Semantic Query Structuring to Enhance Precision of an Information Retrieval System: Application to the Medical Domain

Mohannad ALMASRI [1]

UJF-Grenoble 1, LIG laboratory, MRIM group
mohannad.almasri@imag.fr

ABSTRACT. *Most Information retrieval systems represent a query, also a document, as a bag of indexing terms without any relation between each other. This bag-based representation causes a problem for specialists when they deal with a specific domain like medical one. 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 has 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.*

RÉSUMÉ. *La plupart des systèmes de recherche d'information représentent la requête, et les documents, comme un sac de termes d'indexation sans aucune relation entre eux. Cette représentation pose problème pour les spécialistes d'un domaine spécifique comme le domaine médical. Nous proposons une alternative au sac de termes d'indexation, en fonction de la structuration requête sémantique, afin de répondre à ce besoin de précision dans un domaine spécifique. Cette structuration est obtenue en regroupant les termes d'indexation des requêtes à l'aide des catégories prédéfinies appelées dimensions. Ces dimensions représentent les différents aspects qui pourraient apparaître dans une requête ou un document. Les résultats expérimentaux montrent une amélioration de précision liée à la granularité des dimensions et de sa distribution.*

KEYWORDS: *Sematic Query, Structured Query, Conceptual Indexing, Domain Ontology*

MOTS-CLÉS : *Requête Sémantique, Requête Structurée, Indexation Conceptuelle, Ontologie*

[1] PhD Thesis directed by: Jean-Pierre Chevallet and Catherine Berrut

1. Introduction and Related Works

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 have a good knowledge about the related domain, and they are capable to build a precise or well-structured queries instead of simple keyword-based queries. The main problem of nowadays Web search engines and IRSs is the flat representation of queries and documents, or in other words, a bag of indexing terms [1] representation. This representation exhibits some lack of precision for specialists when they deal with a specific domain like medical. For example, the query number 4 in the ImageCLEF2011(Kalpathy-Cramer *et al.,* 2011) 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 normally called semantic categories or *dimensions*. The previous example shows that a simple keyword-based query is not sufficient to express the whole semantics within specialists' queries. This type of query partitioning or structuring requires an external resource, e.g. an ontology, a knowledge base, which is capable of separating indexing terms over semantic categories or dimensions. The notion of dimensions is proposed in order to navigate a base of images or textual documents (Eero Hyvönen *et al.,* 2003). Each dimension corresponds to a point of view according to which one can explore the base.

Semantic query structuring is used for different purposes in information retrieval. Li et al. (Li *et al.,* 2009) use semantic query structuring in order to search structured data. They tag each term in a query using pre-defined dimensions. Another example of semantic query structuring is to find multiple facets or aspects of a query (Dou *et al.,* 2011). These facets (called dimensions) are used for reformulating a query and improve the diversity of top results. Radhouani et al.(Radhouani *et al.,* 2010), propose a model of semantic query structuring based on conceptual indexing. Basically, they represent documents and queries by means of concepts. Then, they structure these concepts using dimensions from a domain knowledge.

In this paper, we present a semantic query structuring framework as an alternative to the bag of indexing terms. 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. The rest of this paper is organized as follows. In section 2 we talk about our framework for semantic query structuring. We report the experimental results in section 3 and conclude in section 4.

[1] Word, noun phrase, n-gram, or concept (Chevallet *et al.,* 2007).

2. 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, this concept-based representation is obtained using conceptual indexing (Chevallet *et al.,* 2007). Therefore, we need an additional component, containing concepts, which is the external resource. This external resource not only helps IRS in the conceptual indexing, 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 called *dimensions*. This semantic categorization feature for concepts should be supplied by our external resource. For example, assume that a document contains the two terms "Adrenal Cortical Hypofunction" and "Hodgkin Disease", in UMLS [2], 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 *dimension-level*. We have two proposals, in order to take advantage of this structuring idea: 1) *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*. Therefore, the similarity between a document and a query takes into account concept-level and dimension-level. 2) *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 where each sub-document corresponds to a specific dimension and contains the document concepts that belong to this dimension. The same for queries.

Our query structuring framework is the tuple (D, E, F, RSV), where D is the document collection; E is an external resource; F is a conceptual indexing function; RSV is a matching function. We now detail the components of our framework.

2.1. *External Resource* E

An external resource is modeled by $E = (C, M, H)$, where C is a set of concepts, M is a set of dimensions, H is a mapping function that maps each concept $c_i \in C$ into its set of dimensions $H(c_i)$.

$$C = \{c_1, \cdots, c_n\}; \quad M = \{m_1, \cdots, m_k\}; \quad H : C \to 2^M$$

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

[2] Unified Medical Language System. It is a meta-thesaurus in medical domain.

2.2. *Query and Document Model*

The conceptual indexing can be represented as a function: $F\colon D \cup \{q\} \rightarrow 2^C$ where 2^C is the power set of C. At this point, each document $d \in D$ is represented by a set of concepts $d_c = F(d)$, and this is the first level of a document representation in our framework (concept-level).

The second level (dimension-level) aims to represent documents and queries depending on dimensions. By applying the mapping function H to each concept $c \in d_c$ in the document, we obtain the second level d_m of a document d as follows:

$$d_m = \bigcup_{c_i \in d_c} H(c_i)$$

In our framework, we also look at documents and queries from another point of view. A document d is a set of composed dimensions d_m^c and each composed dimension m_k^d is the concepts of d_c that belong to the dimension m_k. Hence, we define:

$$m_k^d = \{c | c \in d_c, m_k \in H(c)\}; \quad d_m^c \ = \{m_k^d | m_k \in d_m\}$$

We apply the previous steps 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.

2.3. *Matching Model*

Documents and queries are represented using two semantic levels: a fine-grain level which is concept-level and a coarse-grain level which is dimension-level. We have, according on our proposals, two ways to compute $RSV(d,q)$:
1) Semantic Levels Matching (SLM): In order to compute $RSV(d,q)$, we combine 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, using equation 1.

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

where $\alpha \in [0,1]$ is a tuning parameter and represents the importance of each level, and similarities Sim_c and Sim_m can be computed using any IR model (e.g. language models or BM25). Whereas, 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.
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.

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

3. Experiments

In this section, we validate our two proposals SLM and SDM against the case-based test collection of CLEF 2011 (Kalpathy-Cramer *et al.,* 2011). We use MetaMap (Aronson, 2006) and UMLS 2011 for conceptual indexing operation. UMLS is an external resource contains concepts which are categorized using two different possibilities of dimensions called: *semantic groups* and *semantic types*. We use three models for computing the similarity between a document and a query: Dirichlet (DIR), Jelinek-Mercer (JM), and BM25 (Zhai *et al.,* 2004, Robertson *et al.,* 1994). We validate our two proposals for semantic query structuring using three experiments: the baseline results (BL) are obtained using queries and documents as a bag of concepts without applying semantic query structuring.

		MAP			P@5		
	Model	BL	SQS-ST	Gain	BL	SQS-ST	Gain
SLM	JM	0.1247	**0.1299***	+4.17%	0.20	0.26	+30%
	Dir	0.1036	0.1070	+3.28%	0.20	0.24	+20%
	BM25	0.0956	0.1116	+16.73%	0.18	0.22	+22.22%
SDM	JM	0.1247	0.1166	-6.57%	0.20	0.19	-5%
	Dir	0.1036	0.0791	-23.64%	0.20	0.15	-25%
	BM25	0.0956	0.1043	+9.1%	0.20	0.20	+10%

Table 1. *Improvement using semantic types as dimension with our SLM and SDM proposals.* best MAP in CLEF2011 campaign is 0.1297.*

3.1. *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 with their frequencies. Whereas, dimension can be one of two possible categorization from UMLS:

- Using UMLS semantic groups as dimensions (SLM-SG): RSV is computed using equation 1, where m is a UMLS semantic group and Sim_c and Sim_m are one of the following models: JM, Dir, and BM25. Our results show that there is no improvement obtained by using UMLS semantic groups as dimension, because the distribution of semantic groups over the test collection is uniform. In other words, all documents nearly contain concepts from all groups.
- Using UMLS semantic types as dimensions (SLM-ST): RSV is computed using equation 1, where m is a UMLS semantic type. The results obtained by different models are summarized in Table 1. Using UMLS semantic types as dimensions gives the potential for precision improvement, because the distribution of semantic types is less uniform than the distribution of semantic groups. In addition, α determines the importance of dimension-level and concept-level in the matching process, and the value of α (0.9 in our case) seems to be model independent and corpus dependent.

3.2. *Validation Using Semantic Dimension Matching (SDM)*

In this third experiment, we structure a query and a document using our second proposal SDM with UMLS semantic types as dimensions. For computing $RSV(d, q)$ between a document d and a query q, we use the equation 2, results in Table 1. As we split documents into dimensions and using language model on these dimensions, the results for Jelinek-Mercer and Dirichlet are less than baseline. We think that language models give a better probability estimation for long documents than short documents. In the other hand, the results of BM25 are better than baseline.

4. 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 semantic 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 (Kalpathy-Cramer *et al.,* 2011). Our results show 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 on other test collections, other domains, and study the relation between the value of our tuning parameter α with the properties of studied collections.

5. References

Aronson A. R., « MetaMap: Mapping Text to the UMLS Meta-thesaurus », 2006.

Chevallet J.-P., Lim J.-H., « Domain knowledge conceptual inter-media indexing: application to multilingual multimedia medical reports », CIKM '07, p. 495-504, 2007.

Dou Z., Hu S., « Finding dimensions for queries », CIKM '11, p. 1311-1320, 2011.

Eero Hyvönen A. S., Saarela S., « Ontology-Based Image Retrieval », 2003.

Kalpathy-Cramer J., Müller H., Bedrick S., « Overview of the CLEF 2011 Medical Image Classification and Retrieval Tasks », *CLEF*, 2011.

Li X., Wang Y.-Y., Acero A., « Extracting structured information from user queries with semi-supervised conditional random fields », SIGIR '09, ACM, p. 572-579, 2009.

Radhouani S., Kalpathy-Cramer, « Using media fusion and domain dimensions to improve precision in medical image retrieval », CLEF'09, p. 223-230, 2010.

Robertson S. E., Walker S., « Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval », SIGIR '94, p. 232-241, 1994.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to information retrieval », *ACM*, vol. 22, n° 2, p. 179-214, April, 2004.