Actes de l'atelier Recherche d'Information SEmantique, RISE 2013

J.P. Chevallet, Catherine Roussey, H. Zargayouna, J.P. Chevallet, Catherine Roussey, H. Zargayouna

Cinquième Atelier Recherche d'Information SEmantique RISE, Lille 01 juillet 2013

Associé à la Plate-forme IA 2013

ACTES DE L'ATELIER RECHERCHE D'INFORMATION SEMANTIQUE RISE 2013

Édité par
Jean-Pierre CHEVALLET, LIG, Grenoble (France)
Catherine ROUSSEY, IRSTEA, Clermont Ferrand (France)
Haïfa ZARGAYOUNA, LIPN, Paris (France)

Cinquième édition de l'atelier Recherche d'Information SEmantique

Atelier Recherche d'Information SEmantique
RISE, Lille 01 juillet 2013

Associé à la Plate-forme IA 2013

---

1. Introduction

Nous avons le plaisir d'organiser à Lille la cinquième édition de l'atelier Recherche d'Information SEmantique, RISE 2013, associé à Plate-forme d'Intelligence Artificielle et avec le soutien de l'ARIA (Association francophone de Recherche d'Information et Applications).

Le but de l'atelier est de proposer un espace d'échange autour de la synergie entre acquisition et gestion de ressources sémantiques (ontologies, terminologies, thesaurii, ...) et la Recherche d'Information. Ces thématiques sont à la croisée du Web Sémantique, de l'Ingénierie des Connaissances, du Traitement Automatique des Langues et de la Recherche d'Information.

Les thèmes couverts par les contributions acceptées à RISE sont les suivants :

- Modèles de Recherche d'Information Sémantique
- Ontologies et Annotation Sémantique
- Mesures de similarité sémantique.

Les propositions ont porté sur des domaines variés : agriculture, médecine, botanique, patrimoine culturel et juridique.

La conférence invitée de cette année a pour titre « Semantic Search Evaluations: Gaps, Challenges and Best Practices ». Elle met l'accent sur les questions d'évaluation de l'accès à l'information dans le Web Sémantique avec des premiers retours d'expériences et un ensemble de bonnes pratiques. Nous remercions Khadija Elbedweihy, de l'université de Sheffield, d'avoir accepté notre invitation.

2. Comité de programme

- BELLOT Patrice, LSIS Avignon (France)
- BERTIN Marc, STIH Paris (France), CIRST Montreal (Canada)
- CABANAC Guillaume, IRIT Toulouse (France)
- CALABRETTO Sylvie, LIRIS Lyon (France)
- CHEVALLET Jean-Pierre, LIG, Grenoble (France)
- DAMAS Luc, LISTIC, Annecy (France)
- GRAU Brigitte, ENSIIE (France)
- HERNANDEZ Nathalie, IRIT Toulouse (France)
- KAMEL Mouna, IRIT Toulouse (France)
- PINET-SAUVAGNAT Karen, IRIT Toulouse (France)
- ROCHE Christophe, LISTIC, Annecy (France)
- ROUSSEY Catherine, IRSTEA, Clermont Ferrand (France)
- SCHWAB Didier , LIG-GETALP, Grenoble (France)
- SERASSET Gilles, LIG, Grenoble (France)
- ZARGAYOUNA Haïfa , LIPN, Paris (France)
- ZWEIGENBAUM Pierre, LIMSI (France)

Semantic Search Evaluations: Gaps, Challenges and Best Practices

Khadija ELBEDWEIHY

University of Sheffield, Regent Court, 211 Portobello, Sheffield, UK
k.elbedweihy@dcs.shef.ac.uk

Résumé : Recent work on searching the Semantic Web has yielded a wide range of approaches with respect to the underlying search mechanisms; result management and presentation; and indeed the style of input. Each approach impacts upon the quality of the information retrieved and the user's experience of the search process. Despite the wealth of experience accumulated from a variety of Information Retrieval (IR) evaluations, evaluations for searching the Semantic Web have largely been developed in isolation with no coherent overall design. This has led to slow progress and low interest when compared to other established evaluation series, such as TREC for IR or OAEI for Ontology Matching. Thus, part of this talk will discuss the missing aspects in current semantic search evaluations and the challenges they are facing and present a set of best practice procedures for designing semantic search evaluations which are motivated by the IR literature and our experience in running semantic search evaluations. Additionally, it is acknowledged that usability and user satisfaction are of paramount importance when designing interactive software solutions. Furthermore, the optimal design can be dependent not only on the task but also on the type of user. Evaluations can shed light on these issues; however, there has been a limited focus on assessing the usability of semantic search systems in current evaluation initiatives. Therefore, the other part of this talk will present the methodology and results of a first-time user-based study that assessed the usability and user satisfaction of different semantic search query input approaches (natural language and view-based) from the perspective of different user types (experts and casuals).

Mots-clés :semantic search, evaluation, user-based study

Language Model: Extension to the Similarity of Non-matching Terms in Retrieval Time

Kian-Lam TAN, Jean-Pierre CHEVALLET, Philippe MULHEM

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS,
LIG UMR 5217, Grenoble, F-38041, France
{Kian-Lam.Tan,Jean-Pierre.Chevallet,
Philippe.Mulhem}@imag.fr

Résumé : With the explosive growth of online information such as email messages, news articles, scientific literature and many kinds of information on the web, a powerful Information Retrieval System (IRS) is required to manage and provide the best result for users that are both effective and efficient. Most of the IRS use the term intersection approach to develop the matching function. However, this approach does not have a good coverage to solve the problem of term mismatch where different terms are used to describe the same meaning of an object. This paper presents an approach to solve the problem of term mismatch by incorporating the Term Similarity Matrix into the Language Model. Our approach is tested on the Cultural Heritage in CLEF (CHiC) collection which consists of short queries and documents. The results show that the proposed approach is effective and has improved the accuracy in retrieval time.
Mots-clés : Information Retrieval, Term Similarity, Term Mismatch.

1 Introduction

Numerous cultural heritage materials are now accessible through online digital library portals. This made the available resources comparatively burden less to be obtained internationally. However, these materials heavily dependent on the annotator from different cultural heritage institutions and indirectly cause the problem of inconsistency and incompleteness in the metadata. For example if an annotator uses “18th century” to describe an object created in year “1756”, then this annotation is incomplete. If another annotator uses the specific year like “1756” to describe

the same object, then two the annotations are inconsistent regarding the content of the annotation. Inconsistency may also refer to the structure of the annotations themselves : for instance some annotators might insert all the descriptions into the same metadata field, and others may split it into multiple metadata fields. In such cases, the information of an object may differ depending on the human annotators.

In this paper, we try to take into account such problems in the context of Information Retrieval (IR). Language Models for IR has been proven that it is a very effective on text retrieval based on [10] and [18]. The extension that we propose in this paper is to integrate the term links (Term Similarity Matrix) into the Language Model based on Dirichlet Smoothing (the most effective Smoothing technique according to [18]). Our proposal has the following advantages: a) it is easy and simple to generate the term links (Term Similarity Matrix) based on statistical information if compare to synthetic queries in [2] or mutual information [7] which we considered as heavy method, and b) it is easy to integrate the term links (Term Similarity Matrix into the Language Model. Moreover, we show that our approach is better than the Dirichlet Smoothing as shown in Section 5.

The rest of this paper is organized as follows. Firstly, we present the state of the art in Section 2. Then, we discuss our idea and approach in Section 3 follow by experiment in Section 4. Finally, we conclude our results and future works in Section 5.

2 State of the Art

In the past, many IR models such as Vector Space Model (VSM) [13, 14], Probabilistic Model [12, 1] and Language Model [10, 18] have been proposed which based on term intersection approach. Term intersection is the approach where both document and query should share the same terms. Although this approach provides a good result in terms of speed and accuracy, but it does not cover the problem of term mismatch which the document does not compromise the same term with the query.

There are a number of approaches that have been proposed to solve the mismatch of terms' problem by using Language Model such as the recent work from [7] and [2] that proposed to use Statistical Translation Model. The main difference between these two works is Berger and Lafferty in using synthetic queries [2] while Karimzadehgan and Zhai proposed to use mutual information to generate term links [1] [7].

1. Term links refer to the relationship between the term t and t'

2.1 Term Links

As mentioned earlier, our goal of this research is to integrate a Term Similarity Matrix into the Language Model. Before we build the Term Similarity Matrix, we need to find the links between all the terms in the collection naming V this vocabulary.

$$\forall(t, t') \in V, 0 \leq Sim(t, t') \leq 1 \tag{1}$$

1. $Sim(t, t') = 0$, there is no link between the term t and t'
2. $Sim(t, t') < 1$, there is a link between the term t and t'
3. $Sim(t, t') = 1$, there is exact match between the term t and t'

Basically, we make the assumption that two terms are considered link to each other if both terms co-occur in the same context. We use this assumption to build the Term Similarity Matrix. For example, a user is searching for the information about a "temple in ceylon". The user then submits the query :

$$q = (temple, ceylon)$$

and an IRS considering the documents below:

$d_1 = (temple, india, buddhist, god)$
$d_2 = (india, temple, sri, lanka)$
$d_3 = (roman, temple)$
$d_4 = (india, gautama)$

First and foremost, the IRS assigns a very similar Retrieval Status Value (RSV) to d_1 and d_2, and d_3 (depends on the indexing weights) because these documents contain similar terms as the query which is the "temple". However, we know that d_3 is surely not relevant since d_3 contains the information of roman temple and not the information of ceylon temple. In addition, we can argue that d_4 is more relevant than d_3 if we compare d_3 with d_4.

If we have the Term Similarity Matrix which contains the link between the term of "ceylon" and "india", then the IRS will return d_1, d_2, d_3 an d_4. Based on this example, it motivates us to consider the non-matching terms by exploiting the term similarity between the query and the document.

Several techniques in [8] have been proposed and the most important methods are : 1) dimension reduction (stemming approach, manual thesaurus, latent semantic indexing), 2) query expansion (automatic, manual

or interactive query expansion) and 3) relevance feedback (explicit, implicit or blind feedback). The main differences between these techniques and our proposed approach is that we do not add any extra terms to the query or the document. Our approach only interferes the Relevance Status Value (RSV) value during the matching process.

2.2 Exploiting Term Similarity

2.2.1 Vector Space Models

Crestani [5] proposed a general framework to exploit the term similarity into the matching process based on the variant where $w_d(t)$ is the weight which assigned to term t in the context of document d, and $w_q(t)$ is the weight assigned to term t in the context of query q as shown below :

$$RSV(d,q) = \sum_{t \in q} w_d(t) w_q(t) s \tag{2}$$

In order to visit the non-matching terms in the document, Crestani [5] proposed to exploit the term similarity by utilizing a Sim function. If $Sim(t_i, t_j) = 1$, it means that t_i and t_j are using the same term or we can rewrite it as $t_i = t_j$. If $Sim(t_i, t_j)$ is close to 1, t_i and t_j are strongly related that t_i and t_j can be used to express the same concepts and otherwise is 0.

The proposed idea by Crestani [5] is an extension of the RSV formula (2). The main idea was to extend the matching process that includes a new intermediate term t^* which contains the link between the term t from the document with the term t from the query. Essentially, the term t^* does not need to appear in the query, but as long as the term t^* contains the link with the term t, then it can be used to extend the matching process.

Given a term t from the query [2] means we need to consider all the terms in the document. The main idea of this approach is to consider the term t^* which is the maximum or highest value based on Sim function for a given t in the query as shown below:

$$t^* = \underset{t' \in d}{argmax}(Sim(t, t')) \tag{3}$$

This means that t^* is chosen among the terms that belong to document d because t^* is the maximum or highest value from the Sim function. If

2. "t from the query" or "from document" refer to the weight of t that is not null in the query, or in the document i.e. $w_q(t) > 0$ or $w_d(t) > 0$

the term t appears in the document d, then the best term t^* should be t itself or we called it as exact match. If t does not appear in the document d, then the weight of t is substituted by a term t^* in the document if there is a similarity value between t and t'. The similarity value should be lower than the value of the exact match which is $t^* = 1$ and it changes the formula in the following way:

$$RSV_{max(q \triangleright d)}(d, q) = Sum_{t \in q} Sim(t, t') w_d(t') w_q(t) \tag{4}$$

or based on (3), it changes the formula in the following way:

$$RSV_{max(q \triangleright d)}(d, q) = Sum_{t \in q}(t^*) w_d(t') w_q(t) \tag{5}$$

In other words, if t^* exists in document d, then the similarity score should be 1 and the formula will be the same as (2). For this reason, we can conclude that the proposed idea by Crestani [5] is an extension of the inner product of the vector d where the term does not appear in the document, but through the Sim function.

Furthermore, Crestani [5] proposed another solution which considers the total value of all non-matching terms in the evaluation of the RSV and this new value $RSV_{tot(q \triangleright d)}$ is defined by:

$$RSV_{tot(q \triangleright d)}(d, q) = Sum_{t \in q} \left(Sum_{t' \in d} Sim(t, t') w_d(t') \right) w_q(t) \tag{6}$$

The main difference for this approach is to use the total value instead of the maximum or highest value for the term t^*. All possible similar terms are used to compute the new weight $w_d^{tot}(t)$:

$$w_d^{tot}(t) = Sum_{t' \in d} Sim(t, t') w_d(t') \tag{7}$$

From the matrix point of view, the computation above equivalent to a matrix product between the similarity matrix Sim and the document vector d and it produces a new extended document vector d_1:

$$d_1 = Sim \times d \tag{8}$$

From the graph point of view, if we consider the Sim matrix as a weighted graph, then this is equivalent to move one step into the graph which is the total value for the term. Hence, it is equivalent to *extent the document* d by using the similarity graph (matrix) and the formula can then be rewritten in:

$$RSV_{tot(q \triangleright d)}(d, q) = (Sim \times d)^\top \times q = d_1^\top \times q \tag{9}$$

2.2.2 Language Models

Recently, Language Models have received considerable attentions because it is based on statistical foundation and a good empirical performance [10][18] and this motivated us to integrate the Term Similarity Matrix into such model. A reference paper from Karimzadehgan and Zhai which proposed to integrate the term similarity (translated into probabilities) into the Language Model based on Statistical Translation Model [7]. In addition, they rely on data from the corpus itself like synthetic queries as in [2] and not from other resources. In some ways, we can consider that their proposal is related to the second proposition from Crestani [5] which the idea is to consider the similarity between each term from query term and the terms from document. The results obtained by Karimzadehgan and Zhai [7] showed that the integration between the term similarity and Language Model is more efficient and more effective than the existing approaches in Information Retrieval.

However, Karimzadehgan and Zhai [7] noticed that the self-translation probabilities lead to non-optimal retrieval performance because it is possible that the value of $P(w|u)$ is higher than $P(w|w)$ for a term w. In order to overcome this problem, Karimzadehgan and Zhai [7] defined a parameter to control the effect of the self-translation.

In a nutshell, we can remark that 1) the normalization of the mutual information is rather artificial and required a parameter to control the effect of the self-translation, and 2) the regularization of the initial transition probabilities may look uncertain.

3 Proposal

As mentioned earlier, our goal is to integrate the Term Similarity Matrix into the Language Model. After the reviews of Crestani [5], Karimzadehgan and Zhai [7] and Zhai [17], we had considered the problems and propose to use the approach as shown below:

- We propose to use the maximum or highest value instead the total value from the term similarity between the terms from query with the terms from document. Besides, we only consider the point we view of a query if we cannot find a term in the document, then we consider the closest semantic terms from the document.
- We propose to use statistical approach rather than probability approach in order to avoid the value of $P(w|u)$ is higher than $P(w|w)$ for a term w obtained by Karimzadehgan and Zhai [7].

3.1 Extended Dirichlet Smoothing

The Language Model approach in IR was proposed by Ponte and Croft [10]. The basic idea of Language Model is to assume that a query q which is generated by a probabilistic model based on a document d as shown below:

$$P(d|q) \propto P(q|d).P(d) \tag{10}$$

$P(q|d)$ is the query likelihood for the given document d matches with the query q. If we consider that every document is equally relevant to any other query, then we can discard the $P(d)$ parameter and we can rewrite the above formula as shown below:

$$P(d|q) = \sum_{w_i \in C} c(w,d).P(w|d) \tag{11}$$

where $c(w,q)$ is the count of words w in query q and C is a set of vocabulary. Based on the multinomial distribution, the simplest way to estimate $P(w|d)$ is through the maximum likelihood estimator:

$$P_{ml}(w|d) = \frac{c(w,d)}{|d|} \tag{12}$$

where $|d|$ is the total length of the document d. Due to the data spareness problem, the maximum likelihood estimator directly assign $null$ to the unseen words in a document. Smoothing is a technique to assign extra probability mass to the unseen words in order to solve this problem.

Basically, Dirichlet [18] is one of the smoothing technique which based on the principle of adding an extra pseudo term frequency which is $\mu P(w|C)$. The Dirichlet smoothing is obtained by taking into account the extra pseudo term frequency distribution:

$$P_{\mu}(w|d) = \frac{c(w;d) + \mu P(w|C)}{\sum_w c(w;d) + \mu} \tag{13}$$

The main idea of this research is to integrate the Term Similarity Matrix into the current Dirichlet formula. Firstly, we need to assume that a term w is $w' \in d$ can play the role of w where w is $w \in q$ during the matching process. More specifically, we consider that if w does not occur in the initial document d, but it occurs in the *document* d_{ext}, which is the result of the extension of d according to the query and some knowledge [3]. Then,

3. The knowledge refers to the Term Similarity Matrix

the probability of the term will define according to the extended document d_{ext}.

The knowledge assumes to form a symmetrical similarity function which is $Sim : V \times V \rightarrow [0, 1]$, that denotes the strength of the similarity between two terms from the vocabulary (the larger the value, the higher the strength). We propose that : $\forall w \in V, Sim(w, w') = 1$ if exact matching between w with w', and $\forall w \in V, Sim(w, w') = 0$ if w does not contain any link with w'.

In order to avoid any complex extensions (see the state of the art), we define the following constraints :

– one query term w must only impact occurrences of one document term w' ;

To achieve this, we use some simple and sensible heuristics:

1. If a query term w occurs in a document d, then the term will not change the length of the document.
2. If a query term w does not occur in a document d but the term w contains a link with w' (term from document), then we define $w'' = argmax_{w' \in d, w' \neq w} Sim(w, w')$ as the term from the document will serve as the basis count of the pseudo occurrences of w in d as $c(w''; d).Sim(w'', w)$. This pseudo occurrences of the term w'' are then included into the size of the extended document.
3. If a query term w does not occur in the document and does not contains any link, then it's occurrences is counted in the extended document.

Eventually, using usual set of notations for the terms that occur in the document and the query, then the new length of the document ($|d_{ext}|$) is:

$$|d_{ext}| = \sum_{w \in d \cap q} c(w; d) + \sum_{w'' \in d \setminus q; Sim(w,w'') \neq 0} c(w''; d).Sim(w'', w) + \sum_{w' \in d \setminus q; Sim(w,w')=0} c(w'; d)$$

with w" defined above for one query term w so that:

$$w'' = argmax_{w' \in d, w' \neq w} Sim(w, w') \tag{14}$$

Using the fact above, the expression of $|d_{ext}|$ can be easily simplified into:

$$|d_{ext}| = |d| + \sum_{w'' \in d \setminus q; Sim(w,w'') \neq 0} c(w''; d).Sim(w'', w) \tag{15}$$

Remind that our proposal is to extend the document according to the query. With all the elements described above, the extended Dirichlet Smoothing leads to the following probability for the term w of the vocabulary V in the document extended d_{ext} according to a query q, noted that $p_\mu(w|d_{ext})$ is defined as:

1. if $w \in d \cap q$:

$$P_\mu(w|d_{ext}) = \frac{c(w;d) + \mu P(w'|C)}{|d_{ext}| + \mu} \tag{16}$$

2. if $\exists w'' \in d \setminus q; Sim(w, w'') \neq 0$:

$$P_\mu(w|d_{ext}) = \frac{c(w";d).Sim(w, w") + \mu P(w"|C)}{|d_{ext}| + \mu} \tag{17}$$

with $w'' = argmax_{w' \in d, w' \neq w} Sim(w, w')$.

3. if $\nexists w'' \in d \setminus q; Sim(w, w'') \neq 0$

$$P_\mu(w|d_{ext}) = \frac{c(w;d) + \mu P(w|C)}{|d_{ext}| + \mu} \tag{18}$$

with $w'' = argmax_{w' \in d, w' \neq w} Sim(w, w')$.

In the specific case when all the query terms from q occur in the document d, the first case in the above is used where $|d_{ext}| = |d|$ leads to $p_\mu(w|d) = p_\mu(w|d_{ext})$.

3.2 Term Similarity Matrix Based on Statistical Approaches

In this section, we propose an easier and a more efficient way to compute the Term Similarity Matrix based on statistical approach which can have a better coverage.

Similarity between terms can be represented in a variety ways. In our approach, we used Confidence Coefficient, Tanimoto Similarity, Dice Coefficient, Cosine Similarity and Overlap Coefficient to generate the statistical information [11][6]. The Confidence Coefficient between term w_i and w_j are calculated as follows:

$$Sim_{conf}(w_i, w_j) = \frac{n(w_i \cap w_j)}{n(w_i)} or \frac{n(w_i \cap w_j)}{n(w_j)} \tag{19}$$

where $n(w_i)$ is the number of term(w_i) in the corpus, and $n(w_i \cap w_j)$ is the number of terms that term w_i co-occur together with w_j in the corpus.

The Tanimoto Similarity between term w_i and w_j are calculated as follows:

$$Sim_{tani}(w_i, w_j) = \frac{n(w_i \cap w_j)}{n(w_i) + n(w_j) - n(w_i \cap w_j)} \tag{20}$$

where $n(w_i)$ is the number of term(w_i) in the corpus, and $n(w_i \cap w_j)$ is the number of terms that term w_i co-occur together with w_j in the corpus.

The Dice Coefficient [6] between term w_i and w_j are calculated as follows:

$$Sim_{dice}(w_i, w_j) = \frac{2n|w_i \cap w_j|}{n(w_i) + n(w_j)} \tag{21}$$

where $n(w_i)$ is the number of term(w_i) in the corpus, and $n(w_i \cap w_j)$ is the number of terms that term w_i co-occur together with w_j in the corpus.

The Cosine Similarity between term X and Y is represented using a dot product and magnitude as follows:

$$Sim_{cosine}(w_i, w_j) = \sqrt{\frac{n(w_i \cap w_j)}{n(w_i).n(w_j)}} \tag{22}$$

where $n(w_i)$ is the number of term(w_i) in the corpus, and $n(w_i \cap w_j)$ is the number of terms that term w_i co-occur together with w_j in the corpus.

The Overlap Coefficient between term X and Y are calculated as follows:

$$Sim_{over}(w_i, w_j) = \frac{n(w_i \cap w_j)}{min(n(w_i), n(w_j))} \tag{23}$$

where $n(w_i)$ is the number of term(w_i) in the corpus, and $n(w_i \cap w_j)$ is the number of terms that term w_i co-occur together with w_j in the corpus.

4 Experiments

4.1 Data Set

We use the CHiC 2012 to test our proposed idea. CHiC 2012 contains fifty queries and one million documents. The proposed model is a generic solution to all application domains. However, CHiC 2012 was chosen as our test collection because the proposed model is more dedicated to the

subject of heritage. By using CHiC, the proposed model returns the best results and thus it could be the baseline benchmark when thus generic model is applied to another application domains.

In this corpus, the metadata inside the documents is quite variable from large to limited data. We use external resources such as Wikipedia to generate the Term Similarity Matrix. Our idea is to compute all the terms which co-occur in the Wikipedia. We use the English Wikipedia (version 2012-01-01) which contains 3.835 million articles in the corpus. For this paper, we only use the first paragraph of each article from the Wikipedia to generate the Term Similarity Matrix. Basically, the first paragraph of each article in the Wikipedia pertains the most critical idea of an article and it can stand on it owns as a concise version of this article according to the guideline from Wikipedia. In the experiments, we only use the title without any description from the queries. As for pre-processing, we remove all the stop words which contains 571 words and non-character, and apply the Porter Stemming. Besides, we convert all the upper case to lower case in order to reduce the term dimension.

All the experiments are done by using the XIOTA engine [4]. The performance is measured by Mean Average Precision (MAP). The optimal value for Dirichlet prior smoothing for baseline is 100 and 350 for all the Extended Dirichlet. Besides, we applied student's paired t-test (at the $p < 0.06$) to assess the significance of the difference measurement between the several types of statistic approach.

Table 1, shown clearly that our approach outperforms the baseline result. The most statistical significant improvement is with the Extended Dirichlet and Dice Coefficient from 0.5273 to 0.5450 at Table 1 while the most depreciation is with the Extended Dirichlet with Overlap Coefficient. The reason for these bad result for (Extended Dirichlet with Overlap Coefficient) is that most of the non-null values of the similarity matrix equal "1" which is abnormal because the value of "1" should represent exact match. Overall, 16 queries show increments, 8 queries show fluctuations and 11 queries remain the same as shown in Figure 1. The most significant increment is in Query 25 which increase around $+2008\%$ from 0.0025 to 0.0547 in terms of Average Precision(AP). The most significant decrements is in Query 28 which decrease around -13.33% from 0.0015(AP) to 0.0013(AP). We may notice that these extreme variation occur at rather low values of AP.

TABLE 1 – Performance with Various Types of Statistic from the First Paragraph of the Articles from Wikipedia (* = statistical significance at $p < 0.06$ using the Student's Paired T-Test)

Types of Approaches	MAP	MAP Gain/Lost
LM + Dirichlet (BL)	0.5273	
LM + Extended Dirichlet + Confidence Coefficient	0.5196	-1.48%
LM + Extended Dirichlet + Tanimoto Similarity	0.5395	+2.31%
LM + Extended Dirichlet + Dice Coefficient	**0.5451***	**+3.38**%
LM + Extended Dirichlet + Cosine Similarity	0.5418	+2.75%
LM + Extended Dirichlet + Overlap Coefficient	0.4929	-6.97%

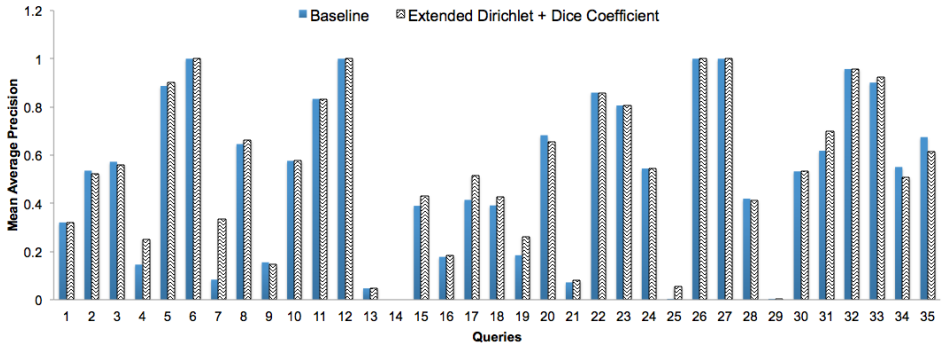


FIGURE 1 – Comparison between the baseline and Dice Coefficient result.

5 Conclusion and Future Work

We have presented a model to exploit the term similarity of non-matching terms during the retrieval time. Our experiments result indicate that the propose approach which is Term Similarity Matrix based on the statistical approach is more efficient and effective than the term intersection approach. For future work, we would like to compute more Term Similarity Matrix from other external resources and not only limited to Wikipedia. If we have more Term Similarity Matrix from different resources means we have higher degree of knowledge to build the link between two different terms.

Références

[1] Gianni Amati and Cornelis Joost van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transaction on Information Systems*, 20(4) :357–389, October 2002.

[2] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 222–229, New York, NY, USA, 1999. ACM.

[3] Guihong Cao, Jian-Yun Nie, and Jing Bai. Integrating word relationships into language models. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 298–305, New York, NY, USA, 2005. ACM.

[4] Jean-Pierre Chevallet. X-iota : An open xml framework for ir experimentation. In SungHyon Myaeng, Ming Zhou, Kam-Fai Wong, and Hong-Jiang Zhang, editors, *Information Retrieval Technology*, volume 3411 of *Lecture Notes in Computer Science*, pages 263–280. Springer Berlin Heidelberg, 2005.

[5] Fabio Crestani. Exploiting the similarity of non-matching terms at retrieval time. *Journal of Information Retrieval*, 2 :25–45, 2000.

[6] William B. Frakes and Ricardo Baeza-Yates. *Information Retrieval : Data Structures and Algorithms*. Prentice Hall PTR, June 1992.

[7] Maryam Karimzadehgan and ChengXiang Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 323–330, New York, NY, USA, 2010. ACM.

[8] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[9] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Wordnet : An on-line lexical database. *International Journal of Lexicography*, 3 :235–244, 1990.

[10] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. pages 275–281, 1998.

[11] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[12] S. E. Robertson. Overview of the okapi projects. *Journal of Documentation*, 53(1) :3–7, 1997.

[13] Gerard Salton. The smart project in automatic document retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 356 – 358, Chicago,

Illinois, United States, 1991.

[14] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523, 1988.

[15] W. Tannebaum and A. Rauber. Acquiring lexical knowledge from query logs for query expansion in patent searching. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 336–338, 2012.

[16] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, 1996.

[17] ChengXiang Zhai. *Statistical Language Models for Information Retrieval*. Now Publishers Inc., Hanover, MA, USA, 2008.

[18] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval.

Une approche basée sur des relations pour la RI sémantique

Marie-Noëlle Bessagnet[1], Davide Buscaldi[2], Albert Royer[1], Christian Sallaberry[1]

[1] LIUPPA, Université de Pau et des Pays de l'Adour,
F-64000 Pau
{marie-noelle.bessagnet,albert.royer,christian.sallaberry}@univ-pau.fr
[2] LIPN, Université Paris XIII,
F-93430 Villetaneuse
davide.buscaldi@lipn.univ-paris13.fr

Résumé :
Dans cet article, nous comparons trois méthodes de RI basées sur des mots-clés, sur des concepts et, pour la dernière, sur des concepts et des relations sémantiques entre concepts. La dernière méthode met en œuvre un algorithme de calcul de similarité conceptuelle implémenté par un prototype. Notre évaluation démontre que cette méthode améliore la précision par rapport à une RI basée uniquement sur des concepts.
Mots-clés : recherche d'information, ontologie, relation sémantique, mesure de similarité

1 Introduction

Les ontologies et les annotations sémantiques qu'elles permettent ont contribué fortement au développement récent de moteurs de recherche plus efficaces. On peut citer les systèmes KIM [16], Hakia [1] ou encore tous les systèmes cités dans [18].

En accord avec [7], le but de ces systèmes est d'améliorer les performances en dépassant les modèles de recherche basés sur les mots clés.

Pour améliorer les performances des systèmes de recherche d'information (**RI**) sur les documents, la prise en compte de la sémantique des termes

1. http://www.hakia.com

est privilégiée. Ainsi, l'indexation se situe au niveau des concepts (les sens des mots) et permet de mieux décrire le contenu du document et le contenu de la requête. Dans ces approches, des ressources sémantiques telles que des thésaurus ou des ontologies sont utilisées dans les phases d'indexation et de recherche. L'idée d'intégrer des ressources sémantiques dans la recherche d'information a été développée parallèlement à l'évolution du web sémantique. Les premiers travaux proposant une architecture pour la RI sémantique [11] utilisent des ontologies pour l'annotation et l'indexation sémantique des textes dans le contexte du web sémantique.

Cependant, la RI sémantique est plus efficace dans des domaines limités, étant donné les problèmes de production de la ressource sémantique associée.

Les expériences positives ont toujours été menées dans des domaines limités comme le domaine médical [1] et [24] ou le domaine de la biologie [14].

La recherche d'information sémantique (**RIS**) s'appuie donc sur une ontologie du domaine et sur un corpus reflétant ce domaine. Elle est basée sur un processus d'indexation des concepts des documents et de la requête utilisateur, sur un processus d'appariement des documents et de la requête pour fournir la meilleure réponse. Le plus souvent, elle utilise également des formules de calcul de similarité qui exploitent la relation hiérarchique classique. Nous souhaitons étendre cette exploitation des relations structurelles et intégrer dans un système de RIS des relations sémantiques.

Nos travaux contribuent au domaine sur deux points : (1) le processus d'annotation est basé sur une ressource termino-ontologique (cf. 2.1) ; et (2) un algorithme de ranking prend en compte non seulement les concepts mais également les relations sémantiques (cf. 3.3.2).

Dans cet article, nous exposons notre système de RIS et une évaluation de ce dernier. Ainsi, dans la section 2, nous présentons des travaux connexes. Dans la section 3, nous décrivons notre approche pour la recherche d'information sémantique. Dans la section 4, nous détaillons la mise en œuvre de cette proposition dans le cadre du projet MOANO [2]. Dans la section 5, nous rendons compte de l'expérimentation et de l'évaluation menées. Enfin, dans la section 6, nous concluons et nous proposons quelques perspectives.

2. `http://moano.liuppa.univ-pau.fr/`

2 Annotation, indexation et RI sémantique

2.1 Définitions

Aujourd'hui de nombreux systèmes annotent, indexent et supportent la RIS [11] : ces trois processus sont qualifiés de sémantique car cette recherche utilise une ressource termino-ontologique ou **RTO**. La notion de RTO apparaît pour la première fois dans le rapport RTP-DOC [15].

Le but est de modéliser le contenu de documents sélectionnés ou formant une collection sous la forme de réseaux de termes afin d'améliorer l'accès à la connaissance. Ces modèles ou représentations sont connus : des thésaurus, des terminologies, des langages documentaires, des index ou encore des ontologies. Ils sont généralement regroupés sous la notion de ressources terminologiques ou ontologiques [2]. [12] en propose une première définition « *Ressource informatique décrivant le vocabulaire et les concepts spécifiques à un domaine, à une communauté pour le traitement de l'information* ». Cette notion est ensuite concrétisée par les travaux de [21]. Enfin, les travaux récents de Cimiano et al.[3], de McCrae et al.[13], et de Roche et al.[22] ont proposé d'associer une partie terminologique et/ou linguistique aux ontologies afin d'établir une distinction claire entre la composante terminologique et la composante conceptuelle. Les RTO [22] sont utilisées dans plusieurs tâches liées à l'ingénierie des connaissances, par exemple, pour l'étiquetage sémantique de corpus, pour l'indexation, pour la recherche d'information ou encore pour la navigation.

Nous distinguons deux types de travaux de recherche en RIS, même si les techniques sont identiques : les systèmes prenant en compte un corpus documentaire spécialisé (notre cas) et les travaux qui s'intéressent à des collections très larges de documents (génériques) comme [11].

Concernant l'**annotation sémantique**, nous pourrions la définir comme le processus qui fixe l'interprétation d'un document en lui associant une sémantique formelle et explicite [11].

Une fois l'annotation effectuée, deux classes d'usages peuvent en découler : annoter pour extraire des connaissances ou annoter pour indexer. Nous nous plaçons dans ce deuxième usage visant l'**indexation sémantique**. Elle permet d'établir une nouvelle représentation de documents d'un corpus à partir des concepts ainsi annotés.

Concernant la **recherche sémantique**, en accord avec [25], il n'existe pas de modèle partagé. Ainsi, plusieurs définitions peuvent être trouvées concernant la RIS. Nous retiendrons celle de [5] qui proposent « ... *La RIS vise à mieux satisfaire les besoins en information des utilisateurs en*

exploitant le sens des termes utilisés. La RIS repose pour cela sur un processus d'indexation destiné à obtenir une représentation sémantique des documents et des requêtes . . . » La RIS cherche à dépasser les limites d'une recherche classique par mots clés.

2.2 Mesures de similarité sémantique

L'utilisation de ressources sémantiques, lors d'une recherche, permet de retrouver les documents qui partagent le maximum de concepts avec la requête. Dans ce cadre, des travaux portent sur le calcul de la similarité sémantique.

Aussi, l'évaluation de la similarité sémantique entre concepts est un problème connu dans le domaine de la RI. Plusieurs méthodes ont été proposées dans ce sens. On peut les classer selon trois catégories :

1. les approches basées sur la distance, c'est-à-dire sur la structure de l'ontologie, que l'on appelle mesures structurelles. Elles sont fondées sur l'analyse et l'exploitation de la structure sémantique des graphes conceptuels où les nœuds représentent les concepts et les arcs représentent la relation *is-a*. D'une manière générale, la distance est caractérisée par le plus court chemin qui fait intervenir un ancêtre commun, le plus petit généralisant, connectant potentiellement deux objets. Parmi les travaux qui implémentent ces mesures on peut citer par exemple : Rada et al. [17], Resnik [19], Wu et Palmer [27] ou encore Dudognon et al [5].
2. les approches utilisant le contenu informatif des concepts, que l'on appelle mesures conceptuelles. Elles stipulent que la distance entre deux concepts est une fonction des instances communes entre eux. La plupart de ces mesures sont fondées sur la notion de contenu informationnel d'un concept, introduite par Resnik [20]. Selon cette approche, le contenu informationnel traduit la pertinence d'un concept en tenant compte de la fréquence de son apparition dans la collection ainsi que de la fréquence d'apparition des concepts qu'il subsume. Ces mesures conceptuelles sont détaillées dans [4].
3. les approches dites hybrides combinent les approches basées sur les arcs et celles basées sur le contenu informationnel qui est considéré comme facteur de décision. On peut citer comme exemple la formule de Jiang et Conrath[10].

Ces mesures sont intégrées dans différentes applications, telles que le

calcul de similarité entre documents, le clustering de documents, la désambiguïsation sémantique, l'indexation, etc.

Les relations hiérarchiques classiques, *part-of*, *is-a* ne suffisent pas à exprimer la sémantique contenue dans les documents et les requêtes. Aussi, il faut modéliser des relations sémantiques et trouver des méthodes permettant d'évaluer la similarité sémantique. C'est ce que nous proposons dans la section suivante.

3 Identification des concepts et des relations sémantiques

Dans cette partie, nous allons présenter notre démarche pour l'annotation de concepts et de relations dans un document ou dans une requête ainsi que l'appariement de ces derniers. La méthode présentée a pour origine les travaux menés dans le groupe MELODI de l'IRIT [3] dans le cadre du projet ANR DYNAMO [6]. Nous avons repris et étendu ces travaux de manière à prendre en compte l'annotation des relations que nous allons détailler.

3.1 Notation

Nous adoptons les notations suivantes pour la formalisation du processus d'annotation de concepts et de relations, d'une part, et du processus d'appariement qui exploite des annotations, d'autre part.

c	pour un concept de l'ontologie,
d	pour un document du corpus,
f	pour un champ (titre, section, paragraphe) d'un document,
r	pour une relation sémantique,
t	pour un terme rencontré dans un document.

3.2 Formalisation des différentes notions

3.2.1 Ontologie, RTO, concepts, relations sémantiques et corpus

Les concepts de notre ontologie sont des classes d'un domaine spécifique ; par exemple, pour le domaine botanique, le concept *gladiolus* est une classe dont le végétal *glaïeul de Colville* est une sous-classe. En plus des relations hiérarchiques classiques comme *part-of* ou *is-a*, sont modélisées des relations sémantiques ; ainsi, nous pourrons noter que le *glaïeul de Colville* se *plante* en *octobre-novembre*.

3. Institut de Recherche en Informatique de Toulouse

C'est pourquoi nous définissons une **ontologie** O par l'ensemble C des concepts du domaine et par l'ensemble R des relations entre concepts et nous écrivons $O = (C, R)$. On note $R = \{r_\nu\}$ avec $r_\nu = (\delta, \nu, \rho)$ où la relation r_ν de nom ν a pour domaine de classe δ et pour co-domaine de classe ρ.

Dans une **RTO**, à chaque concept est associée une liste de termes qui *dénotent* le concept ; on note T l'ensemble des mots pouvant dénoter un concept (c'est-à-dire, des termes dont la présence dans un texte implique automatiquement la présence du concept dénoté). Rappelons que la rédaction de la liste associée à chaque concept de l'ontologie est du ressort d'un spécialiste du domaine.

Sur les **concepts**, plusieurs prédicats sont nécessaires :

$subsumes(c_i, c_j)$,	où $c_i \in C$ et $c_j \in C$, est interprétée c_i subsume c_j ;
$has\_label\_c(c, t)$	pour $c \in C$ et $t \in T$ indiquant que c a pour label le terme t.

Pour les **relations**, les prédicats sont :

$has\_label\_r(r, t)$	pour $r \in R$ et $t \in T$, indique que r a pour label le terme t ;
$relation(c_\delta, t, c_\rho)$	pour $c_\delta \in C, t \in T$ et $c_\rho \in C$ indique une relation révélée par t entre c_δ et c_ρ ;

Le **corpus** est vu comme un ensemble de documents D. Chaque document est composé de plusieurs champs. L'ensemble des n champs d'un document $d \in D$ sera noté $F_d = \{f_0, \ldots, f_n\}$. On note avec $P(F_d) = f_p$ le champ porteur du concept *pivot* c_p pour un document. Le concept pivot correspond à la classe qui caractérise le sujet principal du document (sous l'hypothèse que le document est écrit en style encyclopédique) ; par exemple, la classe *Plante* pour le domaine botanique.

3.2.2 Annotation de concepts

On note $T_{f,d}$ l'ensemble des termes présents dans le champ f du document d. L'annotation se fait par champ. On note $A_C(f) = \{c_0, \ldots, c_m\}$ l'annotation d'un champ f avec les concepts $c_0, \ldots, c_m$.

Nous définissons le prédicat suivant :

$holds\_c(f, c) \iff \exists t \in T_{f,d} \mid has\_label\_c(c', t) \ \wedge \ subsumes(c, c')$

Ainsi, un champ f contient un concept c si et seulement si un terme t dénotant un concept c' existe et si le concept c' est un descendant de c ou si $c = c'$.

3.2.3 Annotation de relations

On note $A_R(f) = \{r_0, \ldots, r_n\}$ l'annotation d'un champ f avec les relations $r_0, \ldots, r_n$.

Nous définissons le prédicat suivant :

$$
\begin{aligned}
holds\_r(f, r) \iff & \exists\, t_r, t_\delta, t_\rho \in T_{f,d} \mid \\
& \quad has\_label\_r(r, t_r) \ \wedge \ relation(c_\delta, r, c_\rho) \\
& \wedge has\_label\_c(c'_\delta, t_\delta) \ \wedge \ subsumes(c_\delta, c'_\delta) \\
& \wedge has\_label\_c(c'_\rho, t_\rho) \ \wedge \ subsumes(c_\rho, c'_\rho) \\
& \bigvee \\
& \exists\, t'_r, t'_\rho \in T_{f,d} \mid \\
& \quad has\_label\_r(r, t'_r) \wedge relation(c_p, r, c_\rho) \\
& \wedge has\_label\_c(c'_\rho, t'_\rho) \wedge subsumes(c_\rho, c'_\rho)
\end{aligned}
$$

Ainsi, un champ f contient une relation r dans deux cas :

- soit que trois termes du champ dénotent, l'un la relation et les deux autres les concepts du domaine c_δ et du co-domaine c_ρ correspondants à cette relation,
- soit que deux termes du champ dénotent la relation et le concept du co-domaine c_ρ correspondant à cette dernière ; le concept du domaine de la relation étant le concept pivot.

3.3 Appariement

3.3.1 Appariement basé sur la similarité de concepts

Soit Q l'ensemble des concepts dans une requête et D l'ensemble des concepts dans un document, et soit $F(C)$ la fonction qui donne le coefficient de dominance d'un concept C (les coefficients sont donnés par l'usager et sauvegardés dans un fichier de configuration), le poids pour un document est calculé de la façon suivante :

$$
w(Q, D) = \frac{\sum_{c_1 \in Q} (F(c_1). \max_{c_2 \in D} s(c_1, c_2))}{\sum_{c_1 \in Q} F(c_1)} \tag{1}
$$

où $s(c_1, c_2)$ est la mesure de similarité.

Cette mesure correspond à une mesure classique de similarité de concepts détaillée dans [5].

3.3.2 Appariement basé sur la similarité de concepts et de relations

Nous prenons en compte les relations pour étendre cette approche fondée sur l'appariement des concepts.

Soit R_Q l'ensemble des relations $r_1, \ldots, r_k$ trouvées dans une requête avec comme ensemble de concepts Q. Chaque relation est un triplet $r = (c_\delta, \nu, c_\rho)$ où c_δ est le concept relatif au domaine, ν le nom de relation et c_ρ le concept relatif au co-domaine. On définit $d(r) = c_\delta$, $n(r) = \nu$ et $e(r) = c_\rho$. Deux relations r_1, r_2 sont comparables uniquement si $n(r_1) = n(r_2)$. On définit l'ensemble R_D comme l'ensemble des relations trouvées dans le document D. Le poids d'un document est calculé comme $w(Q, D) + b(R_Q, R_D)$, où :

$$b(R_Q, R_D) = \frac{\sum\limits_{\substack{r_1 \in R_Q, r_2 \in R_D \\ \text{et } n(r_1)=n(r_2)}} (F(d(r_1)).s(d(r_1), d(r_2)) + F(e(r_1)).s(e(r_1), e(r_2)))}{\sum\limits_{\substack{r_1 \in R_Q, r_2 \in R_D \\ \text{et } n(r_1)=n(r_2)}} (F(d(r_1)) + F(e(r_1)))} \quad (2)$$

De manière transparente pour l'utilisateur, nous qualifions $b(R_Q, R_D)$ de *boost* qui augmente le poids d'un document D qui contient tout ou partie des relations détectées dans R_Q et, par conséquent, replace de tels documents en début de la liste résultat. Ici, pour chaque couple de relations $r_1 \in R_Q, r_2 \in R_D$ de même nom, nous calculons la similarité des concepts relatifs aux domaines et aux co-domaines de ces relations, respectivement. La somme de ces mesures de similarité, normalisée par les coefficients de dominance correspondants, détermine le *boost*.

Nous allons décrire, dans la section suivante, la mise en œuvre de notre méthode implémentée dans le cadre du projet MOANO.

4 Mise en œuvre dans le cadre de MOANO

Notre processus de RIS prend en compte :

– la collection de documents (fiches xml) et la requête de l'utilisateur (expression de son besoin),
– les différentes opérations d'annotation, d'indexation et d'appariement dont le but est la sélection des documents à présenter à l'utilisateur.

À travers l'étape d'indexation (1), le système organise la collection de documents sous la forme d'une représentation sémantique (concepts et relations). L'interrogation du fonds documentaire à l'aide d'une requête nécessite également la représentation de celle-ci sous une forme compatible (concept et relation) avec les documents (2). L'appariement requête-document (3) permet de sélectionner la liste des documents en s'intéressant à la similarité des concepts. Cette liste de documents est réordonnée selon le boost donné par les relations dans une deuxième étape d'appariement (3bis) qui tient compte de la similarité des relations. Ainsi les documents sont proposés à l'utilisateur par ordre de pertinence.

Dans le cadre du projet, le prototype développé annote et indexe les concepts et les relations. Nous avons mené une expérimentation décrite ci-après.

5 Expérimentation

Afin d'évaluer le système de RIS défini dans le cadre du projet MOANO, nous avons mis en place une expérimentation basée sur les deux versions d'appariement proposées : $ThemaSTream_1$ (similarité des concepts) et $ThemaStream_2$ (similarité des concepts et des relations).

5.1 Cadre d'évaluation de systèmes de RI sémantique

Nous nous sommes inspirés des travaux de [25], des campagnes TREC [26], ainsi que de [8] pour la préparation du protocole d'analyse et de la collection de test.

La tâche évaluée est une recherche qualifiée de *ad-hoc* dans TREC : le système de RI (**SRI**) répond à un besoin d'information par une liste de documents ordonnée par pertinence décroissante. L'évaluation vise à mesurer l'efficacité relative des SRI suivants :

– $Lucene$, SRI classique basée sur les mots clés ;
– $ThemaSTream_1$, SRI thématique basée sur les concepts ;
– $ThemaStream_2$, SRI thématique basée sur les concepts et renforcée par les relations.

Pour un *topic* donné, chaque SRI fournit une liste de couples (d, s) représentant le score s de chaque document d restitué. Classiquement, l'efficac-

ité d'un SRI est évaluée grâce aux mesures *Average Precision* (**AP**) pour chaque *topic* et *Mean Average Precision* (**MAP**) globalement. Pour faciliter la phase de jugement de pertinence des documents, nous avons choisi les métriques *Mean Relevance Rank* (**MRR**) et *Precision à 10* (P@10) qui, selon [23], correspondent à des mesures comparables de la qualité des réponses d'un SRI.
À l'image du protocole d'expérimentation de TREC, nous proposons deux niveaux de granularité d'évaluation d'un SRI : le premier niveau *topic* en calculant P@5, P@10 et MRR et le second niveau global en calculant la moyenne arithmétique des n valeurs de P@5, P@10 et de MRR, fournissant ainsi la mesure globale de performance du SRI.
À chaque niveau, les n différences observées $\langle m_i^1 - m_j^1, \ldots, m_i^n - m_j^n \rangle$ sont rapportées en pourcentage (d'amélioration ou de détérioration), où m_s^t représente la valeur de la mesure m obtenue par le système s pour le *topic* t. La significativité des tests statistiques calculée pour les différences observées est également rapportée : les p-valeurs de significativité sont calculées avec le test t de *Student* apparié (la différence est calculée entre les paires de valeurs m_i^t et m_j^t). Lorsque $p < \alpha$ avec $\alpha = 0{,}05$ la différence entre les deux échantillons testés est qualifiée de statistiquement significative [9].

Notre collection de test est composée :

- de 25 « topics » (besoin d'information) issus de questions posées sur le site *Yahoo Answers* décrivant des besoins d'information dans le domaine de la botanique. Les questions ont été sélectionnées de manière à ce qu'un concept et une relation soient présents dans notre ontologie (cf. 5.2.1) ;
- d'un « corpus » comprenant un échantillon de 1 000 fiches-plante du guide Clause Vilmorin dont certaines sont pertinentes pour les *topics* proposés ;
- de « qrels » (jugements de pertinence) désignant, pour chacun des 25 *topics*, l'ensemble des documents pertinents du corpus. Nous nous sommes ici limités à l'évaluation des dix premiers résultats restitués par chaque SRI pour chacun des topics ;
- de ressources ontologiques, décrivant un point de vue relatif au domaine de la botanique sous forme de concepts et de relations sémantiques.

5.2 Expérimentation et évaluation des prototypes ThemaStream

Notre objectif est ici de comparer le SRI $Lucene$ (système de référence) avec les SRI sémantiques $ThemaStream_1$ et $ThemaSream_2$.

5.2.1 Observation quantitative des résultats

Les résultats de la figure 1 confirment l'hypothèse selon laquelle, quand, dans les ressources ontologiques mobilisées, il existe des concepts et des relations spécifiant le contenu du besoin exprimé, une approche de RI sémantique permet d'obtenir de meilleurs résultats qu'une approche classique de RI basée sur des mots-clés.

	Moyenne arithmétique			
25 topics	P@5	P@10	MRR	nombre de résultats
Lucene	0,43	0,46	0,56	405
ThemaStream$_1$	0,53	0,58	0,65	910
Gain//Lucene	23,26%	26,09%	16,07%	
ThemaStream$_2$	0,74	0,74	0,83	910
Gain//Lucene	72,09%	60,87%	48,21%	
Gain//ThemaStream$_1$	39,62%	27,59%	27,69%	

FIGURE 1 – Analyse globale des résultats.

En effet, $ThemaStream_1$ donne de meilleurs résultats que $Lucene$ et $ThemaStream_2$ donne de meilleurs résultats que $ThemaStream_1$ et que $Lucene$. Dans le cas de la comparaison de $ThemaStream_2$ avec $Lucene$ le test t de *Student* apparié donne $p = 0{,}001$ pour P@10 et $p = 0{,}005$ pour MRR, ce qui valide la significativité des résultats obtenus.

Une analyse par topic montre une forte variabilité des résultats, la première fiche pertinente étant mieux classée par $ThemaStream_2$ (ou ex æquo) dans vingt-deux cas sur vingt-cinq et $ThemaStream_2$ assurant une meilleure précision à 10 (ou ex æquo) dans vingt-trois cas sur vingt-cinq. L'observation du nombre de documents pertinents distincts, sur les dix

premiers restitués respectivement par $Lucene$ et $ThemaStream_2$, montre la complémentarité potentielle des deux systèmes. En moyenne, seul un document pertinent sur les dix premiers restitués est commun à $Lucene$ et à $ThemaStream_2$, tous les autres étant distincts. De plus, dans certains cas $Lucene$ pourrait améliorer le MRR correspondant aux résultats de $ThemaStream_2$. Notons toutefois que la non exhaustivité de la taxonomie décrite dans l'ontologie peut être un biais à cette analyse de complémentarité.

5.2.2 Observation qualitative des résultats

Il est nécessaire de rappeler que, pour chacun des 25 *topics* expérimentés, il existe au moins un concept et une relation correspondante spécifiés dans notre ontologie botanique. La présence de plusieurs relations améliore systématiquement la qualité des résultats. La seule exception constatée correspond à la présence d'une négation dans un document restitué.

6 Conclusion et perspectives

Nous avons relaté dans cet article les travaux qui contribuent au domaine de la RIS sur deux points : (1) le processus d'annotation est basé sur une ressource termino-ontologique ; et (2) un algorithme de ranking prend en compte non seulement les concepts mais également les relations sémantiques.

De nombreux autres systèmes de RIS utilisent une ontologie et travaillent sur des collections très larges de documents. Nous nous intéressons à un corpus documentaire spécialisé et donc à une ressource terminologique dédiée. Nous avons inclus dans cette ressource des relations sémantiques pour obtenir une annotation plus fine.

L'algorithme de classement des documents développé dans notre projet tient compte de ces relations sémantiques par le biais d'un calcul de boost. Ainsi, les documents sont classés de manière plus précise.

Cependant notre expérimentation présente certaines limites. En effet, la non exhaustivité de l'ontologie qui ne décrit pas toutes les plantes du corpus, peut conduire à un biais dans les résultats entre la recherche classique et la recherche sémantique. Ainsi, Lucene, par les mots clés, annote tous les documents alors que notre SRI annote les documents avec les seuls taxons existants dans l'ontologie. Aussi, pour lever cette ambiguïté, il faudrait mettre en place deux scénarios d'évaluation :

– une mesure en prenant comme hypothèse que l'ontologie décrit toutes les plantes du corpus (scénario utilisé dans ce papier),
– une mesure en éliminant les documents annotés par Lucene et non annotés par nos systèmes du fait de l'absence de taxons dans l'ontologie.

Nos perspectives sont d'améliorer les prototypes $ThemaStream$ de manière à permettre à un utilisateur de bâtir des requêtes plus expressives et d'obtenir ainsi des résultats encore plus fins.

Remerciements

Cette recherche a été réalisée dans le cadre du projet Moano « Modèles et Outils pour Applications NOmades de découverte de territoire » (`http://moano.liuppa.univ-pau.fr/`), en partie financé par l'Agence Nationale de la Recherche (ANR-2010-CORD-024-01).

Références

[1] ABASOLO J. M. & GOMEZ M. (2000). Melisa. an ontology-based agent for information retrieval in medicine. In *In : Proceedings of the First International Workshop on the Semantic Web (SemWeb2000*, p. 73–82.

[2] AUSSENAC-GILLES N. & CONDAMINES A. (2004). Documents électroniques et constitution de ressources terminologiques ou ontologiques. *Information - Interaction - Intelligence*, **4**(1). Article publié dans le numéro thématique de la revue i3 consacré au document numérique.

[3] CIMIANO P., BUITELAAR P., MCCRAE J. & SINTEK M. (2011). Lexinfo : A declarative model for the lexicon-ontology interface. *Web Semant.*, **9**(1), 29–51.

[4] CORLEY C. & MIHALCEA R. (2005). Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, p. 13–18, Stroudsburg, PA, USA : Association for Computational Linguistics.

[5] DUDOGNON D., HUBERT G., MARCO J., MOTHE J., RALALASON B., THOMAS J., REYMONET A., MAUREL H., MBARKI M., LAUBLET P. & ROUX V. (2010a). Dynamic ontology for information retrieval. In *RIAO*, p. 213–215 : CID - Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

[6] DUDOGNON D., HUBERT G. & RALALASON B. J. V. (2010b). ProxiGénéa : Une mesure de similarité conceptuelle (regular paper). In *Colloque Veille Stratégique Scientifique et Technologique (VSST), Toulouse,*

25/10/2010-29/10/2010, p. (support électronique), http ://www.ups-tlse.fr : Université Paul Sabatier - Toulouse.

[7] FERNÁNDEZ M., CANTADOR I., LOPEZ V., VALLET D., CASTELLS P. & MOTTA E. (2011). Semantically enhanced information retrieval : An ontology-based approach. *J. Web Sem.*, **9**(4), 434–452.

[8] HARMAN D. K. (2005). The TREC Test Collections. In [26], chapter 2, p. 21–53.

[9] HULL D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In *SIGIR'93 : Proceedings of the 16th annual international ACM SIGIR conference*, p. 329–338, New York, NY, USA : ACM Press.

[10] JIANG J. & CONRATH D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, p. 19–33.

[11] KIRYAKOV A., POPOV B., TERZIEV I., MANOV D. & OGNYANOFF D. (2004). Semantic annotation, indexing, and retrieval. *J. Web Sem.*, **2**(1), 49–79.

[12] LORTAL G. (2006). Annotations dans les activités coopératives : élaboration d'un modèle générique multi-points de vue et utilisation des technologies du web sémantique pour sa mise en œuvre. *Doctorat en informatique, Université de Technologie de Troyes*.

[13] MCCRAE J., SPOHR D. & CIMIANO P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web : research and applications - Volume Part I*, ESWC'11, p. 245–259, Berlin, Heidelberg : Springer-Verlag.

[14] MÜLLER H.-M., KENNY E. E. & STERNBERG P. W. (2004). Textpresso : An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, **2**(11), e309.

[15] PÉDAUQUE R. & SALAÜN J. (2006). *Le document à la lumière du numérique*. C&F Editions.

[16] POPOV B., KIRYAKOV A., OGNYANOFF D., MANOV D. & KIRILOV A. (2004). Kim - a semantic platform for information extraction and retrieval. *Natural Language Engineering*, **10**(3-4), 375–392.

[17] RADA R., MILI H., BICKNELL E. & BLETTNER M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, **19**(1), 17–30.

[18] RAMÍREZ R. C. M. & R. V. M. R. (2007). A semantic web approach to enrich information retrieval answers. In *ICEIS (4)*, p. 299–302.

[19] RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, p. 448–453 : Morgan Kaufmann.

[20] RESNIK P. (1999). Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural lan-

guage. *J. Artif. Intell. Res. (JAIR)*, **11**, 95–130.

[21] REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2007). Modélisation de Ressources Termino-Ontologiques en OWL. In F. TRICHET, Ed., *Actes des Journées Francophones d'Ingénierie des Connaissances (IC 2007)*, p. 169–180, Grenoble, France : Cépaduès Editions.

[22] ROCHE C., CALBERG-CHALLOT M., DAMAS L. & ROUARD P. (2009). Ontoterminology : A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development*, p. 321–326, Madeira, Portugal.

[23] SANDERSON M., PARAMITA M. L., CLOUGH P. & KANOULAS E. (2010). Do user preferences and evaluation measures line up ? In *SIGIR*, p. 555–562.

[24] TRIESCHNIGG R., PEZIK P., LEE V., DE JONG F., KRAAIJ W. & REBHOLZ-SCHUHMANN D. (2009). Mesh up : effective mesh text classification for improved document retrieval. *Bioinformatics*, **25**(11), 1412–1418.

[25] UREN V. S., SABOU M., MOTTA E., FERNÁNDEZ M., LOPEZ V. & LEI Y. (2010). Reflections on five years of evaluating semantic search systems. *IJMSO*, **5**(2), 87–98.

[26] VOORHEES E. M. & HARMAN D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval*. Cambridge, MA, USA : MIT Press.

[27] WU Z. & PALMER M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, p. 133–138, Stroudsburg, PA, USA : Association for Computational Linguistics.

A Multi-level Dimension-based Semantic Query and Document Structuring

Mohannad ALMASRI[1] and Jean-Pierre CHEVALLET[2]

[1] Université Joseph Fourier - BP 53 38041 Grenoble Cedex 9
[2] Université Pierre-Mendès - BP 47 38040 Grenoble Cedex 9
{mohannad.almasri,jean-pierre.chevallet}@imag.fr

Résumé : Most Information retrieval systems represent a query, also a document, as a bag of indexing terms without any relation between each other. This representation causes a problem for specialists when they deal with a specific domain like medical one. This bag based representation may have some lack of precision. We present an alternative to the bag of indexing terms representation depending on semantic query structuring, in order to fulfill this need of precision in a specific domain. This structure of a query is obtained by grouping indexing terms using pre-defined categories called *dimensions*. These dimensions represent the different aspects that could appear in a query or a document. By using this notion, the relevant document to a given query should not only have a maximum number of shared indexing terms but also have a similar structure. Experimental results show precision improvement related to the granularity of dimensions and its distribution over the whole corpus.

Mots-clés : Semantic Query, Structured Query, Conceptual Indexing, Domain Ontology.

1 Introduction

Information Retrieval Systems (IRS) are important tools to help domain specialists to retrieve valuable information from huge quantities of available documents. Specialists of a domain, e.g. the medical domain, are the people who have a good knowledge about the related domain, and they are capable of building a precise or a well-structured queries, instead of simple bag of indexing terms [1] queries.

1. Indexing terms differ from system to another, so it can be : word, noun phrase, n-gram, or concept [5].

The main shortcoming of nowadays Web search engines and IRSs is the flat representation of queries and documents, or in other words, a bag of indexing terms representation. This representation exhibits some lack of precision for specialists when they deal with a specific domain like medical. As an example of a well-structured query in the medical domain, assume the fourth query in the ImageCLEF2011 [2] collection, q_4 is "chest CT images with emphysema". q_4 searches images satisfying the following properties : their modality is CT (Computerized Tomography), diagnose emphysema, and concern the chest. In other words, this query can be structured in three distinct parts : *modality* represented by "CT images", *pathology* represented by "emphysema" and *anatomy* represented by "chest". Anatomy, pathology and modality are called semantic categories or dimensions [7, 3, 12, 13]. The previous example shows that a simple bag of indexing terms (keywords in this case) query is not sufficient to express specialists' queries which have a clear structure. This type of query partitioning or structuring requires an external resource, e.g. a meta-thesaurus, a knowledge base, which can separate indexing terms over semantic categories or dimensions.

In this paper, we present a semantic query structuring framework as an alternative to the bag of indexing terms representation. This new framework aims to fulfill the need of precision in a specific domain like medical. In addition, it can be used in different domains. We also study the effect of dimension distribution within a corpus on the retrieval precision. The rest of this paper is organized as follows. We first present some related works in semantic query structuring in section 2. In section 3 we talk about conceptual indexing. In section 4, we present our framework for semantic query structuring. We report the experimental results in section 5 and conclude in section 6.

2 Semantic Query Structuring in Literature

Semantic query structuring is used for different purposes in information retrieval, like searching structured data, reformulating user queries, and entity search.

The notion of dimensions is proposed in order to navigate a base of images [7] or a base of textual documents [3]. This navigation is achieved using an interface based on an ontology. This ontology is divided into different hierarchies and each node in these hierarchies called dimension.

2. http ://www.imageclef.org/

Each dimension corresponds to a point of view according to which one can explore the base.

Li et al. [9] use semantic query structuring in order to search structured data. They tag each term in a query using pre-defined dimensions. Figure 1 shows an example of this tagging operation, where (Brand, Model, Type, Attribute) are examples of dimensions.

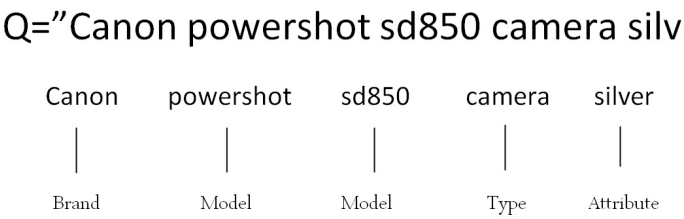


FIGURE 1 – Semantic query structuring example. Each query term (canon, powershot, sd850, camera, silver) is tagged by a dimension (Brand, Model, Type, Attribute).

The tagging operation is achieved using a semi-supervised learning method with Conditional Random Fields, and using two sources of knowledge and a small amount of manually-labeled queries.

Another example of semantic query structuring is to find multiple facets or aspects of a query [6]. These facets (called dimensions) are mined out from top results of a search engine for a given query. For example, the query “watches", using this method, has five dimensions : brands, gender categories, supporting features, styles, and colors. These dimensions are used to improve search experience in many ways : 1- help users to clarify their intention by reformulating his query, 2- improve the diversity of top results by re-ranking search results to avoid showing pages that are nearly duplicated in query dimensions, or 3- can be used in semantic search or entity search. This method is only suitable for HTML documents.

Radhouani et al.[12, 13], propose a model for semantic query structuring based on conceptual indexing. Basically, they represent documents and queries by means of concepts [3]. Then, they structure these concepts using dimensions. A dimension of a domain corresponds to a point of view according to which one can see this domain, e.g. Diseases in the medical

3. “Concepts" can be defined as “Human understandable unique abstract notions in dependent from any direct material support, independent from any language or information representation, and used to organize perception and knowledge" [5].

domain. The purpose of dimensions is to enhance the precision of information retrieval system using a domain knowledge.

In this section, we reviewed three works on semantic query structuring. These works aim, either to help searching in structured data [9], or to help users to rewrite their queries [6]. The works that talked about semantic query structuring to enhance the precision were succeeded to do that in a very special case without explaining their results. In addition, they used a version of vector space model in their evaluation which is outdated model in information retrieval [12, 13]. Our work presented in this paper belongs to the last work category. Our approach differs from previous works in four important points : first, it is a precision oriented approach. Second, it does not need user supervision or training data. Third, we propose a framework for query structuring with two ways for matching between a structured query and a document. Last, our experiments are made using up to date models in information retrieval and with studying the effect of dimensions distribution over the whole corpus.

3 Conceptual Indexing

Classical techniques for indexing represent documents and queries as a bag of words or phrases without taking into account the semantics, meaning or the correlation between these words . The main disadvantage of these techniques is that they depend on the text signal, and not on the meaning [5, 10]. For example, in the medical domain, the two phrases “Atrial Fibrillation" and “Auricular Fibrillation" have the same meaning. However, by using phrases to represent a document and a query, if one phrase appears in a document and another one appears in a query that leads to unmatched document and query. So over the last 20 years, several approaches attempted to use available knowledge bases and natural language processing techniques in order to overcome this problem and produce more meaningful answers [4]. These approaches represent documents and queries by means of concepts. This representation is obtained using conceptual indexing. Conceptual indexing is the process of mapping text into the concepts of an *external resource*. Therefore, it needs a resource out of documents and queries and containing concepts and information about them.

The purpose of conceptual indexing is to represent queries and documents by means of concepts instead of words or phrases. In our framework, queries and documents are represented by means of concepts. Therefore, we use conceptual indexing in order to obtain this concept-based represen-

tation. For a detailed information about conceptual indexing see [5].

4 Semantic Query Structuring Framework

In any IRS, there are three essential components : a query model, a document model, and a matching function. In our case, we use concepts for representing queries and documents, so we need an additional component, contains concepts, which is the external resource. This external resource not only helps our information retrieval system in the conceptual indexing process, but also helps it in the semantic query structuring process.

Semantic query structuring aims to build a structured query, instead of a simple bag of concepts representation. This structure is obtained by mapping each concept in a query to a pre-defined semantic category. Therefore, it requires that our external resource contains a semantic categorization for concepts. This categorization attaches each concept to a more abstract semantic category. For example, assume that a document contains the two terms “Adrenal Cortical Hypofunction" and “Hodgkin Disease", in UMLS [4], these two terms correspond to two concepts, and these two concepts belong to the same semantic category called :“Disease or Syndrome" . Using this idea, documents and queries can be represented by two semantic levels : concept-level and semantic category-level. We call these semantic categories *dimensions*. Therefore, the matching process between a query and a document will be at *concept-level*, and also at *dimension-level*.

In order to take advantage of this structure, we have two proposals :

- *Semantic Levels Matching* (SLM), which is based on the following paradigm : *relevant documents to a given query should share not only the maximum number of concepts but also the maximum number of dimensions*. This method takes into account the similarity between a document and a query represented by concepts and by dimensions. Therefore, The Relevance Status Value $RSV(d, q)$ is the fusion of these two similarities (similarity at concept-level and similarity at dimension-level). Figure 2 shows an example using this proposal.
- *Semantic Dimension Matching* (SDM), which depends on the following hypothesis : *each document dimension answers the part of the query which corresponds to the same dimension*. We partition each document into sub-documents according to its dimensions. Each sub-

4. Unified Medical Language System. It is a meta-thesaurus in medical domain. http ://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi ?book=nlmumls

document corresponds to a specific dimension and contains the document concepts that belong to this dimension. The same for queries. Figure 3 shows an example using this proposal.

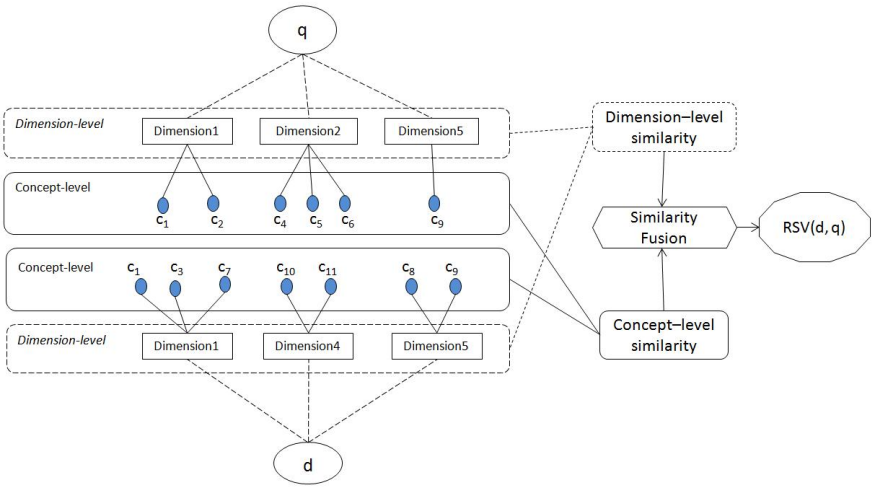


FIGURE 2 – Semantic Levels Matching : RSV is the fusion between the concept-level similarity of a document d and a query q, and the dimension-level similarity between d and q.

In the following we formally define all the components of our query structuring framework. This framework is the tuple (D, E, φ, RSV), where D is the document collection ; E is an external resource ; φ is a conceptual indexing function ; RSV is a matching function. We now detail the components of our framework.

4.1 External Resource E

An external resource E contains concepts, dimensions, and the mapping between them. Each concept can belong to one or more dimensions and each dimension owns several concepts. The external resource is used in the conceptual indexing to map a text into concepts. An external resource is modeled by $E = (C, M, \psi)$, where C is a set of concepts, M is a set of dimensions, ψ is a mapping function that maps each concept $c \in C$ into its set of dimensions $\psi(c)$.

$$\begin{array}{lcl} C & = & \{c_1, \cdots, c_n\} \\ M & = & \{m_1, \cdots, m_k\} \\ \psi & : & C \rightarrow 2^M \end{array}$$

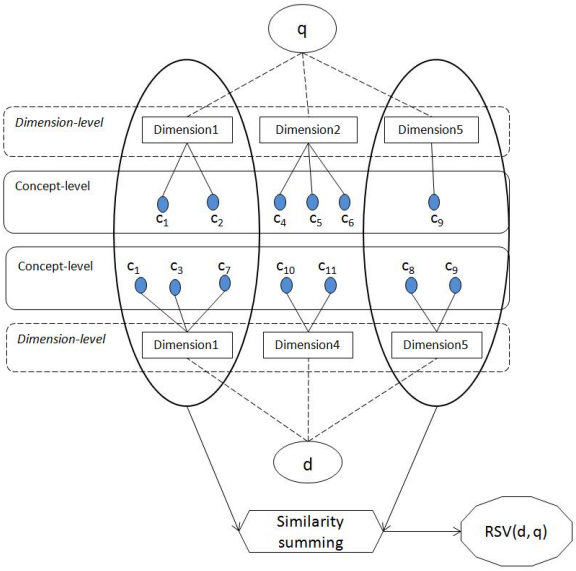


FIGURE 3 – Semantic Dimension Matching : Document d and a query q are splitted into dimensions $(1, 2, 5)$ and $(1, 4, 5)$, respectively, and each dimension contains the concepts of d or q that belong to this dimension. $RSV(d, q)$ is a sum of the similarity for all shared dimensions $(1, 5)$.

where c_i is a concept in external resource E. m_i is a dimension in E. 2^M is the power set of M.

For example, in UMLS, the concept $C0796561$ belongs to the following two dimensions : $\psi(C0796561) = \{T121, T129\}$, where $C0796561$ corresponds the medical term “melanoma" and the dimensions $T121$ and $T129$ correspond “Pharmacologic Substance" and “Immunologic Factor".

4.2 Query and Document Model

Conceptual indexing process converts documents and queries from their original form (e.g. text, image, etc.) to another form, which can be easily processed by machines. The conceptual indexing is the function :

$$\varphi : D \cup \{q\} \rightarrow 2^C$$

where 2^C is the power set of C. At this point, each document $d \in D$ is represented by a set of concepts $d_c = \varphi(d)$, and this is the first level of a

document representation in our framework (concept-level). The same for query q, it corresponds a set of concepts $q_c = \varphi(q)$.

The second level (dimension-level) aims to represent documents and queries depending on dimensions. Dimensions can be extracted from the external resource E using the function φ. For example, assume that a document d contains the two terms “Adrenal Cortex" and “Heart". In UMLS, these two terms correspond two concepts, these two concepts have the same dimension called :“Body Part, Organ". By applying the mapping function ψ to each concept $c \in d_c$ in the document, we obtain the second level d_m of a document d as follows :

$$d_m = \bigcup_{c \in d_c} \psi(c)$$

In our framework, it is possible to look at documents and queries from another point of view. A document d is a set of composed dimensions and each composed dimension contains the document concepts from d_c which is mapped on to this dimension. Hence, we define :

$$d_m^c = \{\delta(d_c, m) | m \in d_m\}$$

$$\delta \colon 2^C \times M \rightarrow 2^C$$

$$\delta(x, m) = \{c \in x | m \in \psi(c)\}$$

So the function δ is used to partition document or query concepts into composed dimensions.

We apply the same process used with documents, to queries. Therefore, for a query q we have a set of concepts q_c, a set of dimensions q_m, and a set of composed dimensions q_m^c. Concerning our two proposals, SLM is applied to d_c, q_c and also to d_m, q_m. Wheres, SDM is applied to d_m^c and q_m^c.

4.3 Matching Model

According to the previous section, we represent documents and queries by two semantic levels. These levels differ in their granularity or abstraction : a fine-grain level which is concept-level and a coarse-grain level which is dimension-level.

According to our two proposals, there are two ways to compute $RSV(d, q)$ between a query q and a document d. Each of them differently takes advantage of semantic query and document structuring.

4.3.1 Semantic Levels Matching (SLM)

In this proposal, to evaluate $RSV(d, q)$ between a document d and a query q, we take into account the similarity at concept-level computed between d_c and q_c, and the similarity at dimension-level computed between d_m and q_m. Then we combine these two similarities using equation 1.

$$RSV_{SLM}(d, q) = \alpha \times Sim_c(d_c, q_c) + (1 - \alpha) \times Sim_m(d_m, q_m) \quad (1)$$

where $\alpha \in [0, 1]$ is a tuning parameter and represents the importance of each level : normalized concept-level similarity $Sim_c(d_c, q_c)$, and normalized dimension-level similarity $Sim_m(d_m, q_m)$. These similarities Sim_c and Sim_m can be computed using any IR model (e.g. language models or BM25). Each concept $c_i \in d_c$ or $c_j \in q_c$ has a frequency reflecting its count in d or q. In addition, each dimension $m_i \in d_m$ or $m_j \in q_m$ has a frequency equals the sum of all concepts frequencies in this dimension.

4.3.2 Semantic Dimension Matching (SDM)

In this second proposal, each document is represented by a set of dimensions, and each dimension is described by a set of concepts. Thus, to evaluate $RSV(d, q)$ between a document d and a query q, we take into account the similarity of the shared dimensions between d and q. We combine these similarities using equation 2.

$$RSV_{SDM}(d, q) = \sum_{m_i \in d_m \cap q_m} Sim(m_i^d, m_i^q) \quad (2)$$

where the similarity $Sim(m_i^d, m_i^q)$ can be computed using one of any IR model (e.g. language models or BM25). These unnormalized similarities mean that a document dimension, which has more shared concepts with its correspondent query dimension, has a greater importance in the RSV. In addition, as we do not divide this sum on the number of shared dimension so the document which has more shared dimensions is more relevant in this proposal.

5 Experiments

In this section, we validate our two proposals SLM and SDM against the test collection CLEF 2011 and using the meta-thesaurus UMLS 2011. First, we present the context of our validation and then we show and analyze the obtained results.

FIGURE 4 – An example of case structure document from CLEF2011 collection.

5.1 Validation Context

5.1.1 CLEFMed 2011

CLEF is Cross-Language Evaluation Forum, which is a yearly campaign for evaluation of multilingual information retrieval since 2000. CLEF concerns searching medical text and images depending on multilingual documents that contain text and images.

The test collection CLEF 2011 contains two collections : image-based and case-based [8]. The goal of the image-based retrieval task is to retrieve an ordered set of images from the collection that best meet the information need specified as a textual statement and a set of sample images. The goal of the case-based retrieval task is to return an ordered set of articles that best meet the information need provided as a description of a “case".

Our validation is made on the case-based collection. The case-based topics are reused from previous years. 10 topics are available based on existing cases from the file Casimage. This file contains cases (including images) from radiological practice that clinicians write mainly for using them in teaching. The diagnosis and all information on the chosen treatment were then removed from the cases so as to simulate the situation of the clinician who has to diagnose the patient. In order to make the judging more consistent, the relevance judges were provided with the original diagnosis for each case. Figure 4 shows an example of a case document. This collection contains 55634 documents. The average document length in this collection is 2594.49 words, where the average query length is 19.7 words.

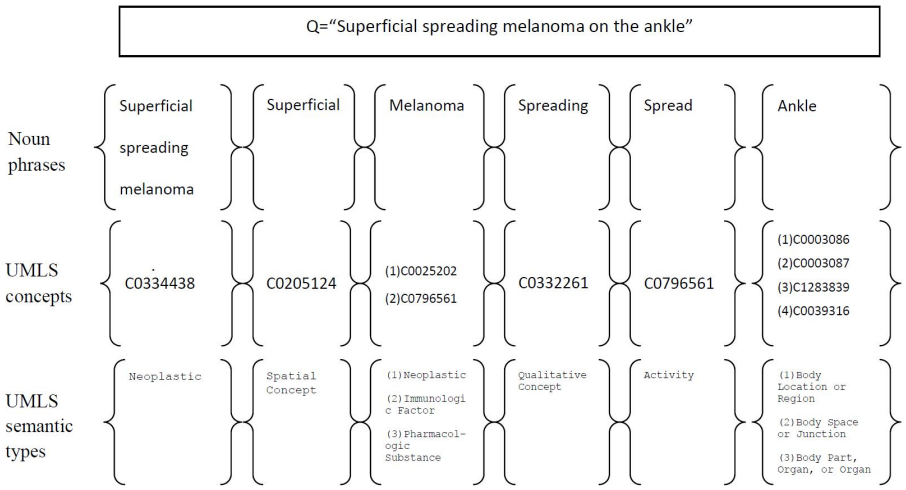


FIGURE 5 – MetaMap mapping example for the query "Superficial spreading melanoma on the ankle".

5.1.2 Conceptual Indexing

We use MetaMap [5] to do conceptual indexing operation. MetaMap is a tool that, when given a piece of text, finds and returns the relevant UMLS Metathesaurus concepts with this text. We also use UMLS 2011 as an external resource. UMLS contains concepts, these concepts are categorized using two different possibilities of dimensions called : *semantic groups* and *semantic types*. UMLS contains 16 different semantic groups and 135 semantic types. The difference between these two categorization is that semantic groups are more abstract than semantic types. Therefore, by moving to concepts, the average document length in the collection is 5752.38 concepts, where the average query length is 57.5 concepts. As we see, the average length in concepts is greater than words because it is normal to map a word or a phrase into more than one concepts using MetaMap (no disambiguation phase applied on MetaMap output and we do not use the relation between concepts in matching phase). Figure 5 shows an example of mapping a query Q using MetaMap.

5. Highly configurable program to map bio-medical text to the UMLS Metathesaurus :http ://metamap.nlm.nih.gov/.

5.1.3 Matching Models

We use three models for computing the similarity between a document and a query : Dirichlet (DIR), Jelinek-Mercer (JM), and BM25. Dirichlet and Jelinek-Mercer are two variations of language models [11, 15]. Language model is an up to date way for achieving matching process in information retrieval. This model represents new approach in information retrieval, and it gives better performance than many of other information retrieval models. One of the goals of our work is to experiment this model on concepts, instead of text. BM25 is a probabilistic model in information retrieval [14]. We used it to compare its results with previous two models, and to enlarge our model test.

Adapting classical models in order to apply them to concepts is a problem recently discussed [1, 2]. However, our simple adaption is just, for example if we talk about language modeling, we assume that query concepts q_c is generated by a probabilistic model based on document concepts d_c. Then, we have $count(c, d_c)$ the count of a concept c in a document d. $|d_c|$ is the number of concept in d. $|C_c|$ is the number of concepts in the collection. $p(c, C_c)$ is the collection language model for the concept c.

5.2 Results

In order to validate our two proposals for semantic query structuring, we define the following three experiments :

- Validation without semantic query structuring (baseline) : there is no structuring step for a query and a document, i.e. we only depend on concepts to compute RSV between a document and a query .
- Validation using Semantic Levels Matching : we structure queries and documents using our first proposal (SLM). In this experiment, the dimensions have two possible categorizations from UMLS which are semantic groups and semantic types.
- Validation using Semantic Dimension Matching : we structure queries and a documents using our second proposal. We use in this experiment UMLS semantic types as dimensions.

5.2.1 Validation without query structuring (baseline)

In this experiment, we leave queries and documents as a bag of concepts without applying our semantic query structuring approach. Concepts are extracted using MetaMap. We compute for each concept its frequency.

Then, we compute the RSV between a document and a query using the following IR models : Jelinek-Mercer, Dirichlet, BM25. This experimentation serves as a baseline for our evaluation. The MAP (Mean Average Precision) and the precision at 10 are used to evaluate the results Table 1.

Model	MAP	$P@10$
JM	0.1247	0.1600
Dir	0.1036	0.1500
BM25	0.0956	0.1400

TABLE 1 – MAP and $P@10$ of the three models, which are used in our evaluation (Jelinek-Mercer, Dirichlet, BM25).

5.2.2 Validation Using Semantic Levels Matching (SLM)

In this second experiment, we use our first semantic structuring proposal : SLM. Documents and queries are represented using two levels : concept-level and dimension-level. These two levels are extracted using MetaMap. Then we compute their frequencies. Frequency of a concept in a document or a query is the number of times this concept appears in this document or query, where frequency of a dimension is the sum of all concepts frequencies which belong to this dimension in a document or a query. In this experiment, dimension can be one of two UMLS categorization :

– Using UMLS semantic groups as dimensions (SLM-SG) : we consider UMLS semantic groups as dimensions. In order to compute the RSV between a document and a query we use the equation 1, where m is a UMLS semantic group in this case and Sim_c and Sim_m are one of the following models : JM, Dir, and BM25. The results obtained by different models are summarized in Table 2. This table contains the value of MAP (Mean Average Precision) and the improvement regarding the baseline.
We notice by using UMLS semantic groups as dimensions, there is no improvement obtained, because the distribution of semantic groups over the test collection is uniform. In other words, all documents nearly contain concepts from all groups as shown in Figure 6.

– Using UMLS semantic types as dimensions (SLM-ST) : we consider UMLS semantic types as dimensions. In order to compute the RSV between a document and a query, we also use the equation 1, where m is a UMLS semantic type in this case and Sim_c and Sim_m are one

Model	Baseline MAP	SLM-SG MAP	Gain	Baseline $P@10$	SLM-SG $P@10$	Gain
JM	0.1247	0.1256	+0.72%	0.1600	0.1600	0.0%
Dir	0.1036	0.1036	0.0%	0.1500	0.1500	0.0%
BM25	0.0956	0.0956	0.0%	0.1400	0.1400	0.0%

TABLE 2 – MAP improvement using UMLS semantic groups as dimension with our first proposal : SLM.

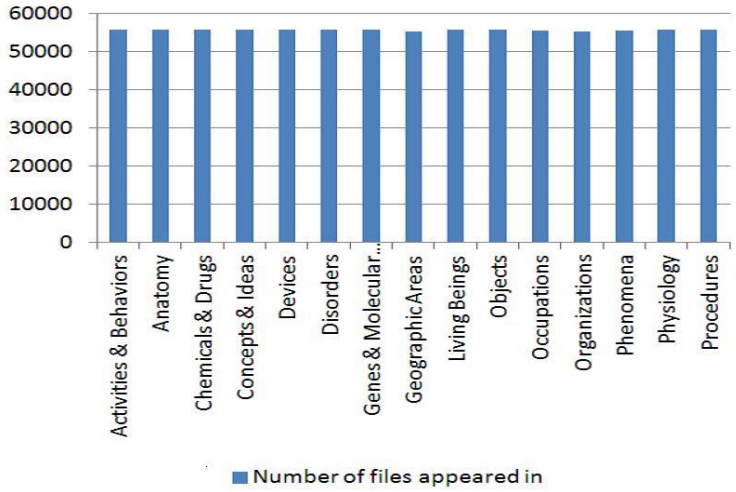


FIGURE 6 – Semantic groups distribution. This histogram shows that each document appears in all UMLS semantic groups or each document contains all UMLS semantic groups. In other words, these semantic groups are not able to discriminate the corpus.

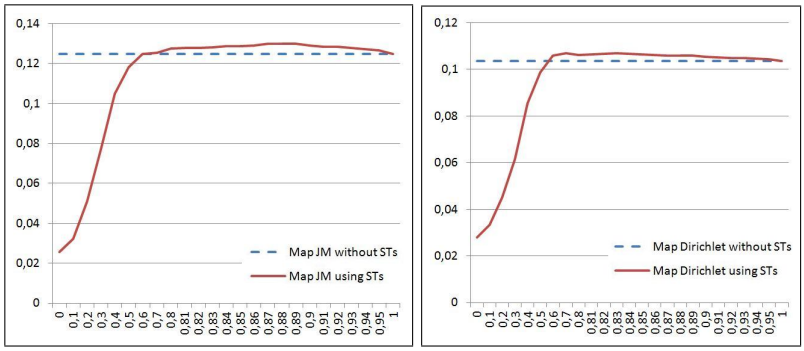
of the following models : JM, Dir, and BM25. The results obtained by different models are summarized in Table 3.

We notice by using UMLS semantic types as dimensions, there is an improvement obtained, because the distribution of semantic types over the test collection is less uniform than the distribution of semantic groups as shown in Figure 7. This distribution gives the potential for precision improvement. In addition, the α value plays an important role in this improvement. It determines the importance of each semantic level : concepts and dimensions in the matching process. Figure 7 shows MAP changes with α changes. As concepts are less abstract than semantic types, we should give a high value (close to 1)

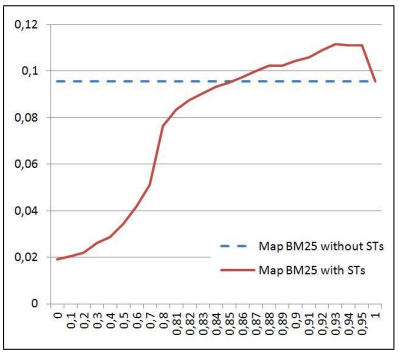
Model	Baseline MAP	SLM-ST MAP	Gain	Baseline $P@10$	SLM-ST $P@10$	Gain
JM	0.1247	**0.1299***	+4.17%	0.1600	0.1800	+12.5%
Dir	0.1036	0.1070	+3.28%	0.1500	0.1800	+20%
BM25	0.0956	0.1116	+16.73%	0.1400	0.1700	+22.22%

TABLE 3 – MAP and $P@10$ improvements using semantic types as dimensions and our SLM proposal. * the best value in CLEF2011 campaign for the case-based collection is 0.1297 [8].

for α in order to reflect the relative importance of concept-level comparing to dimension-level. For the results in Table 3, we fix $\alpha = 0.9$. In addition, α seems to be model independent and corpus dependent.



(a) Map changes for Jelinek-Mercer. (b) Map changes for Dirichlet.



(c) Map changes for BM25.

FIGURE 7 – MAP changes, using our first proposal Semantic Levels Matching and UMLS semantic types as dimensions, with different values of α and with the three IR models (Jelinek-Mercer, Dirichlet, BM25).

5.2.3 Validation Using Semantic Dimension Matching (SDM)

In this third experiment, we structure a query and a document using our second proposal SDM. Therefore, a document and a query consist of a set of dimensions and each dimension contains document or query concepts which belong to this dimension. Here, we only use UMLS semantic types as dimensions. However, we did not validate this proposal against UMLS semantic groups, because they are uniformly distributed over our test collection Figure 6. For computing $RSV(d, q)$ between a document d and a query q, we use the equation 2. Sim is one of the following models : JM, Dir, and BM25. The results obtained are summarized in Table 4.

Model	Baseline MAP	SDM-ST MAP	Gain	Baseline $P@10$	SDM-ST $P@10$	Gain
JM	0.1247	0.1166	-6.57%	0.1600	0.0.1600	0.0%
Dir	0.1036	0.0791	-23.64%	0.1500	0.1100	-26.6%
BM25	0.0956	0.1043	+9.1%	0.1400	0.1600	+14.3%

TABLE 4 – MAP and $P@10$ improvements using semantic types as dimensions and our semantic dimension matching proposal.

As we split documents into dimensions and use language model on these dimensions, the results for Jelinek-Mercer and Dirichlet are less than baseline. We think that language models give poor results for very short documents. In other words, language models give a better probability estimation for long documents than short documents. In the other hand, the results of BM25 is better than baseline.

6 Conclusion

In this paper, we present a semantic query structuring framework for replacing the flat representation of a query and a document by a structured query and document in a specific domain. This approach aims to help domain specialists in their searching task by providing more precise results. We propose two ways in order to take advantage of this structuring approach : *Semantic Levels Matching* and *Semantic Dimension Matching*.

The best result obtained has about 17% improvement in MAP and 30% in precision at the first five results . In addition, one of our result is better than the best result obtained in CLEF2011 campaign for cased-based collection [8]. The analysis of our results shows that the improvement in

precision depends on the distribution of dimensions over the studied collection and the granularity of these dimensions. Future work will focus on validating our work to other test collections and other domains. Besides, we will study the relation between the value of our tuning parameter α and the properties of studied collections.

Références

[1] ABDULAHHAD K., CHEVALLET J.-P. & BERRUT C. (2012). MRIM at ImageCLEF2012. From Words to Concepts : A New Counting Approach. Working notes - CLEF 2012.

[2] ABDULAHHAD K., CHEVALLET J.-P. & BERRUT C. (2013). Revisiting the Term Frequency in Concept-Based IR Models. In *24th International Conference on Database and Expert Systems Applications (DEXA 2013), Prague, Czech Republic*.

[3] AUSSENAC-GILLES N. & MOTHE J. (2004). Ontologies as background knowledge to explore document collections.

[4] BAZIZ M., BOUGHANEM M. & AUSSENAC-GILLES N. (2005). Conceptual indexing based on document content representation. CoLIS'05, p. 171–186, Glasgow, UK.

[5] CHEVALLET J.-P., LIM J.-H. & LE D. T. H. (2007). Domain knowledge conceptual inter-media indexing : application to multilingual multimedia medical reports. CIKM '07, p. 495–504, Lisbon, Portugal.

[6] DOU Z., HU S., LUO Y., SONG R. & WEN J.-R. (2011). Finding dimensions for queries. CIKM '11, p. 1311–1320, Glasgow, Scotland, UK.

[7] EERO HYVÖNEN A. S. & SAARELA S. (2003). Ontology-based image retrieval.

[8] KALPATHY-CRAMER J., MÜLLER H., BEDRICK S., EGGEL I., DE HERRERA A. G. S. & TSIKRIKA T. (2011). Overview of the clef 2011 medical image classification and retrieval tasks. In *CLEF (Notebook Papers/Labs/Workshop)*.

[9] LI X., WANG Y.-Y. & ACERO A. (2009). Extracting structured information from user queries with semi-supervised conditional random fields. SIGIR '09, p. 572–579, Boston, MA, USA.

[10] LIN J. & DEMNER-FUSHMAN D. (2006). The role of knowledge in conceptual retrieval : a study in the domain of clinical medicine. SIGIR '06, p. 99–106, Seattle, Washington, USA.

[11] PONTE J. M. & CROFT W. B. (1998). A language modeling approach to information retrieval. p. 275–281.

[12] RADHOUANI S. & FALQUET G. (2006). Using external knowledge to solve multi-dimensional queries. p. 426–437, Amsterdam, The Netherlands, The Netherlands.

[13] RADHOUANI S., KALPATHY-CRAMER J., BEDRICK S., BAKKE B. & HERSH W. (2010). Using media fusion and domain dimensions to improve precision in medical image retrieval. CLEF'09, p. 223–230, Corfu, Greece.

[14] ROBERTSON S. E. & WALKER S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. SIGIR '94, p. 232–241, Dublin, Ireland.

[15] ZHAI C. & LAFFERTY J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, **22**(2), 179–214.

Les Bulletins de Santé du Végétal : spécification d'une base d'annotations pour la recherche d'information sémantique en français.

Catherine ROUSSEY[1], Jean-Pierre CHANET[1], Stéphan BERNARD[1]

[1] UR TSCF Irstea,
24 avenue des Landais CS 20085 63 178 Aubière
prenom.nom@irstea.fr
http://www.irstea.fr

Résumé : Dans cet article nous décrivons les différents processus d'annotation envisagés pour annoter un corpus sur les bulletins d'alertes agricoles. Notre but est de publier ces annotations sur le web de données pour permettre à des applications environnementales d'enrichir au besoin nos annotations.

Mots-clés : Annotations sémantiques, annotations spatio-temporelles, recherche d'information sémantique, benchmark, bulletins d'alerte agricole.

1 Introduction

Pour être plus respectueuse de l'environnement, l'agriculture doit modifier ses pratiques. Pour se faire, le plan Ecophyto s'appuie notamment sur le système de surveillance des pratiques agricoles, dont les Bulletins de Santé du Végétal (BSV) sont un des moyens de communication. Ce corpus du domaine agricole contient des informations sur les attaques des bio-agresseurs des cultures par région (par exemple : la DRAAF de la région PACA signale une explosion des attaques de la rouille du blé sur les cultures de blé dur en vallée du Rhône, dans son bulletin du 23 mai 2011).

Nous souhaitons mettre en place plusieurs processus d'annotations spatio-temporelles, notamment pour permettre à des acteurs du domaine agricole de retrouver les BSV répondant à leur besoin, ou à des fins d'étude des évolutions spatio-temporelles des attaques sur les cultures.

Cet article présente un premier état des lieux de nos besoins en annotations. Tout d'abord nous présenterons nos motivations et le corpus des BSV. La section suivante définira notre processus d'annotation,

partant de l'extraction des bulletins de santé du végétal et allant jusqu'aux annotations, chacune de ces étapes permettant la collecte d'informations. Nous terminerons par des explications sur la mise en place d'un système de recherche d'information sémantique dédié au monde agricole. Nous souhaitons mettre en place plusieurs expérimentations pour utiliser ce corpus comme benchmark pour la recherche d'informations en français.

2 Motivations

Au fil des dernières décennies, les pratiques agricoles ont fortement évolué sous le jeu de diverses contraintes : enjeux sociétaux et environnementaux, cadre réglementaire, changement climatique… Parallèlement, le rôle des données en agriculture a également fortement évolué : d'abord utilisées à des fins de traçabilité et de sécurité alimentaire, les données, qui sont de plus en plus nombreuses, contribuent maintenant directement au changement des pratiques agricoles par leur aide à une meilleure compréhension de celles-ci. La multiplication des équipements embarqués, des smart-phones, des capteurs aux champs, etc., permet à l'heure actuelle de disposer de grands volumes de données spatio-temporelles (Steinberger et al. 2009). Le prochain enjeu est de rendre ces données disponibles à l'ensemble des acteurs de la filière afin qu'ils puissent les mobiliser dans les outils d'aide à la décision et dans les outils d'analyse (Xie et al 2008; Goumopoulos et al. 2009). Le web de données est une opportunité pour accélérer cette mutualisation des données et, par voie de conséquence, pour faire évolutier les pratiques agricoles.

Afin de pouvoir publier ces données sur le web de données, il convient de structurer les relations entre les différents concepts manipulés par la thématique agriculture : plantes, maladies, ravageurs, pesticides, rotation, … Un certain nombre de ressources sont disponibles et mobilisables : ontologies de taxons, thésaurus, bases de données, corpus de documents. Mais lorsqu'on s'intéresse à un thème particulier, comme par exemple la protection des cultures, les ressources se font rares. Il faut en effet mobiliser un ensemble de ressources et les relier entre elles. Nous proposons de créer une méthode à même de construire une base de connaissances agrégeant un ensemble de ressources autour d'une thématique. Cela permettra ensuite de publier sur le web de données les données disponibles de manière pertinente, mais également d'annoter les nombreux documents mobilisables pour faire évoluer les pratiques.

Nous avons à notre disposition un corpus de bulletins d'alertes agricoles francophones que nous souhaitons annoter à l'aide d'une base de connaissances sur la protection des cultures. Ainsi, les acteurs du monde agricole pourront rechercher dans ce corpus les bulletins qui les intéressent particulièrement. Dans un second temps nous aimerions

transformer ce corpus en un benchmark pour la recherche d'informations sémantiques francophones, en déployant plusieurs campagnes d'évaluation avec différents acteurs du monde agricole (agriculteurs, agronomes, jardiniers, conseillers agricoles).

3 Les Bulletins de Santé du Végétal

Dans nos travaux, nous nous focalisons sur la protection des cultures. Nous avons à disposition une collection de documents intitulés « Bulletins de Santé du Végétal » (BSV). Nous souhaitons constituer une base archivant les données des attaques des bio-agresseurs sur les cultures (par exemple : attaque de rouille sur une culture de blé dur).

Le Grenelle de l'environnement et le plan Ecophyto ont renforcé les réseaux de surveillance sur les cultures et les pratiques agricoles. Les Bulletins de Santé du Végétal sont une des modalités mises en place par ces réseaux de surveillance.

Le Bulletin de Santé du Végétal (BSV) est un document d'information technique et réglementaire, rédigé sous la responsabilité d'un représentant régional du ministère de l'agriculture. Ce représentant peut être par exemple la Chambre Régionale d'Agriculture ou bien la Direction Régionale de l'Alimentation, de l'Agriculture et de la Forêt (DRAAF). Ce représentant doit mettre à disposition ses bulletins sur son site internet afin d'en permettre un accès public. La conséquence est que les BSV sont répartis sur différents sites web (un par région). À notre connaissance, il n'existe pas encore de système donnant un accès uniforme à l'ensemble des BSV.

Les BSV sont rédigés en collaboration avec de nombreux partenaires impliqués dans la protection des cultures. La liste des auteurs des BSV varie en fonction de la région et de la filière agricole. Par conséquent leur contenu et leur présentation ne sont pas uniformes et varient en fonction des auteurs.

Les BSV diffusent des informations relatives à la situation sanitaire des principales productions végétales de la région et proposent une évaluation des risques encourus pour les cultures. Des données générales concernant les stratégies de lutte (notes nationales, …) ou sur la réglementation peuvent figurer également dans les BSV.

Selon l'actualité sanitaire et/ou la culture, le rythme de parution des BSV est variable, allant d'une parution hebdomadaire à mensuelle.

Les BSV sont une synthèse des observations effectuées sur les cultures. Il existe des bases de données d'observations mais la rédaction des BSV oblige leurs auteurs à décider si une observation est un phénomène unique non représentatif ou un phénomène important représentatif d'une réalité. Les BSV ne sont pas une agrégation

BULLETIN DE SANTE DU VEGETAL
MIDI-PYRENEES

Grandes Cultures - n°29

15 juin 2011

A retenir

MAÏS

Pyrale *: Période de risque maximal terminée. Vol étalé, en particulier dans les zones les plus froides.*

Sésamie *: Fin du vol de première génération.*

Puceron vert *: Risque modéré. A surveiller.*
Vers gris *: Risque modéré. Surveillez les parcelles les moins avancées.*
Cicadelle bleue *: Risque faible.*

MAÏS

• Stade phénologique et état de la culture

Après un ralentissement du développement lié à la diminution des sommes de températures, l'offre climatique est de nouveau au rendez-vous.

Les parcelles les plus avancées atteignent 16 feuilles. Le stade moyen sur la région est de 12 feuilles.

Les pluies de ces dernières semaines ont permis le semis des dernières parcelles, notamment dans l'Aveyron. Ces parcelles présentent des stades autour de 4 feuilles.

• Sésamie

Le vol de première génération est quasiment terminé, quelle que soit la zone.

Peu de parcelles sont touchées par des pieds de ponte. Elles se situent essentiellement sur l'ouest de la région, en particulier dans l'ouest du Gers. Le nombre de larves vivantes observées au niveau de ces pieds de ponte, est faible. Cela résulte probablement de l'humidité ambiante de ces dernières décades, peu favorable à la sésamie.

Évaluation du risque : Il est trop tôt pour prévoir l'importance du deuxième vol.

• Pyrale

A part sur le sud du Tarn et Garonne, le piégeage est en diminution significative, avec quelques individus par piège. Cela peut s'expliquer à la fois par un vol en décroissance, ainsi que par la météo de ces derniers jours peu favorable à une activité des papillons.

Dans les situations les plus froides de Midi-Pyrénées, le vol a tendance à s'étaler du fait de la diminution des températures des semaines précédentes.

Directeur de publication :
Jean-Louis CAZAUBON
Président de la Chambre Régionale d'Agriculture de Midi-Pyrénées
BP 22107 – 31321 CASTANET TOLOSAN Cx
Tel 05.61.75.26.00 – Fax 05.61.73.16.66
Dépôt légal : à parution
ISSN en cours

automatique de données mesurées mais bien une synthèse humaine des jugements sur des observations.

FIGURE 1 – *Exemple de BSV de la région Midi-Pyrénées pour la filière grande culture*

Nous avons récupéré les BSV de l'année 2011 de 19 régions. En moyenne une région publie plus d'une centaine de BSV par an. Au total

nous avons 2825 BSV pour l'année 2011. Dans un premier temps nous limiterons notre analyse aux BSV de la région Bourgogne concernées par la filière « grande culture », c'est à dire 37 BSV.

4 Processus d'annotation des BSV

Ce processus d'annotation de notre collection de bulletins est constitué de trois phases, chacune permettant l'extraction d'informations, de plus en plus fines. La première phase, qui consiste en la récupération et l'extraction des fichiers bruts, ne nécessitera pas d'analyse du contenu des BSV. Il permettra néanmoins d'identifier la région, la date et la filière agricole concernée. La seconde phase est un pré-traitement qui permettra de récupérer les noms des cultures et de leurs agresseurs dans les textes. Elle aura pour résultat une première série d'observations spatio-temporelles des attaques des agresseurs sur les cultures. La dernière phase sera l'annotation en elle-même, et visera à récupérer les niveaux de risques liés aux attaques des agresseurs dans les cultures. Ces différentes phases du processus sont décrites plus précisément dans les sections suivantes.

4.1 Extraction brute

4.1.1 Principe

Les BSV sont disponibles sur les sites des chambres d'agricultures ou des DRAAFT. Les pages contenant les liens permettant le téléchargement des fichiers pdf contenant les BSV nous donnent déjà un certain nombre d'informations :

- La région : elle est indiquée par le nom du site web
- La filière agricole concernée : le site a généralement une page par filière, donc le nom de la page web nous permet de récupérer cette information
- Le numéro du BSV : le nom du fichier téléchargé contient souvent le numéro du BSV.
- La date de parution : le nom du fichier téléchargé contient souvent cette date.

Ce premier niveau d'annotations a été réalisé en analysant les différentes pages web contenant les liens vers les fichiers pdf.

4.1.2 Problèmes rencontrés

Il est important de constater que les listes des filières agricoles disponibles ne sont pas normalisées d'une région à une autre. En effet, tout dépend des productions agricoles de la région. Certaines régions vont

regrouper plusieurs filières. Par exemple, la Haute-Normandie a une filière « arboriculture et petits fruits ». D'autres régions ont au contraire une production très spécialisée et ne vont indiquer que la production qui les intéresse : la Bourgogne, par exemple, a une filière « cassis » sans avoir de filière « petit fruit ». Il est donc nécessaire de normaliser les noms des filières entre les régions, et de constituer une classification homogène des filières de productions agricoles en France. Une classification des filières agricoles françaises est déjà disponible sur le site français de Wikipedia[1]. Cette classification a été enrichie manuellement pour couvrir l'ensemble des noms de filières utilisées dans les 20 régions.

Suivant les cas, il est possible qu'un BSV soit associé à plusieurs filières.

4.2 Pré-traitement

4.2.3 Principe

Ce second niveau d'analyse, au contraire du précédent, a pour but de travailler sur le contenu textuel des BSV. Nous souhaitons récupérer dans le texte, le nom des plantes cultivées et le nom de leurs agresseurs.

Dans l'exemple de la **Fig 1**, les noms des cultures sont présents dans les titres des sections (maïs), et les agresseurs parmi les sous-sections (sésamie, pyrale, …). La structure logique du document nous donne l'indication que la sésamie et la pyrale sont des agresseurs du maïs. Cette régularité éditoriale est vérifiée sur l'ensemble des BSV de toutes les régions. Ainsi la reconnaissance des noms des organismes vivants dans les titres des sections et leur organisation logique dans le texte nous permet de produire nos premières annotations : attaque de la culture maïs par l'agresseur pyrale dans la région Midi-Pyrénées à la date du 15 juin 2011.

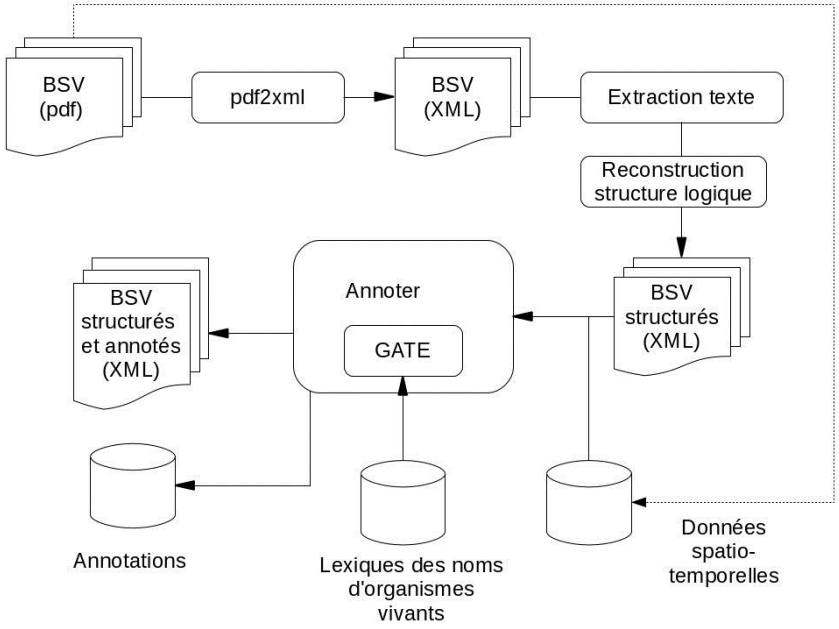
La **Fig 2** présente le processus de pré-traitement, puis le processus d'annotation :

Dans un premier temps il est nécessaire de transformer les fichiers au format PDF en un format qui permet une lecture plus aisée de la structure du document. Le logiciel pdf2xml (Dejean Giguet 2012) nous offre une sortie XML contenant des balises représentant l'organisation physique du document et sa mise en page. Chaque mot est encadré par une balise ayant des attributs de mise en forme et de position dans la page. A partir de ces fichiers XML, nous récupérons l'ensemble des contenus textuels.

Ensuite, nous reconstruisons automatiquement la structure logique des documents, en rassemblant les mots. Le but est d'identifier les sections, leur titre et leurs paragraphes, ainsi que leur inclusion logique (section et sous section). Cette structure logique sera décrite dans des fichiers XML,

[1] http://fr.wikipedia.org/wiki/Classement_en_France_des_cultures_par_groupes_d'usage

et seront les entrées du processus d'annotation. On leur associera les informations spatio-temporelles recueillies lors de l'extraction des



fichiers bruts.

Cette phase de pré-traitement est en cours d'élaboration et a été testée sur les grandes cultures dans quatre régions.

FIGURE 2 – *Processus d'annotation.*

4.2.4 Problèmes rencontrés

La mise en page des BSV n'est pas homogène : certains utilisent par exemple une mise en page en plusieurs colonnes, d'autres ont une mise en page sur une seule colonne. La détection des colonnes complexifie le processus de reconstruction de la structure logique, surtout lorsque certains effets de mise en page, tels que des cadres chevauchant plusieurs colonnes, empêchent d'établir correctement la frontière entre colonnes.

La mise en page des BSV ne varie pas seulement d'une région à l'autre, mais aussi d'une culture à l'autre, les auteurs des bulletins semblant peu contraints sur cet aspect. Cette diversité rend difficile la création d'une structure logique uniforme pour l'ensemble des BSV. Toutefois, l'association des tailles de caractères et de leurs attributs (gras, italique) et d'une description générale des structures des différents BSV

permet d'obtenir, à quelques exceptions près, les informations nécessaires à la phase d'annotation.

4.3 Annotations

4.3.5 Principe

L'annotation des BSV est composée de deux étapes :

- La première étape, appelée simplement « annotation », consiste à reconnaître les noms des organismes vivants dans les titres des sections, à partir d'un lexique.
- La seconde étape, nommée annotation fine, vise à estimer, à partir du contenu de la section, si l'agresseur en question est cité pour avoir été observé, ou au contraire parce qu'aucune attaque n'a été constatée, ou encore pour inciter à prévenir une attaque.

Pour ces deux étapes, la plateforme GATE sera utilisée (Cunnigham et al. 2011).

4.3.6 L'annotation « simple »

4.3.6.1 Principe

Cette première phase d'annotation utilisera en entrée un lexique précédemment défini. Le lexique sera composé des noms de plantes cultivées en France ainsi que les noms de leurs agresseurs (champignons, plantes, insectes, animaux, …). Ces lexiques seront dérivés d'une base de connaissances construite à partir de ressources disponible sur le LOD.

La relation entre une plante cultivée et son agresseur sera dérivée automatiquement de la structure logique du document. L'inclusion logique entre sections implique la présence de l'agresseur sur la culture.

4.3.6.2 Problèmes rencontrés

Pour mettre en place ce processus d'annotations, nous avons besoin de construire des lexiques d'organismes vivants et de les classer en plantes cultivées (culture) et en agresseur des cultures (organisme vivant nuisant au développement d'une plante cultivée). Ces lexiques seront dérivés d'une ontologie des organismes vivants capable de classer automatiquement un organisme en plantes cultivées ou en agresseurs d'une culture. Contrairement à nos attentes, catégoriser une plante en plante cultivée (culture) ou en agresseur n'est pas évident :

1 Une plante peut être associée à différentes filières agricoles. La filière agricole indique l'usage de sa production : huile, céréale panifiable, fourrage, ornementale etc… et donc une même plante peut être utilisée pour différentes cultures.

2 De plus, sa culture est contextuelle. En effet, certaines plantes sont plus adaptées à un climat qu'à un autre ; les usages alimentaires du pays peuvent aussi entrer en compte. Par exemple le riz est consommé partout en France mais il s'agit principalement de riz importé. Les principaux pays producteurs de riz sont la Thaïlande, le Pakistan, les USA et la Chine. En France métropolitaine seule la région PACA produit du riz. La définition d'une culture (d'une filière agricole) est dépendante du pays et donc elle ne peut pas être généralisée.

3 Dernier point, une plante peut à la fois être une plante cultivée ou un agresseur d'une autre culture. Par exemple, sur une parcelle agricole cultivée pour du blé, il peut persister une ancienne culture qui va gêner le développement de la culture en cours. Ainsi d'une année à une autre, la même plante peut être catégorisée sous forme de culture ou d'agresseur.

Pour identifier les plantes cultivées, nous allons nous baser sur la classification des filières agricoles précédemment définies dans le processus d'annotations brutes. Dans notre ontologie, les organismes vivants seront décrits par leurs caractéristiques agronomiques (plantes, champignons, animaux etc.). Cette description est générique, c'est à dire commune quel que soit le pays concerné ou le domaine étudié (agriculture, botanique, épidémiologie). Cette description sera ensuite enrichie par l'usage de la plante en agriculture en indiquant sa filière agricole. Ainsi, un raisonneur pourra automatiquement découvrir les plantes cultivées en France. Dans un second temps, nous allons nous intéresser aux agresseurs des cultures les plus connus qui apparaissent dans des listes d'agresseurs publiées. La notion d'agresseur étant elle aussi contextuelle il est nécessaire de la dériver d'une observation d'agression. Nous allons, de la même façon que précédemment, enrichir la description d'un organisme par ses agressions connues, pour dériver les agresseurs des cultures. De la même manière nous pourrons dériver les auxiliaires des cultures, c'est-à-dire les agresseurs des agresseurs des cultures.

Cette ontologie sera enrichie au cours du temps pour tenir compte des différents noms régionaux que peut avoir un organisme vivant et aussi de l'apparition de nouveaux organismes dans différents pays. Par exemple les changements climatiques permettent à des organismes de se développer dans des lieux où préalablement ils n'avaient pas les moyens de subsister. Les nouveaux moyens de transports et routes commerciales permettent aussi des migrations d'organismes vivants.

4.3.7 L'annotation fine

4.3.7.3 Principe

Dans l'annotation précédente des BSV nous avons fait l'hypothèse que toute apparition dans le texte d'un agresseur indique qu'il y a eu une observation de l'agression d'une culture par cet agresseur. Cette hypothèse est fausse dans certains cas. En effet, il arrive que la section sur l'agresseur indique qu'aucune observation de cet agresseur sur la culture n'a été constatée. Il est possible aussi que le texte indique que l'agresseur n'est pas assez présent sur la culture pour entraîner un risque.

Nous avons donc besoin d'extraire des textes un ensemble d'informations précises comme :

- Le stade de développement de la plante lors de l'observation. Cette information permet une estimation du risque. En effet, selon le stade de développement de la plante, l'apparition d'un organisme extérieur peut nuire à la culture ou avoir un impact limité voir inexistant sur le développement de la culture. Ce stade de développement est aussi nécessaire pour connaître le moyen de lutte adéquat. Les stades de développement des plantes sont publiés dans des listes et nous pourrons, à l'aide d'un lexique approprié, extraire ces informations.
- Le niveau de risque de l'agression. Ce niveau est dépendant du stade de développement de la culture et du nombre d'agresseurs présents dans la culture. Les niveaux de risques sont aussi publiés dans des listes. Nous pourrons donc extraire ces informations à l'aide d'un lexique.
- Le nombre de parcelles agricoles impliquées où une attaque a été constatée. Ce nombre peut aussi être exprimé en pourcentage.

Pour extraire ce type d'information nous avons donc besoin d'une analyse des textes plus fine, avec des processus de traitement automatique du langage naturel. Nous allons mettre en place des règles linguistiques pour extraire des hypothèses qui devront être validées manuellement ensuite.

La **Fig 3** présente le processus d'annotations fines. Nous allons, comme précédemment, utiliser la plateforme Gate. Cette plateforme utilisera plusieurs lexiques dérivés automatiquement de notre ontologie. Le but des lexiques sera de découvrir les termes représentant les stades de développement des cultures et les niveaux de risques. Ensuite nous appliquerons des règles linguistiques pour extraire l'observation d'une attaque d'une culture avec son niveau de risque et le stade de développement de la plante, à partir du contenu de la section.

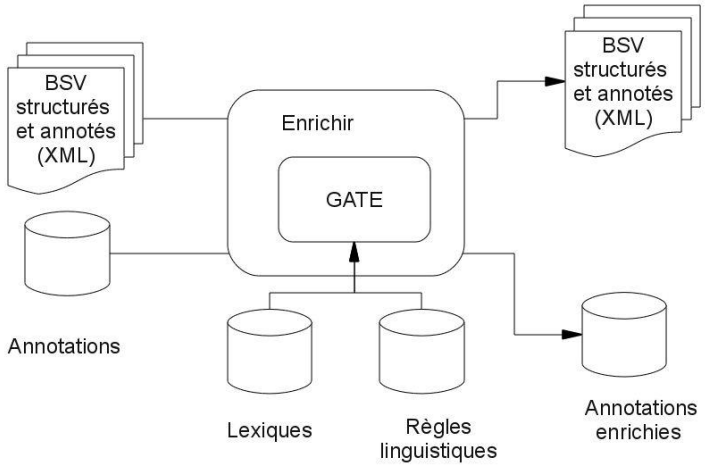


FIGURE 3 – *Processus d'annotations fines.*

4.3.7.4 Problèmes rencontrés

Nous voulons atteindre plusieurs buts dans cette annotation. Dans un premier temps, nous voulons découvrir s'il y a eu effectivement une observation d'une attaque d'un agresseur sur une culture. Ensuite nous voulons enrichir cette observation en indiquant le niveau de risque et le stade de développement de la plante. Ces informations ne sont pas forcement indiquées dans le texte des BSV. Parfois le BSV rappelle juste dans un tableau les niveaux de risque en fonction des stades de développement, sans donner aucune indication sur l'observation réalisée. Donc nous devrons prendre en compte les absences d'observations. Les résultats du processus d'annotations fines devront être validés manuellement pour compléter au besoin les imperfections de ce processus automatique.

5 Système d'interrogation des BSV

Au fur et à mesure, l'ensemble de nos annotations sera mis à disposition sur le web par le biais d'un SPARQL end point. Pour faciliter l'interrogation SPARQL de ces annotations, nous intégrerons, au dessus

de SPARQL end point, le système d'interrogation en langage naturel développé par l'IRIT intitulé SWIP (Pradel et al, 2012). Ce système permet aux utilisateurs de requêter un SPARQL end point à l'aide d'une requête sous forme d'une question en langage naturel, Il se base sur des patrons de requêtes représentatifs du domaine et sur des phrases en langage naturel pour proposer à l'utilisateur plusieurs requêtes SPARQL possibles.

Il est à noter que nous ne sommes pas dépositaires des BSV. Nous avons connaissance d'un autre projet d'annotations des BSV dont l'INRA est l'un des membres partenaires : Le projet Vespa. Ce projet devrait mettre en ligne les versions électroniques des documents. Donc nous ne mettrons en ligne que nos annotations. Nous établirons un lien entre nos annotations et les versions des BSV mises en lignes dans le projet Vespa.

6 Conclusion et perspectives

Dans cet article nous décrivons les différents processus d'annotation envisagés pour annoter un corpus sur les bulletins d'alertes agricoles. Notre but est de publier ces annotations sur le web de données pour permettre à des applications environnementales d'enrichir au besoin nos annotations. Nous souhaitons utiliser ce corpus de bulletins pour mettre en place une première campagne d'évaluation de systèmes de recherche d'information sémantique en français.

Dans le but de développer notre benchmark de recherche sémantique en français, nous souhaitons mettre en place plusieurs expérimentations de recherche d'information sur les BSV. Nous avons en tête d'utiliser plusieurs groupes d'utilisateurs :

- des experts en agronomie, notamment de l'INRA,
- des responsables d'exploitations issus des établissements d'enseignement agricole,
- des internautes agriculteurs identifiés dans les réseaux sociaux agricoles.

Ainsi il sera aussi possible d'effectuer des recherches en fonction du profil de l'utilisateur

Références

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. (2011) Text Processing with GATE (Version 6) University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311

DEJEAN H., GIGUET. (2012) pdf2xml, available at `http://sourceforge.net/projects/pdf2xml/` last update octobre 2012.

GOUMOPOULOS, C., KAMEAS, A.D., CASSELLS, A. (2009). An ontology-driven system architecture for precision agriculture applications. International Journal of Metadata, Semantics and Ontologies 4, p 72–84.

PRADEL, C. HAEMMERLE, O. HERNANDEZ, N. (2012) Des patrons modulaires de requêtes SPARQL dans le système SWIP. Dans : Journées Francophones d'Ingénierie des Connaissances (IC 2012), Paris, 27/06/2012-29/06/2012, juin 2012, p 385-400.

STEINBERGER, G., ROTHMUND, M. & AUERNHAMMER, H. (2009) Mobile farm equipment as a data source in an agricultural service architecture. Computers and electronics in agriculture, 65(2), 2009, p.238−246.

XIE, N., WANG, W., YANG, Y. (2008) Ontology-based agricultural knowledge acquisition and application, in: Computer And Computing Technologies In Agriculture, Volume I. Springer, 2008, p. 349–357.

Une ontologie documentaire pour la recherche d'information relationnelle

Nada Mimouni, Adeline Nazarenko et Sylvie Salotti

LIPN, CNRS (UMR 7030), Université Paris Nord
Sorbonne Paris Cité, F-93430 Villetaneuse
Nada.Mimouni, Adeline.Nazarenko,
Sylvie.Salotti@lipn.univ-paris13.fr

Résumé : Cet article présente un modèle documentaire qui prend en compte non seulement les annotations sémantiques portées par les documents, leurs structures logiques et leurs différentes versions mais aussi la structure d'une collection documentaire composée de différents types de documents reliés entre eux par des types variés de relations.

Le développement de la recherche d'information sémantique dans ses usages professionnels suppose d'exploiter tout cet ensemble de propriétés documentaires. Dans le domaine juridique, notamment, il faut pouvoir retrouver les documents d'un type particulier (par ex. des décrets émis par telle juridiction) qui portent sur une notion spécifique juridique (ex. contrat) et qui précisent un texte de loi donné. Il faut aussi pouvoir retrouver l'ensemble des textes portant sur un sujet donné (ex. le bruit) en vigueur à une date précise et la manière dont ils ont été appliqués, c'est-à-dire la jurisprudence relative à ces textes.

Au moment où les efforts de standardisation et d'ouverture donnent accès aux données publiques, il est essentiel de penser la modélisation des collections juridiques pour offrir des fonctionnalités d'interrogation avancées. L'approche proposée repose sur les standards du web sémantique. Elle a l'originalité d'intégrer les différentes propriétés documentaires dans un modèle unique qui permet de croiser les critères sémantiques, temporels et relationnels dans la recherche d'information.

Mots-clés : Collection documentaire, Modèle ontologique, RI sémantique, Intertextualité, Requêtes relationnelles.

---

0. Ce travail a été partiellement financé par le projet LEGILOCAL (FUI 2010-2013). Nous remercions nos partenaires, notamment Meritxell Fernández-Barrera, Eve Paul et Danièle Bourcier pour l'aide qu'elles nous ont apportée dans la compréhension des enjeux de la recherche d'information juridique.

1 Introduction

Avec l'émergence du web sémantique et le mouvement d'ouverture de données touchant à de plus en plus de domaines, des efforts sont faits pour rendre ces données compatibles avec les standards et normes définies dans le web sémantique (XML, RDF, SPARQL) et définir des modèles sémantiques (ontologies) pour différents domaines. Ces efforts ont pour but d'assurer l'interopérabilité des données et de faciliter leur accès et leur gestion par les utilisateurs.

Par exemple dans le domaine juridique, plusieurs standards XML juridiques ont été définis pour normaliser la structure des textes de loi, assister la production de ces textes et améliorer leur interopérabilité. En parallèle, des initiatives d'ouverture de données gouvernementales se multiplient (ex. UK Government Linked Data). Pourtant, les données mises à disposition restent souvent sous-exploitées.

Peu d'approches en effet ont été définies pour permettre la recherche d'information dans des textes juridiques normalisés et publiés. La multiplicité des sources juridiques, leur technicité, leur fort degré de structuration et d'intertextualité, les enjeux de sécurité juridique imposent des contraintes très particulières en termes de recherche d'information dans ce domaine. Il faut pouvoir retrouver tous les textes en vigueur sur un sujet donné, mais également consolider un texte de loi à une date donné en prenant en compte toutes les modifications apportées à ses différents articles, ou connaître la jurisprudence relative à tel texte juridique. Il faut pouvoir savoir quelles juridictions ont tendance à modifier ou abroger tels types de textes.

Nous défendons l'idée que le développement de fonctionnalités avancées de recherche d'information dans ce domaine repose sur l'intégration de l'ensemble des propriétés documentaires dans un modèle unique. Nous proposons une ontologie documentaire (OWL) qui permet de représenter le contenu sémantique du document (ce dont parle le document), sa structure logique, ses différentes versions et son cycle de vie, ainsi que la structure de la collection documentaire qui organise différents types de documents dans un vaste réseau de liens intertextuels de documents. Il s'agit de donner accès à la complexité des sources juridiques (Bourcier, 2011). Le but est en effet de pouvoir à terme, modéliser l'ensemble d'une collection documentaire sous la forme d'un graphe RDF puis l'interroger de manière sémantique, structurelle, temporelle et/ou relationnelle à l'aide de requêtes SPARQL.

La section 2 présente les approches et les techniques proposées pour

modéliser des documents dans le web sémantique, notamment dans le domaine juridique. La section 3 décrit les besoins auxquels la recherche d'information juridique se trouve confrontée. La section 4 présente l'ontologie documentaire que nous proposons avec les différents modules la composant et leurs dépendances. La section 5 présente des exemples d'utilisation de cette ontologie pour répondre à des requêtes relationnelles.

2 Modélisation des documents dans le web sémantique

L'essor du web sémantique et du web de données repose sur l'évolution des technologies sémantiques qui assurent l'interopérabilité des données (section 2.1) mais aussi sur le développement des ressources pour l'annotation sémantique des documents (section 2.2). Dans ce contexte, un effort est fait pour développer des ontologies documentaires mais nous montrons que les modèles existants sous-estiment la dimension intertextuelle et ne permettent pas de modéliser l'ensemble des propriétés documentaires de manière homogène (section 2.3), ce qui constitue un frein à l'essor des méthodes de recherche d'information sémantique.

2.1 Langages et standards du web sémantique

Pour améliorer l'interopérabilité des données, le langage de balisage XML est utilisé pour représenter la structure des documents et de multiples modèles de documents ont été définis pour modéliser différents types de documents. Dans le domaine juridique, ces standards XML sont des éléments clés de la standardisation des contenus des sources de loi. Différents standards ont été développés par différents états européens, qui peuvent s'articuler avec le standard européen CEN-Metalex [1]. Au niveau international, un ensemble de DTD a été développé par la Chambre des représentants des Etats-Unis et le projet AKOMANTOSO [2] a produit des DTD pour les documents parlementaires, législatifs et judiciaires de plusieurs pays africains. Les détails de ces standards XML ont été décrits dans (Sartor, 2011).

Les données dans le web sémantique sont stockées sous forme de graphes de données, *c.a.d* sous la forme de triplets RDF (*RDF triple store*). Le format RDF (*Resource Description Framework* [3]) définit des déclarations

1. http ://www.metalex.eu/
2. http ://www.akomantoso.org/
3. http ://www.w3.org/RDF/

comprenant un sujet, un prédicat (*property*) et un objet, c'est-à-dire un triplet (sujet,prédicat,objet). Des URIs (*Uniform Resource Identifiers*) sont utilisés pour donner un identifiant unique au sujet, au prédicat et à l'objet.

Pour associer une sémantique aux modèles de données RDF, on peut définir les URIs dans des schémas (RDFS [4]) ou des ontologies (OWL [5]). RDFS et OWL sont des spécifications W3C. En OWL, on définit un concept comme une classe d'individus partageant les mêmes caractéristiques. Les individus sont reliés par des rôles ou des relations (*Object properties*) et ils peuvent comporter des attributs valués (*Datatype properties*). SKOS [6] (Simple Knowledge Organization System) est une famille de langages formels permettant aussi une représentation standard des thésaurus, classifications ou tout autre type de vocabulaire contrôlé et structuré.

Les triples stores RDF sont interrogés à l'aide du langage (protocole) SPARQL [7], de même que les bases de données relationnelles peuvent être interrogées à l'aide du langage SQL. SPARQL est un standard W3C et il est actuellement à sa version 1.1.

2.2 Vocabulaires conceptuels et annotation sémantique

L'approche classique de recherche d'information sémantique (comme par exemple dans AquaLog (Lopez *et al.*, 2007), KnOWLer (Ciorascu *et al.*, 2003) ou MELISA (Abasolo & Gomez, 2000)) dépasse les méthodes à base de mots clés en exploitant les annotations sémantiques qui sont apposées sur les documents pour en modéliser le contenu.

Les termes utilisés comme annotations sont définis dans des vocabulaires ou des ontologies qui sont eux-mêmes définis en SKOS ou OWL. Les ontologies de domaine permettent d'associer aux contenus des documents une description sémantique à la fois explicite et formelle, ce qui facilite l'exploitation sémantique des contenus par des outils automatiques et améliore l'interopérabilité des sources. Dans le domaine juridique, on s'appuie notamment sur des ontologies comme DOLCE (Gangemi *et al.*, 2005) ou LKIF core (Hoekstra *et al.*, 2009). Une initiative récente, Linked Open Vocabularies (LOV) [8] vise à rassembler et fournir un seul point d'en-

4. http ://www.w3.org/2001/sw/wiki/RDFS

5. http ://www.w3.org/2001/sw/wiki/OWL

6. http ://www.w3.org/2001/sw/wiki/SKOS

7. http ://www.w3.org/2001/sw/wiki/SPARQL

8. Publiée le 26/04/2013, http ://lov.okfn.org/dataset/lov/, par Mondeca, Inserm, DataLift project et Open Knowledge Foundation

trée pour les vocabulaires ouverts liés (ontologies RDFS ou OWL) utilisés dans *Linked Data Cloud*. Les vocabulaires sont listés et décrits individuellement par des métadonnées, organisés dans des classes de vocabulaires et inter-liés par le vocabulaire dédié VOAF (Vocabulary Of A Friend [9]).

Des outils d'annotation sont utilisés pour annoter sémantiquement les documents au regard d'une ontologie, c'est-à-dire pour lier certains fragments de textes (des mots, groupes de mots, phrases, etc.) à des entités de l'ontologie, le plus souvent à des instances (Amardeilh *et al.*, 2005; Uren *et al.*, 2006), mais aussi, dans certains cas, à des concepts et à des rôles (Ma *et al.*, 2013).

Le contenu d'un document ainsi que les annotations qui lui sont attachés peuvent ainsi être publiés sous forme de triplets RDF. Les annotations permettent d'identifier les entités et les concepts mentionnés dans les documents d'un domaine donné : littérature scientifique dans le domaine biomédical (Croset *et al.*, 2010) ou celui de la biodiversité (Cui *et al.*, 2010), comptes rendus hospitaliers (Minard *et al.*, 2011), etc. Dans (Mokhtari, 2010), les annotations sémantiques des documents sont stockés sous forme de triplets RDF, générés selon l'emplacement de leurs propriétés dans le texte.Dans (Croset *et al.*, 2010), la modélisation sous la forme de triplets RDF et d'URIs permet également de lier les articles scientifiques et les bases de connaissances du domaine. (Mrabet *et al.*, 2012) propose à l'inverse d'enrichir des bases de connaissances RDF/OWL en utilisant une base de documents HTML annotés par un ou plusieurs outils d'annotations.

Une fois publiées sous forme de triplets RDF, les annotations sont interrogeables par des requêtes SPARQL, même si une phase de transformation est nécessaire si la requête est formulée en langage naturel. Un système de questions réponses basé sur des patrons de requêtes (utilisés par exemple dans (Pradel *et al.*, 2012)) a été proposé comme solution intuitive et expressive au problème d'accès aux données liées publiées en RDF (Unger *et al.*, 2012).

2.3 Ontologies documentaires

Au-delà de la modélisation du contenu, des ontologies ont été produites pour modéliser les propriétés documentaires. Elles s'inspirent naturellement des langages de métadonnées définis dans la tradition des documentalistes, comme le Dublin Core. Ces ontologies sont souvent conçues pour

9. http ://lov.okfn.org/vocab/voaf/v2.2/index.html

des usages particuliers. Dans (Bouzidi *et al.*, 2011) par exemple, la modélisation doit aider la rédaction des documents réglementaires dans le domaine du bâtiment.

Ces ontologies mettent l'accent sur différents types de propriétés documentaires. L'ontologie SDO (*SALT Document Ontology* [10]) décrit la structure d'une publication scientifique, ainsi que ses propriétés identificatoires et les différentes révisions qu'elle comporte. L'ontologie d'annotation SAO (*SALT Annotation Ontology* [11]) permet de lui associer une couche d'annotation sur le contenu en lien avec des ontologies existantes, telles que FOAF, SWRC et l'ontologie bibliographique BIBO. Cette dernière (*Bibliographic Ontology* [12]) décrit en RDF des entités bibliographiques pour le web sémantique.

D'autres ontologies mettent l'accent sur le cycle de vie du document. L'ontologie PDO (*Project Documents Ontology* [13]) modélise la structure des documents de projets, en rendant compte de leurs différents statuts (rapports d'étape, rapports finaux, livrables, etc.). De la même manière, dans le domaine juridique, l'ontologie MetaLex prend en compte le statut du document (ex. document de travail) et les relations qu'ils entretiennent (`resultOf`, `generatedBy`, etc.).

L'ontologie documentaire proposée dans cet article intègre les différents types de propriétés (sémantiques, structurelles et temporelles) dans un même modèle. Elle permet aussi de rendre compte de la dimension intertextuelle qui est peu représentée dans les ontologies documentaires existantes.

3 Enjeux de la recherche d'information juridique

Notre travail se situe dans le cadre du projet Légilocal, qui vise à faciliter l'accès des citoyens aux documents juridiques des collectivités locales.

De fait, l'accès à l'information juridique est aussi problématique pour les citoyens qui essayent de comprendre la norme qui s'applique à leurs cas particulier que pour les juristes professionnels qui doivent déterminer comment la loi s'applique sur des cas de droit. Le champ du juridique pose des questions spécifiques en terme de recherche d'information.

10. http ://salt.semanticauthoring.org/ontologies/sdo
11. http ://salt.semanticauthoring.org/ontologies/sao
12. http ://uri.gbv.de/ontology/bibo/
13. http ://vocab.deri.ie/pdo-Document

En premier lieu, il est essentiel de comprendre que le tri des résultats retournés par un moteur de recherche n'est pas central dans le domaine juridique, où la recherche d'information se doit d'abord d'être exhaustive. La sécurité juridique impose en effet de prendre connaissance de tous les documents qui se rapportent à un cas particulier. Il est préférable de laisser le contrôle au juriste qui peut progressivement raffiner sa requête en fonction de ses besoins plutôt que de lui présenter un sous ensemble de documents sélectionnés en fonction d'un critère de pertinence défini *a priori*. En cela la recherche d'informations juridiques se distingue clairement des moteurs de recherche généraliste sur le web.

La structure du document est essentielle à prendre en compte. Un texte juridique, notamment le texte d'une loi, est composé d'articles qui ont un cycle de vie autonome. Ils peuvent être modifiés ou même abrogés indépendamment de la loi considérée dans son ensemble. Il est essentiel pour un juriste de pouvoir consolider un texte de loi, c'est-à-dire retrouver toutes les modifications qui s'appliquent à ce texte, et retrouver la version en vigueur à une date donnée, parce qu'il faut pouvoir déterminer le droit qui s'applique à un moment particulier du passé. Il faut également pouvoir ajuster la granularité documentaire (texte complet *vs.* article de ce texte) aux besoins de l'utilisateur et prendre en compte la complexité du cycle de vie du document juridique qui peut être signé, publié, entré en vigueur, promulgué, modifié et abrogé à des dates différentes. Les systèmes actuels d'accès à l'information juridique, comme Normattiva [14] ou UK Legislation [15], prennent partiellement en compte ce type de propriétés quand ils proposent un accès temporel aux sources juridiques (*point in time access*).

Le plus souvent cependant, dans ces systèmes, les notions de modification ou d'abrogation qui sont en réalité des relations intertextuelles sont modélisées comme des attributs de documents. On peut savoir quel est le statut d'un document juridique mais pas quel est le texte qui lui confère ce statut. La dimension intertextuelle des collections de documents juridiques est mal prise en compte. Elle est pourtant centrale dans la compréhension du raisonnement juridique : un texte ne s'interprète pas isolément, indépendamment de la jurisprudence et des interprétations auxquelles il a donné lieu, des textes qui sont venus le modifier ou des décrets qui en précisent l'application. La dimension intertextuelle des collections juridiques est reconnue comme un facteur de complexité majeur (Bourcier, 2011) pour la compréhension du droit. Ouvrir cette complexité est aujourd'hui un défi

14. http ://www.normattiva.it/ricerca/avanzata/vigente

15. http ://www.legislation.gov.uk/search/point-in-time

majeur pour l'accès à l'information juridique [16] : cela suppose de pouvoir lancer des requêtes relationnelles sur un moteur de recherche et de retrouver non pas une liste de documents autonomes mais une liste de graphes de documents qui respectent les contraintes relationnelles formulées en entrée par l'utilisateur (*Quels sont les textes de jurisprudence relatifs au texte de loi donné avant la date d'abrogation de ce dernier ?*).

Au-delà de ces besoins particuliers au domaine juridique, il faut également fournir des outils sémantiques d'accès au contenu pour permettre aux utilisateurs de retrouver des documents à partir de leurs métadonnées d'identification (date de publication, titre, type de document, numéro d'un article, etc.) mais aussi de certaines notions clés.

4 Proposition d'une ontologie documentaire

L'ontologie que nous proposons a été conçue sur la base de cette analyse des besoins. Elle permet de représenter de manière homogène toutes les informations relatives aux documents juridiques : 1) la structure d'un document (sections, paragraphes, etc.), 2) le cadre temporel dans lequel il s'inscrit, 3) la caractérisation sémantique de son contenu à l'aide de concepts ou d'entités du domaine considéré, 4) son type (loi, décret, etc.) et 5) les relations qu'il entretient avec d'autres (modification, abrogation, jurisprudence, transposition, etc.).

Notre ontologie de documents est structurée en trois grands modules qui permettent de modéliser les propriétés ci-dessus : le module document (propriétés 1 et 2), le module sémantique (propriété 3) et le module collection (propriétés 4 et 5). Les détails des classes, propriétés et attributs dans chaque module seront décrits dans ce qui suit [17].

16. Les efforts de simplification juridique actuels portent essentiellement sur la normalisation et le contrôle du lexique, à ce jour.

17. Nous avons utilisé Protégé et OWL-DL pour créer ce modèle ontologique

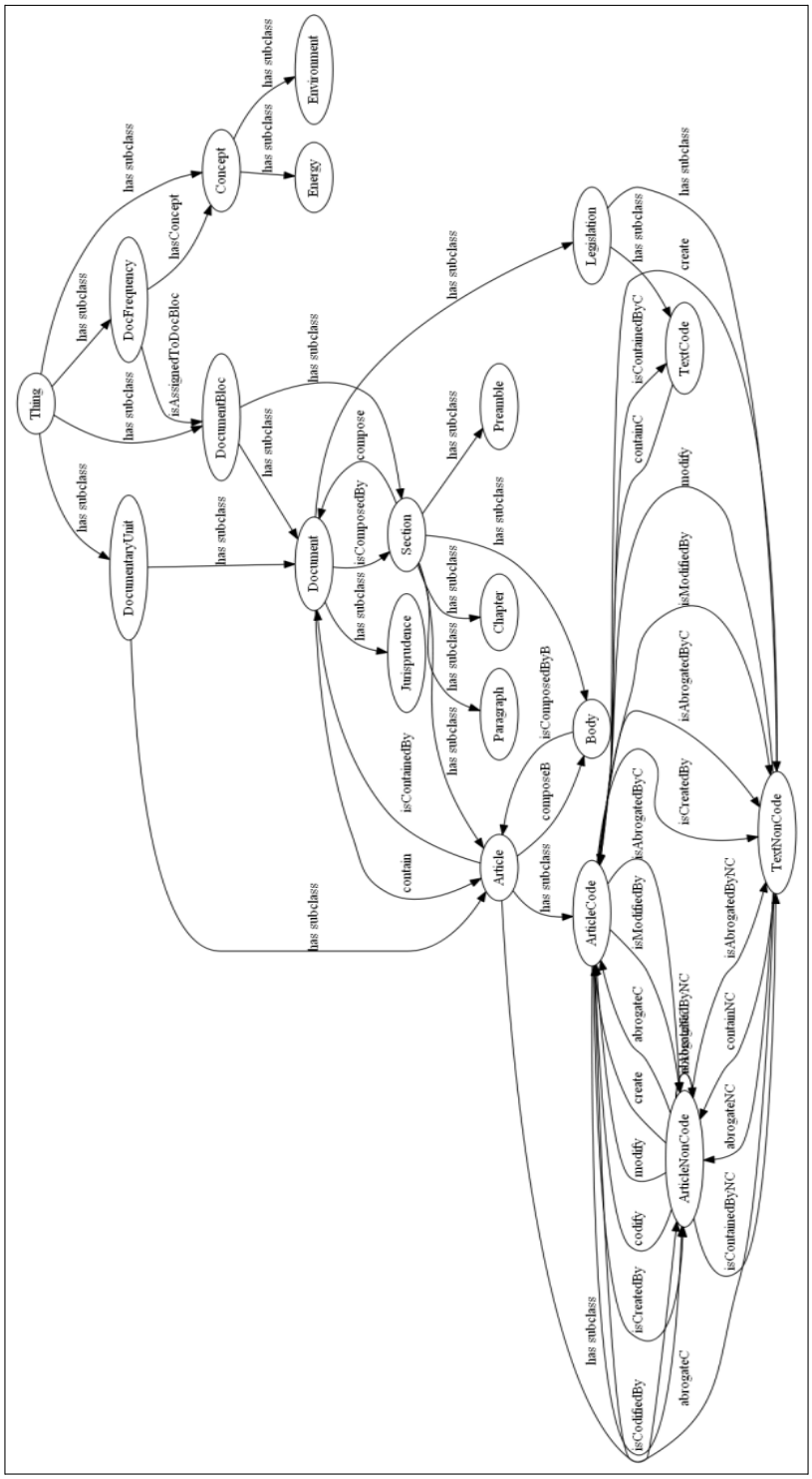


FIGURE 1 – Ontologie de collection documentaire

La figure 1 montre les différents modules de notre ontologie documentaire, sachant que la granularité de la description a été adaptée au cas d'usage Légilocal pour lequel cette ontologie a été initialement développée. Le module document est représenté par les classes `DocumentaryUnit` et `DocumentBloc`. Le module sémantique est représenté par la classe `Concept` et interagit avec le module document *via* la propriété `isAssignedToDoc`. Le module collection est représenté par l'ensemble des types de documents (`Legislation` est une sous classe de `Document`, par exemple) et un ensemble des relations intertextuelles (par ex. `modify`, `isCodifiedBy`, etc.).

Les annotations sémantiques et les relations intertextuelles peuvent porter sur n'importe quel bloc documentaire, que ce soit un document juridique complet ou un de ses composants.

4.1 Module document

Les documents juridiques possèdent une structure riche dont la sémantique est importante à prendre en compte. Les parties d'un document n'ont pas toutes la même importance : le préambule est généralement peu utile alors que les articles qui composent le texte font l'objet de requêtes particulières. L'intérêt de l'utilisateur (citoyen ou expert) porte souvent sur une partie du texte plutôt que sur le texte dans son ensemble. Cela suppose que les métadonnées d'identification et les annotations sémantiques soient attachées non pas au texte globalement mais à ses sous-parties (Hoekstra, 2011).

Le module document est représenté par les classes `DocumentaryUnit` et `DocumentBloc` et leurs sous classes `Document`, `Article` et `Section` (figure 2, figure 3). Dans le contexte de Légilocal, nous considérons en effet que l'unité documentaire de base qui peut être identifiée, qui subit des modifications et qui a un cycle de vie propre est l'article. Un article est lié au document qui le contient par les propriétés `contains` et `isContainedBy`. La structure d'un document peut être décrite plus finement par un ensemble de composants : la classe `Section` est reliée à `Document` par la propriété `compose` ; elle comporte un ensemble de sous-classes qui décrivent les différentes parties d'un document (figure 3).

Le cycle de vie du document juridique (le texte de loi mais aussi chacun des articles qui le composent) est complexe du fait des processus de modi-

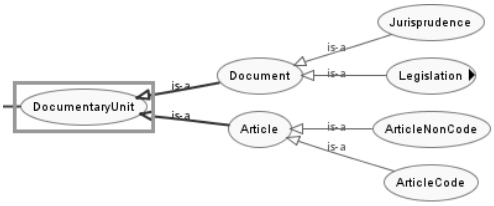


FIGURE 2 – Unité documentaire : Document ou Article

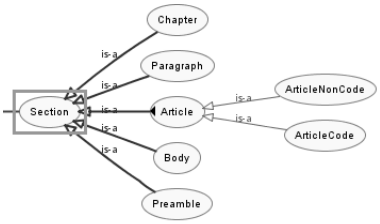


FIGURE 3 – Composants d'un document

fication des textes juridiques, de leur consolidation [18] ou de leur application par les structures gouvernementales. Plusieurs dates sont généralement associées à un document. Nous les représentons par des attributs (*Datatype properties*) dont les valeurs sont de type `date`. C'est la nature de l'attribut qui associe à la date son statut (date de création, d'entrée en vigueur, etc.).

Notons ici que nous considérons toutes les versions d'articles comme des unités documentaires différentes et que la modification d'un article est représentée par le lien existant entre l'article modifieur et l'article modifié, la date de modification étant celle de l'entrée en vigueur de l'article modifieur.

4.2 Module sémantique

Le module sémantique (voir figure 4) est classique. En pratique, on cherche généralement à réutiliser une ontologie existante. Nous définissons le prédicat `hasConcept` et `isAssignedToDoc` entre un concept du module sémantique et un bloc documentaire (document ou section). Nous avons également prévu d'associer une fonction de pondération à la

18. La consolidation consiste à intégrer dans un acte de base tous ces actes modificateurs.

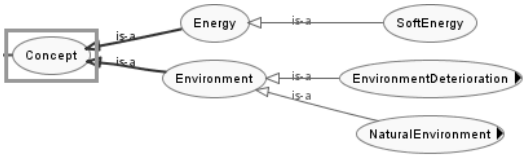


FIGURE 4 – Module sémantique : concepts du domaine

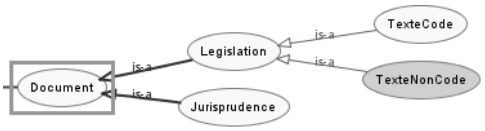


FIGURE 5 – Hiérarchie des types des documents juridiques

relation `hasConcept` pour tenir compte du nombre d'occurrences d'un concept dans un document donné. Pour modéliser cette relation ternaire nous avons créé une classe, `DocFrequency`, sur laquelle sont définis trois prédicats, `isAssignedToDoc` vers la classe bloc documentaire, `hasConcept` vers la classe concept et `hasFrequency` vers une donnée de type entier.

Comme mentionné plus haut, selon la stratégie d'annotation adoptée, les annotations sémantiques portent uniquement sur des instances du module sémantique ou renvoient aussi à des concepts et des rôles. Aucun choix n'a encore été fait à ce stade dans le projet Légilocal.

4.3 Module collection documentaire

Comme argumenté plus haut, il est essentiel de modéliser des différents types des documents juridiques. La figure 5 fait un zoom sur la hiérarchie des types de documents.

Par ailleurs, du fait de la structure hyper enchevauchée de la législation, il faut souvent consulter plusieurs textes et plusieurs versions de ces textes pour interpréter une loi. Une décision publiée comme jurisprudence doit être reliée aux textes législatifs qu'elle met en application. Il est essentiel de pouvoir modéliser cette dimension de l'intertextualité dans l'ontologie documentaire : cela doit permettre d'interroger la collection documentaire en croisant les critères sémantiques, temporels, structurels et relationnels.

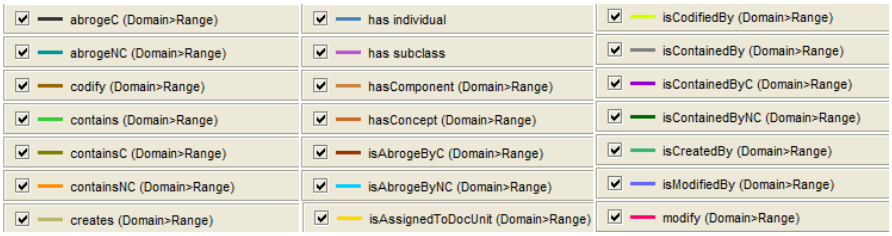


FIGURE 6 – Différents types de relations entre les entités

Dans notre ontologie, l'intertextualité est modélisée par des relations (*Object properties*) qui ont pour sujet une unité documentaire (document ou article) et pour objet une autre unité documentaire (document ou article). Par exemple le prédicat `isCreatedBy` définit la relation de création entre un article de texte non codifié [19] vers un article de texte codifié [20]. La figure 6 donne un aperçu des types de relations que nous avons codés dans ce module.

Nous précisons que nous ne prenons pas en compte ici les visas dit "de forme", c'est-à-dire les références qui figurent dans le préambule des documents juridiques, et qui ont généralement une valeur très générale, comme les références au Code Civil, par exemple.

5 Interrogation

La modélisation d'une collection juridique revient à instancier l'ontologie [21] en produisant un ensemble de triplets RDF : sont ainsi modélisés les documents et leurs types (`Legislation`, `Jurisprudence`, etc), les articles (`ArticleCode`, `ArticleNonCode`), les concepts du domaine (`Energy`, `Environment`, etc.), les relations entre classes (`hasConcept`, `isModifiedBy`, etc). La collection est ainsi représentée comme une base de connaissances qui peut ensuite être interrogée à l'aide de requêtes SPARQL.Ces requêtes peuvent porter sur :

19. Un article non codifié est un article qui appartient à un texte réglementaire autre que les codes : articles de loi, etc.

20. Un article codifié est un article qui appartient à un code : code civil, code de l'environnement.

21. Dans le cadre du projet, le peuplement de l'ontologie se fera automatiquement avec les résultats des outils d'analyse de documents (ex. structuration, repérage de références dans le texte, annotation sémantique).

- le contenu sémantique d'un type donné de documents : la requête *Quels sont les textes qui parlent de la préservation de l'environnement ?* porte par exemple sur les classes `Concept`, et `DocumentBloc` (le concept `DocFrequency` peut également être pris en compte) ;
- l'historique d'une unité documentaire, qu'il s'agisse d'un document ou de l'un de ses articles : la requête *Comment a été abrogé l'article 22 de la loi sur l'enseignement obligatoire ?* fait appel à la classe `Article`, à la propriété `estAbrogéPar` et doit retourner tous les documents qui ont abrogé l'article 22 en question (si la version de l'article 22 considérée n'est pas précisée, tous les textes abrogatifs doivent être retournés) ;
- les types de documents et les types des liens : la requête *Donnez moi les jurisprudences qui ont appliqué l'article 4 actuellement en vigueur de la loi Sapin ?* porte par exemple sur les classes `Jurisprudence`, `ArticleNonCode` (article 4), `TexteNonCode` (loi Sapin), sur la propriété `dateVigueur` de la classe `ArticleNonCode` et sur la propriété `appliedBy` (non encore présentée dans le modèle) reliant les textes de jurisprudence aux textes de loi.

Nous pouvons aussi interroger sur la consolidation d'un texte (mise à jour d'un texte de loi : décret, loi, etc.) à une date donnée. Cela suppose un calcul un peu plus compliqué, puisqu'il faut partir de la structure du texte, identifier la liste des articles qui le composent et retrouver pour chacun la version en vigueur à la date considérée.

Nous avons collecté un ensemble de requêtes utilisateurs suite à une analyse de besoins menée auprès des juristes. En étudiant ces requêtes nous avons constaté qu'elles présentent des structures récurrentes sur lesquelles nous pouvons définir des patrons de requêtes. Cette particularité nous semble un atout en faveur d'une interface d'interrogation plus facile permettant aux utilisateurs d'exprimer leurs besoins en langage naturel puis de faire la traduction de ces requêtes sous forme de graphes (exprimés en SPARQL). Cette solution n'est pas envisageable à ce stade de notre étude. Nous comptons à court terme définir des interfaces d'interrogation en langage contrôlé ou sous forme de formulaires.

6 Discussion

Le modèle ontologique que nous avons présenté peut être affiné mais la version présentée ici correspond au degré de modélisation requis par le projet Légilocal. Ce modèle est en passe d'être adopté par l'ensemble des par-

tenaires : il permet déjà d'élargir largement le cadre de la recherche d'information sémantique à la recherche d'information fine et à la recherche relationnelle. L'ontologie proposée a été définie en s'appuyant sur une expertise du domaine. Sa validation se fera en deux étapes : étape du peuplement avec des documents du projet, étape du déploiement du moteur de recherche et de son évaluation par des communes (ou comités de public). L'ontologie que nous proposons est utilisable par toute autre application de gestion documentaire dans le domaine juridique comme par exemple l'aide à la publication ou la consolidation. Les prochaines étapes de ce travail consistent à instancier l'ontologie sur le corpus de documents de travail du projet Légilocal, à produire les triplets RDF à partir de ces données et à les stocker dans des triples stores, mais aussi, à concevoir un module de traduction des requêtes utilisateurs en requêtes SPARQL et une interface d'interrogation et de visualisation des résultats.

Références

ABASOLO J. M. & GOMEZ M. (2000). MELISA. An ontology-based agent for information retrieval in medicine. In *Proceedings of the First International Workshop on the Semantic Web (SemWeb2000)*, p. 73–82.

AMARDEILH F., LAUBLET P. & MINEL J.-L. (2005). Document annotation and ontology population from linguistic extractions. In *Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05)*, p. 161–168.

BOURCIER D. (2011). Sciences juridiques et complexité. Un nouveau modèle d'analyse. *Droit et Cultures*, **61**(1), 37–53.

BOUZIDI K. R., FARON-ZUCKER C., FIES B., CORBY O. & NHAN L.-T. (2011). Modélisation de documents réglementaires dans le domaine du bâtiment. In *Actes 12e Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance, EGC 2011*, Bordeaux, France.

CIORASCU C., CIORASCU I. & STOFFEL K. (2003). KnOWLer - Ontological Support for Information Retrieval Systems. In *Proceedings of 26th Annual International ACM SIGIR Conference, Workshop on Semantic Web*.

CROSET S., GRABMÜLLER C., LI C., KAVALIAUSKAS S. & REBHOLZ-SCHUHMANN D. (2010). The CALBC RDF Triple Store : retrieval over large literature content. *CoRR*, **abs/1012.1650**.

CUI H., JIANG K. Y. & SANYAL P. P. (2010). From text to RDF triple store : an application for biodiversity literature. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, volume 47, p. 1–2 : American Society for Information Science.

GANGEMI A., SAGRI M.-T. & TISCORNIA D. (2005). A Constructive Framework for Legal Ontologies. In V. BENJAMINS, P. CASANOVAS, J. BREUKER

& A. GANGEMI, Eds., *Law and the Semantic Web*, volume 3369 of *Lecture Notes in Computer Science*, p. 97–124. Springer Berlin Heidelberg.

HOEKSTRA R. (2011). The METALEX document server : legal documents as versioned linked data. In *Proceedings of the 10th International Conference on the Semantic Web*, ISWC'11, p. 128–143, Berlin, Heidelberg : Springer-Verlag.

HOEKSTRA R., BREUKER J., BELLO M. D. & BOER A. (2009). LKIF Core : Principled Ontology Development for the Legal Domain. In *Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web : Channelling the Legal Information Flood*, p. 21–52, Amsterdam : IOS Press.

LOPEZ V., UREN V., MOTTA E. & PASIN M. (2007). AquaLog : An ontology-driven question answering system for organizational semantic intranets. *Web Semant.*, **5**(2), 72–105.

MA Y., LÉVY F. & NAZARENKO A. (2013). Annotation sémantique pour des domaines spécialisés et des ontologies riches. In *Actes de la 20ème conférence du Traitement Automatique du Langage Naturel (TALN 2013)*.

MINARD A.-L., LIGOZAT A.-L. & GRAU B. (2011). Extraction de relations dans des comptes rendus hospitaliers. In *22es Journées Francophones d'Ingénierie des Connaissances*, IC 2011, p. 491–506, Chambéry, France.

MOKHTARI N. (2010). *Extraction et exploitation d'annotations sémantiques contextuelles à partir de texte*. PhD thesis, Université Sophia Antipolis.

MRABET Y., BENNACER N. & PERNELLE N. (2012). Enrichissement contrôlé de bases de connaissances à partir de documents semi-structurés annotés. In *23es Journées Francophones d'Ingénierie des Connaissances*, IC 2012, Paris.

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2012). Des patrons modulaires de requêtes SPARQL dans le système SWIP. In *23es Journées Francophones d'Ingénierie des Connaissances*, IC 2012, Paris, France.

SARTOR G. (2011). *Law, Governance and Technology : Legislative Xml for the Semantic Web : Principles, Models, Standards for Document Management*. Law, Governance and Technology Series, 4. Springer London, Limited.

UNGER C., BÜHMANN L., LEHMANN J., NGONGA A.-C. N., GERBER D. & CIMIANO P. (2012). Template-based question answering over RDF data. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, p. 639–648 : ACM.

UREN V., CIMIANO P., IRIA J., HANDSCHUH S., VARGAS-VERA M., MOTTA E. & CIRAVEGNA F. (2006). Semantic Annotation for Knowledge Management : Requirements and a Survey of the State of the Art. *Journal of Web Semantics*, **4**.

Une mesure de Similarité Sémantique basée sur la Recherche d'Information

Davide Buscaldi[1]

Laboratoire d'Informatique de Paris Nord,
CNRS, (UMR 7030)
Université Paris 13, Sorbonne Paris Cité,
F-93430, Villetaneuse, France
davide.buscaldi@lipn.univ-paris13.fr

Résumé : Dans cet article, nous décrivons une mesure de similarité sémantique basée sur la recherche d'information qui a été utilisée dans le "shared task" SemEval 2013. Habituellement, la similarité sémantique est utilisée pour améliorer la performance des systèmes de recherche d'information. Nous avons essayé de faire le contraire : deux textes à comparer sont traités comme des requêtes et on compare les listes des résultats de la recherche afin d'établir la similarité sémantique des textes. Les résultats préliminaires obtenus dans le "shared task" SemEval 2013 montrent que cette mesure a fonctionné mieux que les mesures de similarité sémantique basées sur WordNet et des mesures classiques comme le cosinus ou la distance d'édition. **Mots-clés** : Similarité Sémantique. Recherche d'Information.

1 Introduction

La détermination du niveau de similarité sémantique entre textes est une tâche très importante pour différentes applications dans le contexte du Traitement Automatique de la Langue (TAL). Une des premières applications de similitude de texte est, peut-être, le modèle d'espace vectoriel utilisé dans la Recherche d'Information (RI), où il faut classer des documents dans une collection par ordre de pertinence par rapport à une requête en entrée (Salton (1971)). La pertinence est calculée en fonction d'une mesure de similarité entre la requête et le document, dont le contenu est représenté par un vecteur de mots. Les mesures de similarité de texte ont été également utilisées pour la désambiguïsation lexicale (Lesk (1986)) et le résumé automatique de texte (Lin & Hovy (2003)), entre autres.

Très récemment, une évaluation comparative pour la tâche de similarité sémantique a été proposée en tant que "pilot task" au SemEval 2012 (Semantic Textual Similarity, STS) par Agirre *et al.* (2012) et confirmée pour SemEval 2013, élevée ainsi au rang de "shared task" du *SEM2013 [1], dans le but d'établir un cadre d'évaluation qui favorise le développement de nouveaux systèmes et mesures de similarité sémantique. L'objectif final des campagnes d'évaluation SemEval est de promouvoir la recherche dans tous les aspects de la sémantique computationnelle et d'améliorer les performances dans toutes les tâches où il est nécessaire d'analyser sémantiquement un texte de façon automatique. Il s'agit de déterminer, sur une échelle de 0 (aucune similarité) à 5 (même signification), la similarité sémantique entre couples de phrases. Cette particularité distingue la tâche STS de la tâche de détection de paraphrases, dont la réponse à la question si deux phrases sont similaires est une réponse binaire.

Dans cet article, nous présentons une mesure de similarité sémantique qui utilise la RI pour déterminer la similarité entre textes ; cette mesure est basée sur l'hypothèse que, si on utilise deux textes en tant que requêtes d'entrée dans un même système de RI, alors leur similarité dépend de la similarité des listes des documents récupérés par le système. Cette mesure a été vérifiée expérimentalement dans le cadre de la participation de l'équipe RCLN du LIPN au STS du SemEval 2013 (Buscaldi *et al.* (2013)), où elle s'est révélée la mesure la plus efficace, parmi celles mis en œuvre par le système LIPN. La suite du papier est articulée comme suit : dans la Section 2, nous présentons des travaux connexes. Dans la Section 3, nous décrivons notre mesure de similarité basée sur la RI. Dans la Section 4, nous rendons compte de l'expérimentation et de l'évaluation menées. Enfin, dans la Section 5, nous concluons et nous proposons quelques perspectives.

2 Travaux connexes

Traditionnellement, les mesures de similarité sémantique ont été définies sur des mots ou des concepts, mais non sur des fragments de texte. La dérivation d'une mesure de similarité sémantique compositionnelle à partir d'une mesure de similarité sémantique lexicale (c'est à dire, déterminer la similarité entre deux phrases à partir des mots qui les composent) n'est pas triviale. Par exemple, les phrases « un chien mord un homme » et « un homme mord un chien » ont la même similarité sémantique mot-à-mot mais sont très différentes globalement. L'accent mis sur les mesures de

1. `http://clic2.cimec.unitn.it/starsem2013/`

similarité mot-à-mot est probablement dû à la disponibilité de ressources sémantiques telles que WordNet ou des ontologies qui codent des relations concept-à concept. Pedersen *et al.* (2004) a implémenté et mis à disposition un *package*, WordNet :Similarity [2], qui code différentes mesures de similarité mot-à-mot, souvent utilisées dans plusieurs applications de TAL, par exemple la détection de l'implication textuelle (Harabagiu & Hickl (2006)) ou la résolution de coréférences (Ponzetto & Strube (2006)). La plupart des systèmes qui participent à SemEval STS ont utilisé des mesures classiques, modèle vectoriel, des extensions du modèle de n-grammes (Buscaldi *et al.* (2012)), la distance d'édition, ou des combinaisons de mesures de similarité sémantique lexicale (Banea *et al.* (2012); Bär *et al.* (2012); Šarić *et al.* (2012)). Les meilleurs résultats ont été obtenus en intégrant les différentes mesures avec des méthodes de régression linéaire (Schölkopf *et al.* (1999)).

3 Mesure de Similarité Basée sur la Recherche d'Information

Étant donnés deux textes p and q, un système de RI S et une collection de documents D indexé par S, cette mesure est basée sur l'hypothèse que p et q sont sémantiquement similaires si les documents récupérés par S pour les deux textes utilisés en tant que requêtes d'entrée, sont classés de façon similaire.

Soient $L_p = \{d_{p_1}, \ldots, d_{p_K}\}$ et $L_q = \{d_{q_1}, \ldots, d_{q_K}\}$, $d_{x_i} \in D$ les ensembles des premiers K documents récupérés par S pour p et q, respectivement. Soient $s_p(d)$ et $s_q(d)$ les scores assignés par S au document d pour les requêtes p et q, respectivement. Le score de similarité est calculé de la façon suivante :

$$sim_{IR}(p,q) = 1 - \frac{\sum\limits_{d \in L_p \cap L_q} \frac{\sqrt{(s_p(d) - s_q(d))^2}}{\max(s_p(d), s_q(d))}}{|L_p \cap L_q|} \quad (1)$$

si $|L_p \cap L_q| \neq \emptyset$, 0 en cas contraire.

La valeur optimale de K a été déterminée avec des expériences menées sur le jeux de données de test du SemEval-2012, composée par 3108 couples de phrases en anglais, avec des jugements de similarité gradués entre 0 et 5. Le score de similarité calculé avec la formule (1) est toujours compris entre 0 et 1, donc pour pouvoir calculer la corrélation avec le gold standard (jugements de similarité), le score calculé pour chaque couple de phrases

2. `http://wn-similarity.sourceforge.net/`

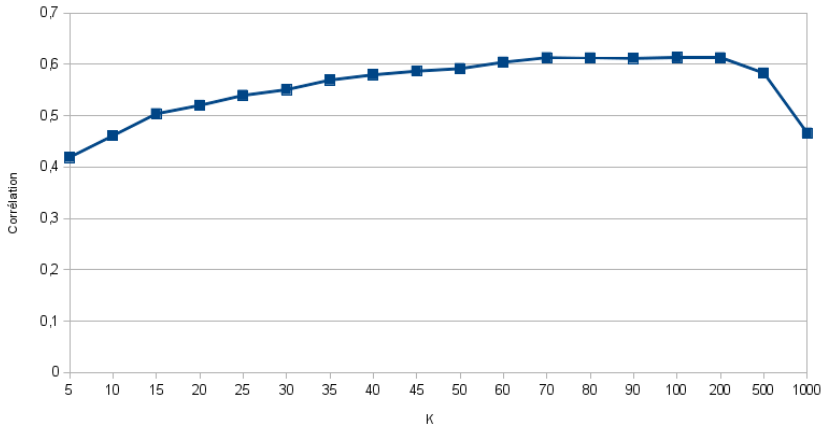


FIGURE 1 – Coefficient de corrélation de Pearson pour différentes valeurs de K, calculés sur le jeu de données SemEval 2012.

a été multiplié par 5. Les résultats en Figure 1 montrent que la valeur optimale se situe entre 70 et 100. Nous avons utilisé une valeur $K = 80$

4 Expérimentation

La mesure de similarité a été testée sur les jeux de données SemEval-2012 et SemEval-2013, et comparée avec des mesures de similarité habituellement utilisées dans la tâche STS, décrites dans la Section 4.1. La tâche STS consiste à comparer deux phrases et déterminer leur similarité sémantique sur une échelle de 0 à 5. Chaque couple de phrases a été jugé par 4 experts, le jugement de similarité du gold standard est la moyenne de tous les 4 jugements. La Figure 2 présente des exemples de phrases à comparer, avec le jugement de corrélation donné par les experts.

La collection de documents utilisée pour les expériences est en anglais et composée par la collection AQUAINT-2 [3] et la collection NTCIR-8 [4]. Le moteur de recherche utilisé est Lucene [5], plus précisément dans sa version 4.2, et la mesure de similarité utilisée pour calculer la similarité entre les requêtes et les documents est la mesure BM25 (Jones *et al.* (2000)). La

3. `http://www.nist.gov/tac/data/data_desc.html#AQUAINT-2`
4. `http://metadata.berkeley.edu/NTCIR-GeoTime/ntcir-8-databases.php`
5. `http://lucene.apache.org/core`

Jugement	Phrase 1	Phrase 2
4.8	Tehran: Discussions with Venezuelan President Hugo Chavez	city of tehran : conversations with president of venezuela hugo chavez
3.8	Anti-Putin protesters form human chain in Moscow	Anti-Putin protesters form human chain
2.6	Egyptians vote to pick president for first time	Egyptians wait for key presidential vote results
1	Dozens killed in Nigerian riots	Dozens killed in Kenyan clashes
0	20-member parliamentary, trade team leaves for India	Commerce secretary to take leave of absence

FIGURE 2 – Examples de phrases de la tâche STS.

valeur K a été mis à 20 pour les expériences du SemEval-2013, pour des raisons de rapidité (il y a un temps limite pour produire les résultats).

4.1 Mesures de similarité pour comparaison

4.1.1 Cosinus

Soient $\mathbf{p} = (w_{p_1}, \ldots, w_{p_n})$ et $\mathbf{q} = (w_{q_1}, \ldots, w_{q_n})$ les vecteurs de poids $tf.idf$ associés aux phrases p et q, alors la distance cosinus est :

$$sim_{cos}(\mathbf{p}, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{p_i} \times w_{q_i}}{\sqrt{\sum_{i=1}^{n} {w_{p_i}}^2} \times \sqrt{\sum_{i=1}^{n} {w_{q_i}}^2}} \tag{2}$$

Où la valeur idf (inverse document frequency) a été calculée sur Google Web 1T.

4.1.2 Distance d'édition

La distance d'édition a été calculé de la façon suivante :

$$sim_{ED}(p, q) = 1 - \frac{Lev(p, q)}{\max(|p|, |q|)} \tag{3}$$

où $Lev(p, q)$ est la distance de Levenshtein entre p et q, au niveau des caractères.

4.1.3 Mesure de similarité basée sur WordNet

Cette mesure de distance a été introduite pour Buscaldi *et al.* (2012). Soient C_p et C_q les ensembles de concepts contenus dans les phrases p et q, avec $|C_p| \geq |C_q|$, alors la similarité entre p et q est me résultat de :

$$sim_{WN}(p,q) = \frac{\sum\limits_{c_1 \in C_p} \max\limits_{c_2 \in C_q} s(c_1, c_2)}{|C_p|} \tag{4}$$

où $s(c_1, c_2)$ est une mesure de similarité conceptuelle calculée sur la hiérarchie de WordNet, selon la formulation de Wu & Palmer (1994).

4.2 Mesure de similarité basée sur N-grammes

Cette mesure a été proposée par Buscaldi *et al.* (2009) pour la tâche de recherche de passages dans le contexte des systèmes question-réponse. La similarité entre p et q est calculée de la façon suivante :

$$sim_{ngrams}(p,q) = \frac{\sum\limits_{\forall x \in Q} h(x,P) \frac{1}{d(x, x_{max})}}{\sum_{i=1}^{n} w_i} \tag{5}$$

où w_i est le poids idf du mot t_i, P est l'ensemble des n-grammes composés par les mots de p qui sont aussi dans q , Q est l'ensemble de tous les n-grammes possibles dans q, et n le nombre de mots dans la phrase la plus longue. $\frac{1}{d(x,x_{max})}$ est un facteur de distance qui réduit le poids des n-grammes en fonction de la distance du plus grand n-gramme. Le poids pour un n-gramme est calculé comme la somme des poids des mots qui le composent :

$$h(x, P_j) = \begin{cases} \sum_{k=1}^{j} w_k & \text{if } x \in P_j \\ 0 & \text{autrement} \end{cases} \tag{6}$$

où w_k est le poids du *k-ième* mot et *j* la taille du n-gramme x ;

4.3 Évaluation

Le coefficient de corrélation de Pearson a été calculé pour chaque mesure, sur les jeux de données de test du SemEval-2012 (Table 1) et du SemEval-2013 (Table 2). Le jeu de données de test SemEval-2012 est composé par différents sous-ensembles de phrases extraites de différents

corpus : *MSRpar* est composé par des couples de phrases extraites du corpus de paraphrases de Microsoft[6] ; *MSRvid* est un corpus composé par des annotations de vidéos ; *OnWN* est composé par des définitions de WordNet ; *Europarl* est composé par des transcriptions de traductions automatiques des sessions du parlement européen ; *News*, *Headlines* sont des corpus composés par des titres d'actualités. Le jeu de données SemEval-2013 contient aussi des définitions de frames de FrameNet (*FNWN*) et des phrases issues de l'évaluation automatique de systèmes de traduction automatique (*SMT*). La Table 4 présente la taille, en nombre moyen de mots pour chaque phrase, dans chaque jeu de données.

	MSRpar	**MSRvid**	**OnWN**	**Europarl**	**News**	**ALL**
sim_{ngrams}	0.419	0.543	0.453	0.505	0.408	0.412
sim_{WN}	0.380	0.784	**0.507**	0.556	0.426	0.609
sim_{ED}	0.251	0.290	**0.507**	**0.625**	0.426	0.300
sim_{cos}	**0.468**	0.688	0.458	0.556	0.349	0.513
sim_{IR}	0.167	**0.785**	0.359	0.584	**0.523**	**0.613**

TABLE 1 – Coefficient de corrélation de Pearson, calculé sur le test set SemEval 2012.

	FNWN	**Headlines**	**OnWN**	**SMT**	**ALL**
sim_{ngrams}	0.285	0.532	0.459	0.280	0.336
sim_{WN}	**0.395**	**0.606**	0.552	0.282	0.477
sim_{ED}	0.220	0.536	0.089	**0.355**	0.230
sim_{cos}	0.306	0.573	0.541	0.244	0.382
sim_{IR}	0.067	0.598	**0.628**	0.241	**0.541**

TABLE 2 – Coefficient de corrélation de Pearson, calculé sur le test set SemEval 2013.

La Table 3 présente les résultats du test d'ablation mené sur le système utilisé pour la participation au SemEval-2013, où chaque mesure a été utilisée en tant que caractéristique pour entraîner un modèle de régression linéaire.

Dans tous les tests on peut observer que la mesure basée sur la RI, en moyenne, s'est montrée meilleure, surtout pour les corpus basés sur les

6. `http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/`

Mesure enlevée	**Corrélation**	**Perte**
aucune	0.597	0
sim_{ngrams}	0.596	0.10%
sim_{WN}	0.563	3.39%
sim_{ED}	0.584	1.31%
sim_{cos}	0.596	0.10%
sim_{IR}	0.510	**8.78%**

TABLE 3 – Test d'ablation sur différentes mesures utilisées dans le test set SemEval-2013.

corpus	**# moyen de mots**
MSRpar	17.71
MSRvid	6.63
OnWN 2012	7.5
Europarl	10.7
News	11.72
FNWN	19.92
Headlines	7.21
OnWN 2013	7.17
SMT	26.39

TABLE 4 – Nombre moyen de mots pour chaque phrase, dans chaque corpus.

actualités (c'est à dire, dans le même domaine du corpus indexé par le moteur de recherche), même si pour certains sous-ensembles de données elle a obtenu des résultats non satisfaisants, en particulier sur les données MSRpar et FNWN. On a supposé que la taille moyenne des phrases joue un rôle dans la qualité des résultats ; apparemment, c'est le cas pour MSRpar et FNWN qui ont une taille en moyenne double que la plupart des autres corpus (Table 4), mais ce n'est pas si évident dans le cas du corpus SMT, où la perte de performances ne semble pas en corrélation avec la taille des phrases.

5 Conclusion et perspectives

Dans ce papier, nous avons présenté une mesure de similarité sémantique entre deux textes à partir d'une collection de documents indexé avec un système de RI. L'évaluation de cette mesure sur les jeux de données de la

tâche SemEval STS 2012 et 2013 montre qu'elle obtient, en moyenne, des meilleures valeurs de corrélation par rapport à d'autres mesures de similarité entre textes, basées soit sur des caractéristiques de surface du texte, soit sur la similarité conceptuelle entre concepts. Cependant, cette mesure montre des limites, surtout si la taille des textes à comparer augmente, et si le corpus de documents indexés n'est pas dans le même domaine que les textes à comparer. Nous souhaitons, en perspective, vérifier si la mesure est indépendante du modèle de RI utilisé par le moteur de recherche, et implémenter la mesure sur le web, pour résoudre le problème lié au domaine du corpus indexé. Dans ce cas, un problème additionnel à résoudre est constitué par le poids à donner à chaque document récupéré.

Remerciements

Le travail présenté a été en partie soutenu par le LabEx EFL (Empirical Foundations of Linguistics) et par le projet OSEO Quaero.

Références

AGIRRE E., CER D., DIAB M. & GONZALEZ A. (2012). A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantcis (*SEM 2012)*, Montreal, Quebec, Canada.

BANEA C., HASSAN S., MOHLER M. & MIHALCEA R. (2012). UNT : A Supervised Synergistic Approach to Semantic Text Similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, p. 635–642, Montreal, Canada.

BÄR D., BIEMANN C., GUREVYCH I. & ZESCH T. (2012). UKP : Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, p. 435–440, Montreal, Canada.

BUSCALDI D., ROSSO P., GÓMEZ J. M. & SANCHIS E. (2009). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems (JIIS)*, **34**(2), 113–134.

BUSCALDI D., ROUX J. L., FLORES J. G. G. & POPESCU A. (2013). LIPN-CORE : Semantic Text Similarity using n-grams, WordNet, Syntactic Analysis, ESA and Information Retrieval based Features. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, GA, USA. to appear.

BUSCALDI D., TOURNIER R., AUSSENAC-GILLES N. & MOTHE J. (2012). Irit : Textual similarity combining conceptual similarity with an n-gram comparison method. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Quebec, Canada.

DUDOGNON D., HUBERT G. & RALALASON B. (2010). Proxigénéa : Une mesure de similarité conceptuelle. In *Proceedings of the Colloque Veille Stratégique Scientifique et Technologique (VSST 2010)*.

HARABAGIU S. & HICKL A. (2006). Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, p. 905–912, Stroudsburg, PA, USA : Association for Computational Linguistics.

JONES K. S., WALKER S. & ROBERTSON S. E. (2000). A probabilistic model of information retrieval : development and comparative experiments part 2. *Inf. Process. Manage.*, **36**(6), 809–840.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, p. 24–26, New York, NY, USA : ACM.

LIN C.-Y. & HOVY E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, p. 71–78, Stroudsburg, PA, USA : Association for Computational Linguistics.

PEDERSEN T., PATWARDHAN S. & MICHELIZZI J. (2004). WordNet : :Similarity : measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, p. 38–41, Stroudsburg, PA, USA : Association for Computational Linguistics.

PONZETTO S. P. & STRUBE M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, p. 192–199, Stroudsburg, PA, USA : Association for Computational Linguistics.

SALTON G. (1971). *The SMART Retrieval System&#8212 ;Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

SCHÖLKOPF B., BARTLETT P., SMOLA A. & WILLIAMSON R. (1999). Shrinking the tube : a new support vector regression algorithm. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, p. 330–336, Cambridge, MA, USA : MIT Press.

ŠARIĆ F., GLAVAŠ G., KARAN M., ŠNAJDER J. & BAŠIĆ B. D. (2012). TakeLab : Systems for Measuring Semantic Text Similarity. In *Proceedings of*

the 6th International Workshop on Semantic Evaluation (SemEval 2012), p. 441–448, Montreal, Canada.

WU Z. & PALMER M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, p. 133–138, Stroudsburg, PA, USA : Association for Computational Linguistics.