Exploiting and Extending a Semantic Resource for Conceptual Indexing

Karam ABDULAHHAD* — Jean-Pierre CHEVALLET — Catherine BERRUT***

** UJF-Grenoble 1, ** UPMF-Grenoble 2, LIG laboratory, MRIM group*

{karam.abdulahhad,jean-pierre.chevallet,catherine.berrut}@imag.fr

ABSTRACT. *Information Retrieval Systems that compute a matching between a document and a query based on terms intersection, cannot reach relevant documents that do not share any terms with the query. The objective of this study is to propose a solution to this problem in the context of conceptual indexing. We study an ontology-based matching that exploits links between concepts. We propose a model that exploits the weighted links of an ontology. We also propose to extend the links of the ontology to reflect the structural ambiguity of some concepts. A validation of our proposal is made on the test collection ImagCLEFMed 2005 and the external resource UMLS 2005.*

RÉSUMÉ. *Les Systèmes de Recherche d'Information qui calculent la correspondance entre un document et une requête à base d'intersection de termes, ne peuvent pas atteindre les documents pertinents qui ne partagent aucun termes avec la requête. L'objectif de ce travail de master est alors de proposer une solution à ce problème dans le cadre d'une indexation par concepts. Nous étudions une correspondance basée sur une ontologie qui exploite les liens entre les concepts. Nous proposons un modèle de correspondance qui exploite la pondération des liens de l'ontologie. Nous proposons également d'étendre les liens de l'ontologie pour tenir compte de l'ambigüité de structure de certains concepts. Une validation de notre proposition est effectuée sur la collection de test ImagCLEFMed 2005 et la ressource externe UMLS 2005.*

KEYWORDS: *term mismatch, concept mismatch, Bayesian matching, conceptual indexing*

MOTS-CLÉS : *variation terminologique, correspondance Bayésien, indexation conceptuelle*

1. Term and concept mismatch

Information Retrieval Systems IRSs based on term [1] intersection to compute a matching between a document and a query, suffer from *term mismatch* problem. This problem appears when users write a query using terms different from terms in relevant document. For example, the following two terms '*Skin Cancer*' and '*melanoma*' have a close meaning in a medical context. In less than 20% of cases, two people use the same term to describe the same meaning (Crestani, 2000). So without an external resource that links these two terms, we cannot retrieve a document containing '*Skin Cancer*' as a response to a query containing '*melanoma*'.

To solve the term mismatch problem the first step is: *using concepts* [2] *instead of terms* (Chevallet *et al.,* 2007). Using concepts solves a part of the problem when different terms correspond to the same concept, e.g. the two terms "*Atrial Fibrillation*" and "*Auricular Fibrillation*" correspond to the same concept "*C0004238*" in UMLS [3]. However, when two terms corresponds to two concepts, and these two concepts have a relation, this relation must be used for matching. For example, the two terms "*B-Cell*" and "*Lymphocyte*" correspond to the two concepts "*C0004561*" and "*C0024264*" respectively, and there is a relation of type "*isa*" between these two concepts. Here, and without exploiting the relations between concepts, we get the same problem but at the conceptual level "*concept mismatch*".

The previous problem can be solved by using conceptual relations during the matching process (Le, 2009).

Using concepts and conceptual relations supposes the existence of external resources that encompass them. However, external resources are *incomplete*. We found out that many potential relations based on the syntax of terms are missing. For example, in UMLS there are five concepts containing the word "*spirochaete*", so we have twenty pairs of concepts that potentially have a linguistic relation, but we did not find any relation (see Table 1).

In this work, we propose to enrich the external resource by adding more relations between concepts, and in this way we hope to enhance system's recall. We must, however, be careful in building and using relations, because building too many relations may decrease precision.

---

1. A term is a noun phrase that has a unique meaning in a specific domain (e.g. medical domain) and that belongs to a terminology (Baziz, 2005) (Chevallet, 2009).

2. "Concepts" can be defined as "Human understandable unique abstract notions independent from any direct material support, independent from any language or information representation, and used to organize perception and knowledge" (Chevallet *et al.,* 2007). In IR domain, to achive the conceptual indexing, each concept is associated to a set of terms that describe it (Baziz, 2005) (Chevallet, 2009).

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

Table 1. *Statistics from UMLS*

word	# concepts	# all concepts pairs	# pairs with relation
device	86,985	7,566,303,240	161,660
activity	22,395	501,513,630	380,052
sedum	98	9,506	122
spirocheate	5	20	0

2. Proposed model

Our model bases on using concepts as indexing elements, enriching external resource by new relations, and then using these relations at matching time. The model consists of three main components:

1) External Resource: The external resource is used in conceptual indexing to map text to concepts. It contains terms $T = \{t_1, t_2, \ldots\}$, concepts $C = \{c_1, c_2, \ldots\}$ linked by relations $R = \{r | r \subseteq C \times C\}$. Each concept corresponds to several terms, then a function can be defined: $\zeta : C \rightarrow T^*$ where T^* is the set of all subsets of T. We enrich the external resource by:
a) Adding relations between concepts: e.g. "*shared-words*": this relation means that there are words in common between two concepts. Formally, there is a shared-words relation between two concepts $C_1 = \{t_1^{c1}, t_2^{c1}, \ldots, t_k^{c1}\}$ and $C_2 = \{t_1^{c2}, t_2^{c2}, \ldots, t_l^{c2}\}$ iff $NSW > 0$, where: $NSW = |C_1 \cap C_2|$ the number of shared words between the two concepts. NSW could be calculated because each concept corresponds to terms and each term corresponds to a sequence of words.
b) Defining a Certainty property to distinguish relations already defined in the external resource R_C from relations added by us $R_{\neg C}$. The Certainty represents how much we are sure that there is a semantic relation between two concepts. The hypothesis here is: if there is a document d contains a concept c_d, a query q contains a concept c_q, and if there is a relation of type R_C (e.g. *isa*) between c_d and c_q. Then it is more probable that d is relevant document for q than if the relation between c_d and c_q is of type $R_{\neg C}$ (e.g. *shared-words*). $R_C \cap R_{\neg C} = \emptyset$.
c) Defining the notion of '*Strength of relation*' which represents the ability of a relation to retrieve relevant documents of a query. In other words, the strength assigned to a relation between two concepts C_1 and C_2 measures the extent to which if a document talks about C_1, it also talks about C_2 (Nie, 1992). We calculate the strength of a relation by using the following formula $\forall r \in R, \forall (c_i, c_j) \in r$:

$$Strength_r(c_i, c_j) = sim_r(c_i, c_j) \times certainty(r) \quad [1]$$

Where:

$$\forall r \in R, \quad certainty(r) = \begin{cases} 1 & r \in R_C \\ x \in]0,1[& r \in R_{\neg C} \end{cases} \quad [2]$$

$sim_r(c_i, c_j)$ represents the semantic similarity between two concepts.

Finally we define the conceptual indexing function $Index$: suppose there are a query q and a collection of documents D then:

$$Index : D \cup \{q\} \rightarrow C^* \quad [3]$$

where, C^* is the set of all subsets of C

2) Bayesian Network: To compute the matching between a document and a query, we use a Bayesian network (Murphy, 1998)(Le, 2009). The network in our model contains three types of nodes: documents D, concepts C, and query q. Nodes are connected by using three types of weighted links:
(1) $L_{DC} = \{(d, c) \,|d \in D, c \in Index\,(d)\}$: links from documents to their concepts, weighted by the importance of concept in its document.
(2) $L_{CQ} = \{(c, q) \,|c \in Index\,(q)\}$: links from concepts to their query, weighted by the importance of concept in the query.
(3) $L_{CC} = \{(c_i, c_j) \,|\exists d \in D, c_i \in Index\,(d)\,, c_j \in Index\,(q)\,, \exists r \in R, (c_i, c_j) \in r\}$: links from documents' concepts to query's concepts, represent relations between concepts, weighted by the *strength* of the relation.

3) Matching function: To calculate RSV (Relevance Status Value), we use the calculation rules of the conditional probability in Bayesian network, according to the following steps:
a) choosing a document $d_{selected}$ from document collection D, then:

$$\forall d \in D, \quad P\,(d) = \begin{cases} 1 & d = d_{selected} \\ 0 & \text{else} \end{cases} \quad [4]$$

b) for concepts that belong to the selected document $\{c_i|\,(d_{selected}, c_i) \in L_{DC}\}$:

$$P\,(c_i|L_{DC}) = \frac{weight_{DC}\,(d_{selected}, c_i)}{\sum_{(d_j, c_i) \in L_{DC}} weight_{DC}\,(d_j, c_i)} \quad [5]$$

c) for concepts that belong to the query and don't belong to the selected document and that are linked to a concept of the selected document $\{c_i|c_i \in Index\,(q)\,, c_i \notin Index\,(d_{selected})\,, \exists c_j \in Index\,(d_{selected})\,, (c_j, c_i) \in L_{CC}\}$:

$$P\,(c_i|L_{CC}) = \frac{\sum_{(c_j, c_i) \in L_{CC}} weight_{CC}\,(c_j, c_i) \times P\,(c_j|L_{DC})}{\sum_{(c_j, c_i) \in L_{CC}} weight_{CC}\,(c_j, c_i)} \quad [6]$$

d) now for the query node $RSV\,(d_{selected}, q) = P\,(q|L_{CQ})$:

$$P\,(q|L_{CQ}) = \frac{\sum_{(c_i, q) \in L_{CQ}} weight_{CQ}\,(c_i, q) \times P\,(c_i|L_{CC})}{\sum_{(c_i, q) \in L_{CQ}} weight_{CQ}\,(c_i, q)} \quad [7]$$

3. Model validation context

We validated the proposed model by applying it to the test collection: ImageCLEFMed2005, and by using the UMLS 2005 as an external resource. We used

MetaMap (Aronson, 2006) tool to identify concepts from raw text, we program a tool to build Bayesian network and calculate correspondence value, and we use the $tf.idf$ measure to calculate the importance of a concept in its document.

The goal of these experiments is showing that by enriching the external resource, more relevant documents could be retrieved. We have tested three variants of the model:
(1) Basic: there is no relations between concepts, i.e. it depends on the shared concepts between a document and a query to find matching.
(2) ISA: the *isa* relation (*isa*: this relation is predefined in UMLS) is used to link documents' concepts and query's concepts. Here,
$isa \in R_C$ then we have $certainty\,(isa) = 1$
$\forall\,(c_i, c_j) \in isa, \quad sim_{isa}\,(c_i, c_j) = \frac{1}{minLen(c_i, c_j)}$ where $minLen\,(c_i, c_j)$ is the path of minimum length between c_i and c_j according to isa.
b) ISA_SW: another relation (*shared-words*: this relation is added by us to UMLS) is added to ISA. Here,
$shared - words \in R_{\neg C}$ then we have $certainty\,(shared - words) = 0.1$ (10% is the value that gives the best result in our experiments)
$sim_{shared-words}\,(c_i, c_j) = mutual\_information\,(c_i, c_j) = \frac{NSW_{ij}}{NW_i \times NW_j}$
Where:
NSW_{ij}: number of shared words between c_i and c_j
NW_i: number of words in c_i
NW_j: number of words in c_j

We got the following results (see Tables 2, 3).

Table 2. *MAP of Basic, ISA, ISA_SW*

	MAP
Basic	0.1240
ISA	0.1395
ISA_SW	**0.1408**

Table 3. *Number of relevant, retrieved, and retrieved-relevant documents of Basic, ISA, ISA_SW*

	# Relevant documents	# Retrieved documents	# Retrieved-Relevent
Basic	2217	58037	1234
ISA	2217	101182	1464
ISA_SW	2217	128342	**1698**

From the previous results we can notice that, by exploiting relations between concepts (ISA), we could retrieve more relevant documents for a query (see Table 3) and

at the same time we gain a small enhancement in the precision of the system (see Table 2).

Also we can notice that, the enrichment of the external resource by adding a very simple relation (ISA_SW), allows us to retrieve more relevant documents (see Table 3) and also gain more enhancement in the precision (see Table 2).

4. Conclusion

We have presented in this paper our model to solve term and concept mismatch problems. In this model, documents and queries are represented by concepts, and we have also modeled the different relations between concepts.

This model also depends on the techniques of Bayesian Network to compute the matching value between a document and a query.

We show in this work that conceptual indexing is insufficient to solve the term mismatch problem. The use of relations from the conceptual resource increase the MAP, but we think that in UMLS, too many potential relation between concepts are missing. When we add these relations, we show an interesting increase in the MAP. In conclusion, these research tend to show that existing resources even very larges ones like UMLS, are not totally adapted to IR because of lack of relations between concepts. This lack can be partly compensated by analysis of terms associated to concepts. Finally there are many points in this work, that need more study, like studying the influence of adding other relations to the model, using properties other than Certainty to describe relations, and validation the model by using another test collections and another external resources.

A detailed version of this paper with different and more comprehensive experiments could be found in (Abdulahhad *et al.,* 2011).

5. References

Abdulahhad K., Chevallet J.-P., Berrut C., « Solving Concept mismatch through Bayesian Framework by Extending UMLS Meta-Thesaurus », *la huitième édition de la COnférence en Recherche d'Information et Applications (CORIA 2011), Avignon, France*, March 16–18, 2011.

Aronson A. R., « Metamap: Mapping text to the umls metathesaurus », 2006.

Baziz M., Indexation conceptuelle guidée par ontologie pour la recherche d'information, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre, 2005.

Chevallet J.-P., « endogènes et exogènes pour une indexation conceptuelle intermédia », Mémoire d'Habilitation a Diriger des Recherches, 2009.

Chevallet J.-P., Lim J. H., Le T. H. D., « Domain Knowledge Conceptual Inter-Media Indexing, Application to Multilingual Multimedia Medical Reports », *ACM Sixteenth Conference on*

Information and Knowledge Management (CIKM 2007), Lisboa, Portugal, November 6–9, 2007.

Crestani F., « Exploiting the similarity of non-matching terms at retrieval time », *Journal of Information Retrieval*, vol. 2, p. 25-45, 2000.

Le T. H. D., Utilisation de ressource externes dans un modèle Bayesien de Recherche d'Information: Application a la recherche d'information médicale multilingue avec UMLS, PhD thesis, Université Joseph Fourier, Ecole Doctorale MSTII, 2009.

Murphy K., « A Brief Introduction to Graphical Models and Bayesian Networks », 1998.

Nie J.-Y., « Towards a probabilistic modal logic for semantic-based information retrieval », *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, ACM, New York, NY, USA, p. 140-151, 1992.