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.

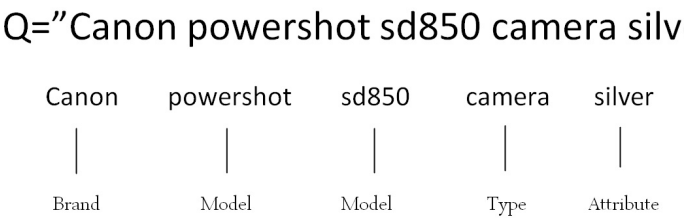


FIGURE 1 – Semantic query structuring example. Each query term (canon, powershot, sd850, camera, silver) is tagged by a dimension (Brand, Model, Type, Attribute).

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.

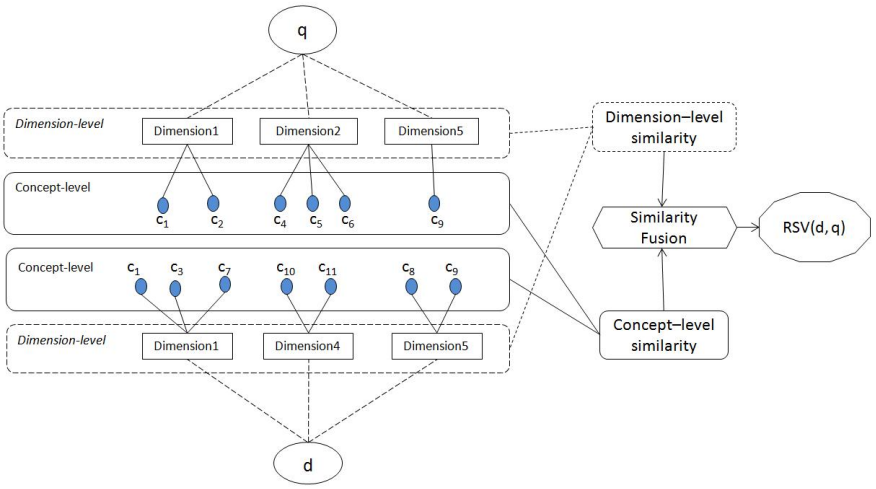


FIGURE 2 – Semantic Levels Matching : RSV is the fusion between the concept-level similarity of a document d and a query q, and the dimension-level similarity between d and q.

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 $\psi(c)$.

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

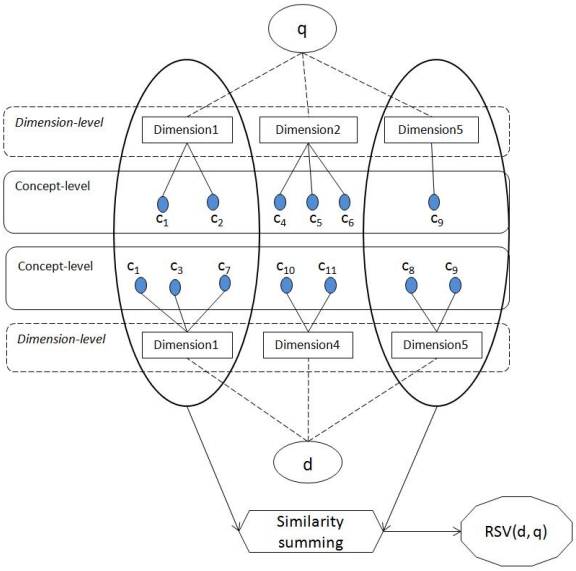


FIGURE 3 – Semantic Dimension Matching : Document d and a query q are splitted into dimensions $(1, 2, 5)$ and $(1, 4, 5)$, respectively, and each dimension contains the concepts of d or q that belong to this dimension. $RSV(d, q)$ is a sum of the similarity for all shared dimensions $(1, 5)$.

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.

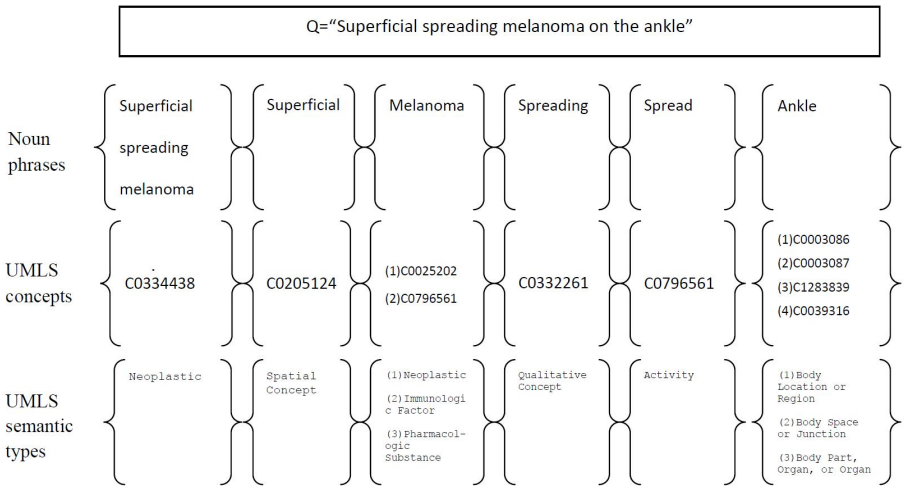


FIGURE 5 – MetaMap mapping example for the query "Superficial spreading melanoma on the ankle".

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.

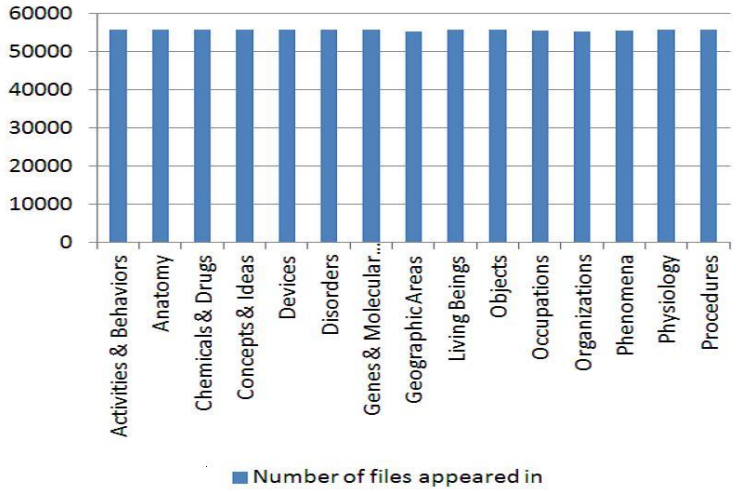


FIGURE 6 – Semantic groups distribution. This histogram shows that each document appears in all UMLS semantic groups or each document contains all UMLS semantic groups. In other words, these semantic groups are not able to discriminate the corpus.

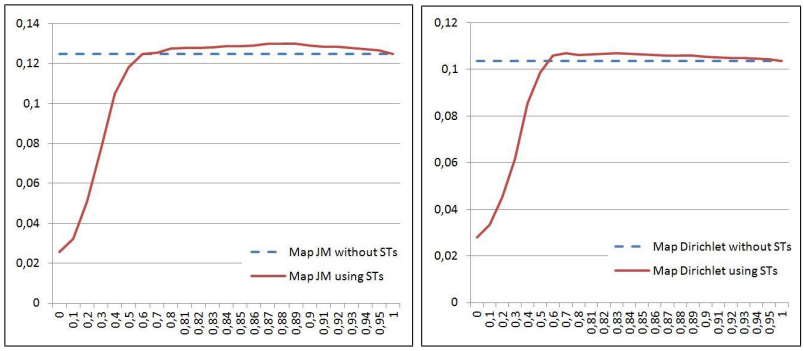
of the following models : JM, Dir, and BM25. The results obtained by different models are summarized in Table 3.

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)

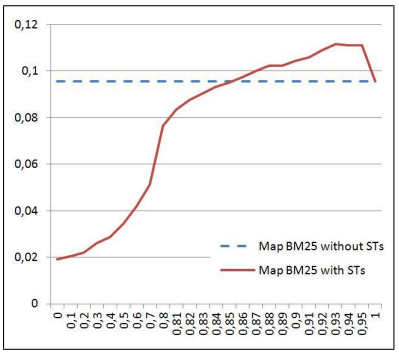
Model	Baseline MAP	SLM-ST MAP	Gain	Baseline $P@10$	SLM-ST $P@10$	Gain
JM	0.1247	**0.1299***	+4.17%	0.1600	0.1800	+12.5%
Dir	0.1036	0.1070	+3.28%	0.1500	0.1800	+20%
BM25	0.0956	0.1116	+16.73%	0.1400	0.1700	+22.22%

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.