by decreasing support of A and C (intuitively, the most important entities in the literature), then by decreasing support of B (intuitively, the strongest link between the entities) and finally by decreasing distance between A and C (intuitively, the most foreign entities in the literature). However, we are still testing other ways of ranking, especially those allowing to take user's feedback into account.

All results are available at the already mentioned address: http://hypolite.termwatch.es.

4.3.1. Based on lexical clustering

8 286 distinct acronyms generated 23 447 frequent item sets and 389 579 hypothesis of scientific gap.

Given the nature of the information conveyed by the acronyms, some examples of relationships found are:

1. The laboratory A uses the experimental technique B. There is another technique C closely related with B which is not used in this laboratory.
2. Company A has a B type reactor. C is a security protocol associated with the use of B but does not appear to be applied by A.
3. The use of some material is linked to certain types of accidents (A). A reactor of type B uses this material. Some company C has one of these reactors but no document relates it to A. Here, A is composed of two acronyms (material and accident).

In all these relations, the generator of the class of acronyms is a singleton, i.e. an acronym which never appears without the other ones in the same class.

Among the top ranked relations acronyms exists that are expansions of others. These potential gaps are irrelevant since there is an obvious relationship between the two entities. This artifact will be avoided using MWT definitions.

A thorough objective analysis of the remaining 15 most relevant unpublished relations between a class A and a class C, has revealed that most of these gaps have a valid reason behind. For example, a laboratory may not have used some technique that seems best suited to a study, simply because it did not exist at the time the measurements were made. It is worth mentioning that these relationships do not explicitly appear in the corpus but can be checked on public data accessible on the web.

But on the other hand, there subsists among these top ranked relationships a small set for which we have not found in the literature the valid arguments to justify the lack of implication between concepts enclosed by the set A and the set C. Are they the true alerts? This question can only be properly answered by those responsible for the techniques involved in these gaps.

4.3.2. Based on syntactic clustering

The 18 240 MWTs have been clustered into 2193 clusters, based on syntactic variations and associations. There are slightly less classes than acronyms. This is mainly due to the fact that MWTs clusters contain not only the definitions of a single acronym, but also those that are expansions of others.

These clusters give rise to 32 583 closed frequent item sets and 5 082 578 lattice gaps.

While the drawback of obvious syntactic relationships is avoided working on clustered MWTs, another drawback appears caused by non clustered MWTs. These ones have much lower frequencies than classes themselves and tend to generate relationships between equivalent definitions. These isolated MWTs dramatically increase the number of lattice gaps. However it is easy to avoid this if we considerer other occurrences of these MWTs than those detected closed to an acronym. We thus have used the standard IR engine Indri (http://www.lemur.edu) that allows to seek phrases in the texts to check the empty intersection between formal concepts extensions.

Once this is done, the top results are the same as for lexical clustering. Moreover, due to the possibility given by the MWTs to back and forth with the full text contain, further ranked relationships better corresponds to unpublished relationships. But a closer analysis shows that most of them can be explained on the time line where new techniques have replaced other ones. Meanwhile this transition is not mentioned in the journals, it can be checked on public available data. The few hypothesis for which we couldn't find any reasonable explanation are the same.

5. Discussion and perspectives

These results we obtained are surprisingly positive and promising, even though we could not get a proper validation by our partners in CEA before submitting this paper.

The techniques we developed have a high level of scalability and they can be applied to any scientific domain for which large full-text publications databases are available. That is to say the large majority of scientific domains.

In the nearest future, we intend to enlarge the corpus of documents in each field we already studied (forest ecology, nuclear energy) in order to decrease the probability of retrieving unpublished relationships in our document collections that are actually published elsewhere.

We are planning to extend the set of MWTs by adding more linguistic markers such as chemical notation of atoms, molecules or alloys. We also plan to increase the number of indicators of *relevance* for our A-B-C hypothesis.

We already started to explore the feasibility of classifying the terms into several semantic categories, represented by the set of verbs of which they can be subject or object. This will allow us to perform a semantic categorization of hypothesis that could be taken into account in the ranking formula.

However, the most exciting perspective is to confront the target domains researchers with our results, to evaluate results accuracy and develop techniques that lead to improvement through relevance feedback and user interaction. First contacts in this direction have been made with researchers of CEA on nuclear energy, of INRA on forest ecosystems, of INRIA on telecommunications networks, but we are still looking forward new partnerships.

We hope these partnerships will lead us to find out a few hypothesis worthwhile exploring further with the use of the more conventional open- and closed-discovery LBD techniques.

References

Barbut M., Monjardet B. (1970) *Ordres et classifications : algèbre et combinatoire*, Paris, Hachette, 174 p.

Carpineto C., Romano G. (2005) Using Concept Lattices for Text Retrieval and Mining. In Bernhard Ganter B., Stumme G., Wille R. (Eds.) *Formal Concept Analysis, Foundations and Applications. Lecture Notes in Computer Science 3626, Springer,* pp: 161-179.

Courtial, J-P (1989) Qualitative Models, Quantitative Tools and Network Analysis, *Scientometrics*, 15(5-6):527-534.

Jacquemin C., Daille B., Royauté J., Polanco X. (2002) In vitro evaluation of a program for machine-aided indexing. *Information Processing & Management*, 38(6): 765-792

Jacquemin, C. (2001). *Spotting and discovering terms through NLP*. Cambridge, MA: MIT Press.

Okazaki N., Ananiadou S. (2006) A term recognition approach to acronym recognition. In *Proc of the COLING/ACL*, pp. 643-650.

SanJuan E., Ibekwe-SanJuan F. (2006) Text mining without document context. *Information Processing & Management*, 42(6): 1532-1552.

Swanson D.R. (2008) Literature-Based Discovery? The Very Idea. In Bruza P., Weeber M. (eds), *Literature-based Discovery,Series: Information Science and Knowledge Management, Vol. 15*, *Springer*, pp 3-12.

Wille R. (2005) Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies In Ganter B., Stumme G., Wille R. (Eds.) *Formal Concept Analysis, Foundations and Applications. Lecture Notes in Computer Science 3626, Springer,* pp: 1-33.

Yetisgen-Yildiz M., Pratt W. (2009) A new evaluation methodology for literature-based discovery systems. *Journal of Biomedical Informatics,*42(4):633-643.