Matching Fusion with Conceptual Indexing

Karam Abdulahhad, Jean-Pierre Chevallet, Catherine Berrut

Matching Fusion with Conceptual Indexing

Karam Abdulahhad*, Jean-Pierre Chevallet**, Catherine Berrut***

* UJF-Grenoble 1, LIG laboratory, MRIM group
karam.abdulahhad@imag.fr
** UPMF-Grenoble 2, LIG laboratory, MRIM group
jean-pierre.chevallet@imag.fr
*** UJF-Grenoble 1, LIG laboratory, MRIM group
catherine.berrut@imag.fr

Abstract. Many studies have been addressed the term-mismatch problem, which arises when using different terms or words for expressing the same meaning. We also introduce another problem: over-specialized document, which is caused when IR systems prefer documents that have poor query-document intersection, but with high weighting value, to those that have rich query-document intersection with low weighting value. In this study, we propose to use, simultaneously, multiple types of indexing elements: ngrams, keywords, and concepts, instead of only keywords. We followed a late data-fusion technique to achieve that. Through our proposed model, we also try to overcome the over-specialized document problem. Experiments for model validation have been done by using ImageCLEF2011 test collection, UMLS2009 Meta-thesaurus, and MetaMap tool for mapping text into UMLS concepts.

1 Introduction

Two terms or words could have the same meaning in a specific context. For example, (atrial, auricular), (apartment, flat), (air pollution, pollution of the air), etc. This is one of the preferable features of the natural languages, and one of features that gives each author the ability to have her/his own writing style. However, in IR field, it is a problematic feature, because the most of IR systems use a type of query-document intersection. Therefore, by using different terms, in queries and documents, for expressing the same meaning, IR systems will not be able to retrieve relevant documents. This problem is well studied in literatures and is called "**term-mismatch**" problem Woods (1997) Crestani (2000) Baziz (2005) Maisonnasse (2008) Chevallet (2009).

Most of IR systems use a type of weighting for estimating the amount of contribution of an indexing element in the overall matching, and subsequently for ranking the retrieved documents. There are many weighting formulas Harter (1975a) Harter (1975b) Robertson and Walker (1994) Lee (1995) Amati and Van Rijsbergen (2002), each with its properties. However, using element weighting in IR systems poses a problem. In general, IR systems can not warrant that documents, which share more number of distinct elements with queries, will be ranked higher than other documents. In other words, documents that have less shared elements with

queries, but with high weighting values, could be ranked higher than documents that have more shared elements, but with low weighting values. For us, this is inconvenient behavior and it is better to retrieve documents that cover more aspects of a query, even with a low weighting values. We will call this problem an "**over-specialized documents**".

In this study, we address these two problems and we try to find a practical and effective solution for them.

The paper will be organized as follow: in section 2, we present some related works. In section 3, we describe in details our proposed model. Section 4 presents some experiments for model validation, then we discuss our results. We conclude in section 5.

2 Related Works

Before we go forward, we should define terms and concepts. A term is a sequence of words that have a unique meaning in a specific domain Chevallet (2009), whereas concepts could 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). Practically, each concept is represented by an identifier in an external resource and is associated with a set of synonym terms Baziz (2005) Chevallet (2009).

The term-mismatch problem was heavily studied by multiple researchers. In literatures, several approaches, to solve this problem, could be identified:

1. Dimensionality reduction: reduces the chance that a query and a document use different terms for representing the same meaning. Among the techniques that are used for achieving this mission, we can mention: Stemming Frakes (1992), Latent Semantic Indexing (LSI) Deerwester (1988) Deerwester et al. (1990), and Conceptual Indexing (using concepts instead of terms) Chevallet et al. (2007).
2. Query expansion: extends the query with new terms to increase the chance of matching with documents Efthimiadis (1996).
3. Using term-term semantic similarity measures: this approach presupposes the existence of a measure capable of estimating the similarity between any two terms Crestani (2000).

$$\forall t_i, t_j \in T, \quad 0 \leq Sim\left(t_i, t_j\right) \leq 1 \tag{1}$$

In our previous studies Abdulahhad et al. (2011b) Abdulahhad et al. (2011a), we used concepts as a solution for the term-mismatch problem. For example, the two terms '*Atrial Fibrillation*' and '*Auricular Fibrillation*' correspond to the same concept '*C0004238*' in UMLS [1]. However, using concepts poses another problem: the **concept-mismatch** problem Abdulahhad et al. (2011b). An example about this problem could be: according to UMLS, the two terms '*Dermatofibroma*' and '*Dermatofibrosarcoma*' correspond to two different concepts '*C0002991*' and '*C0392784*', respectively. Therefore, even by using concepts, the mismatch between a document containing '*Dermatofibroma*' and a query containing '*Dermatofibrosarcoma*' still exist.

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

In addition, external resources, e.g. UMLS, that contain concepts are generally incomplete Bodenreider et al. (1998) Bodenreider et al. (2001) Abdulahhad et al. (2011b). For example, the term '*Osteoporotic*' does not map to any concept in UMLS2009.

Many approaches to solve the concept-mismatch problem could be found in literatures:

1. Exploiting semantic relations between concepts: especially the hyponymy-hypernymy relations Baziz (2005) Maisonnasse (2008) Le (2009) Abdulahhad et al. (2011b) Abdulahhad et al. (2011a).
2. Query expansion: by exploiting concepts, their content, and their position in the external resource Aronson and Rindflesch (1997) Baziz (2005).
3. Domain dimensions: indexing documents and queries by domain dimensions, which are more abstract than concepts Radhouani (2008).

Again in our previous studies Abdulahhad et al. (2011b) Abdulahhad et al. (2011a), we exploit semantic relations and enrich the external resource by new relations for solving the concept-mismatch and external resource incompleteness problems, respectively.

In this study, we use concepts, but we tried **data fusion** as another technique to solve, simultaneously, the two problems: concept-mismatch and external resources incompleteness.

Data fusion is the process of combining different result sets of a certain information need. Different result sets of the same information need, in the same corpus, could be produced by Croft (2000):

1. Using different IR systems.
2. Using the same IR system in conjunction with different configurations:
   (a) Different document representations using: different types of indexing elements, different parts of documents, different weighting schemas, etc.
   (b) Different queries of the same information need.
   (c) Different matching formulas.

Actually, in this study, we used three types of indexing elements: ngrams, words, and concepts. Through the following two examples, we will illustrate the reason of these choices.

Example 1: The query number 21 in the Image-based task of ImageCLEF2009 [2] is: "*osteoporotic bone*". According to MetaMap [3], there is no concept in UMLS corresponds to the word '*osteoporotic*'. Using only concepts without ngrams and words, and as the word '*bone*' is an indiscriminative, the IR system will not be able to retrieve the relevant documents of this query.

Example 2: The query number 16 in the Image-based task of ImageCLEF2010 [4] is: "*images of dermatofibroma*". The word '*dermatofibroma*' does not exist in the corpus, istead the corpus contains the word '*dermatofibrosarcoma*'. Therefore, and as the word '*images*' is an indiscriminative, using only words as indexing elements will not be sufficient. In addition, according to MetaMap, the word '*dermatofibroma*' maps into the concept '*C0002991*' and the word '*dermatofibrosarcoma*' maps into one of the concepts: '*C2697408*', '*C0206647*', or '*C0392784*'. By consulting the UMLS, we did not find any direct relation between the '*dermatofibroma*' concept and one of the '*dermatofibrosarcoma*' concepts. Consequently, and as

2. http://www.imageclef.org/2009
3. is a tool to map text into UMLS concepts. http://metamap.nlm.nih.gov/
4. http://www.imageclef.org/2010

neither the word '*images*' nor the concepts that correspond to it are discriminative, the two types of indexing elements (words and concepts) are not sufficient to retrieve relevant documents. That's what justify the usage of ngrams.

Concerning matching formulas, multiple heuristics could be found in IR literatures. The Fang et al. (2004) recalls some of those heuristics and transforms them to a set of constraints. Any matching formula should satisfy some of these constraints to be effective. Table (1) lists the constraints [5].

TAB. 1 – *The constraints*

Constraints	**Intuitions**
TFC1	to favor a document with more occurance of a query term
TFC2	to favor document matching more distinct query terms
TFC2	to make sure that the change in the score caused by increasing TF (Term Frequency) from 1 to 2 is larger than that caused by increasing TF from 100 to 101
TDC	to regulate the impact of TF and IDF (Inverse Document Frequency): it ensure that, given a fixed number of occurences of query terms, we should favor a document that has more occurences of discriminative terms (i.e. high IDF terms)
LNC1	to penalize a long document
LNC2, TF-LNC	to avoid over-penalize a long document: as it says that if we concatenate a document with itself k times to form a new document, then the score of the new document should not be lower than the original document
TF-LNC	to regulate the interaction of TF and document length: if d_1 is generated by adding more occurences of the query term to d_2, the score of d_1 should be higher than d_2

In our matching formula, we tried to meet some of these constraints, in order to build an effective formula.

3 The Proposed Model

3.1 Three Types of Indexing Elements

After the discussion in the previous section, We believe now that no single type of indexing elements could completely represent the content of documents and queries, because:

1. there is no perfect indexing function Baziz (2005) Aronson (2006) Dozier et al. (2007). It is always an approximate function.
2. concerning concepts, most resources, e.g. UMLS, are incomplete Bodenreider et al. (1998) Bodenreider et al. (2001) Abdulahhad et al. (2011b).

5. The presentation of the constraints is a rendering of the original presentation in Fang et al. (2004)

3. each type of indexing elements covers an aspect of documents and queries Das-Gupta and Katzer (1983). Ngrams cover the morphological aspect, words cover the lexical aspect, and concepts cover the conceptual aspect.

The goal of the indexing function is to convert documents and queries from their original form to another easy to use form. As we have three different types of elements: ngrams (NG), words (W), and concepts (C), three indexing functions are defined (one for each type).

$$Index_{NG} : D \cup Q \rightarrow E^*_{NG} \tag{2}$$

$$Index_{W} : D \cup Q \rightarrow E^*_{W} \tag{3}$$

$$Index_{C} : D \cup Q \rightarrow E^*_{C} \tag{4}$$

Where
D set of documents
Q set of queries
E_{NG} set of ngrams
E_W set of words
E_C set of concepts
E^* the set of all subsets of E

3.2 Matching Function

Our model, as almost all models, depends on some hypotheses. Actually, it depends on the following hypotheses:

1. The more shared elements a document and a query have, the more semantically closer they are. This hypothesis corresponds to the $TFC2$ constraint in Fang et al. (2004) (see Table 1).
2. The descriptive power of an element (local weight): the more frequently an element occurs in a document, the better it describes the document Luhn (1958) Baziz (2005). This hypothesis corresponds to the $TFC1$ constraint in Fang et al. (2004) (see Table 1).
3. The discriminative power of an element (global weight): the less number of documents an element appears in, the more important it is Luhn (1958) Baziz (2005). This hypothesis corresponds to the TDC constraint in Fang et al. (2004) (see Table 1).
4. As we use document length for element frequency normalization, our model is also compatible with the $LNC1$ constraint in Fang et al. (2004) (see Table 1).

One of the future potential works could be to make our model compatible with the other constraints.

By taking these hypotheses into account, our model could be formulated according to each type of elements. The Relevance Status Value (RSV) between a document d and a query q is:

Words

$$RSV_W\left(d,q\right) = \left\|d \cap q\right\|_W \times \left(\sum_{w \in q} \frac{N}{N_w} \times \frac{f_{d,w}}{\left\|d\right\|_W} \times \left\|w\right\|\right) \tag{5}$$

Where
$d = \{w | w \in Index_W\left(d\right)\}$ a document

$q = \{w|w \in Index_W\,(q)\}$ a query
$\|d \cap q\|_W = \|\{w|w \in Index_W\,(d) \cap Index_W\,(q)\}\|$ the number of shared words between a document d and a query q
N the number of documents in the corpus
$N_w = \|\{d|w \in Index_W\,(d)\}\|$ the number of documents that contain the word w
$f_{d,w}$ the number of occurances of a word w in a document d
$\|d\|_W$ the number of words in a document d
$\|w\|$ the number of characters in a word w

We added the last component $\|w\|$ to express the importance of an element itself, in isolation from the document that contains it and from the corpus. In other words, is it possible to say that a document d_1 containing an element e_1 should have higher retrieval score than another document d_2 containing e_2, in isolation from all statistical aspects of the model? Well, we believe that the existence of such measure will improve the effectiveness of IR models.

By using words as indexing elements, we tried to approximate this measure by simply supposing that the longer a word is, the more important it is. For ngrams and concepts, we supposed that all concepts/ngrams are equally important and we omitted this component from the model. Finding an effective measure for each type of elements, maybe, is a good direction for the future works of this study.

Ngrams

$$RSV_{NG}\,(d,q) = \|d \cap q\|_{NG} \times \left(\sum_{ng \in q} \frac{N}{N_{ng}} \times \frac{f_{d,ng}}{\|d\|_{NG}}\right) \tag{6}$$

Concepts

$$RSV_{C}\,(d,q) = \|d \cap q\|_{C} \times \left(\sum_{c \in q} \frac{N}{N_{c}} \times \frac{f_{d,c}}{\|d\|_{C}}\right) \tag{7}$$

3.3 The Three Types in one Model (Matching Fusion)

As we said earlier, no single type of indexing elements could cover all aspects of documents and queries. Therefore, merging all types (aspects) in one model could improve the performance of our model Croft (1981) Belkin et al. (1993) Shaw and Fox (1994). One of the merging formulas is:

$$RSV_{all}\,(d,q) = RSV_{NG}\,(d,q) + RSV_{W}\,(d,q) + RSV_{C}\,(d,q) \tag{8}$$

We used the SUM formula for combining the result sets of the three types of elements. This is a type of **late-fusion** because we used each type of elements alone, then we merged the result sets. Conversely to late-fusion, **early-fusion** means combining the three types of elements in one index structure, then applying the model as we have one element type.

4 Model Validation

4.1 Validation Context

The proposed model is validated by applying it to the corpus of ad-hoc image-based retrieval task of the Medical Retrieval track of ImageCLEF2011, and by using the UMLS2009

as an external resource. We use MetaMap Aronson (2006) tool to identify concepts from raw text.

ImageCLEF is a part of CLEF[6] (Cross-Language Evaluation Forum), which is a yearly campaign for evaluation of multilingual information retrieval since 2000. ImageCLEFMed concerns searching medical images depending on heterogeneous and multilingual documents that contain text and images.

ImageCLEF2011[7] Kalpathy-Cramer et al. (2011) contains four main tracks: 1) medical retrieval, 2) photo annotation, 3) plant identification, and 4) Wikipedia retrieval. Medical retrieval track contains three tasks: 1) modality classification, 2) ad-hoc image-based retrieval which is an image retrieval task using textual, image or mixed queries, and 3) case-based retrieval: in this task the documents are journal articles extracted from PubMed[8] and the queries are case descriptions.

The corpus that we used contains: about $230,000$ images with their text caption and title written in English and 30 queries written in three languages: English, French, and German.

UMLS is a multi-source knowledge base in the medical domain. It contains three sources of knowledge:

1. Metathesaurus: is a vocabulary database in the medical domain, extracted from many sources, each source of them is called "Source Vocabularies". The Metathesaurus is organized in Concepts, which represent the common meaning of a set of strings extracted from different source vocabularies.
2. Semantic Network: consists of a set of Semantic Types linked together by two different types of Semantic Relations (hierarchical, non-hierarchical). The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus.
3. SPECIALIST Lexicon: is a set of general English or biomedical terms and words extracted from different sources.

Moreover, UMLS contains many tools to deal with these different sources (e.g. MetamorphSys, UMLS Knowledge Source Server).

MetaMap is a tool to map text into UMLS concepts. This tool is composed of the following components:

1. Morphology and Syntax: extraction of noun phrases from text using NLP techniques.
2. Variation: construction of different forms (variants) of the noun phrase or part of it.
3. Identification: for each noun phrase variant, it retrieves all concepts that possibly correspond to this variant. The set of concepts that possibly corresponds to the noun phrase, is called "Candidates set".
4. Evaluation: ordering the concepts of candidate set according to an evaluation function (f), which determines: "how much the concept represents the noun phrase?".

6. http://www.clef-campaign.org/
7. http://www.imageclef.org/2011
8. http://www.ncbi.nlm.nih.gov/pubmed/

5. Disambiguation: reduction of the size of the candidates set.

In this study, image captions and titles are used as documents and only the English part of queries is taken into account.

4.2 Text Indexing

We extracted three types of indexing elements:

1. 5gram [9]: before extracting 5grams from documents and queries, we deleted all non-ASCII characters. Then we used five-characters-wide window for extracting 5grams with shifting the window one character each time.
2. Words: before extracting words from documents and queries, we deleted all non-ASCII characters. Then we eliminated the stop words and stem the remaining words using Porter algorithm to get finally the list of words that index documents and queries.
3. Concepts: before mapping the text of documents and queries to concepts, we deleted all non-ASCII characters. Then we mapped the text to UMLS's concepts using MetaMap.

4.3 Model Variants

Actually we experimented seven variants of our model in this study, which are:

$$RSV_{5G*}(d,q) = \left(\sum_{5g \in q} \frac{N}{N_{5g}} \times \frac{f_{d,5g}}{\|d\|_{5G}}\right) \tag{9}$$

$$RSV_{5G}(d,q) = \|d \cap q\|_{5G} \times \left(\sum_{5g \in q} \frac{N}{N_{5g}} \times \frac{f_{d,5g}}{\|d\|_{5G}}\right) \tag{10}$$

$$RSV_{W*}(d,q) = \left(\sum_{w \in q} \frac{N}{N_w} \times \frac{f_{d,w}}{\|d\|_{W}} \times \|w\|\right) \tag{11}$$

$$RSV_{W}(d,q) = \|d \cap q\|_{W} \times \left(\sum_{w \in q} \frac{N}{N_w} \times \frac{f_{d,w}}{\|d\|_{W}} \times \|w\|\right) \tag{12}$$

$$RSV_{C*}(d,q) = \left(\sum_{c \in q} \frac{N}{N_c} \times \frac{f_{d,c}}{\|d\|_{C}}\right) \tag{13}$$

$$RSV_{C}(d,q) = \|d \cap q\|_{C} \times \left(\sum_{c \in q} \frac{N}{N_c} \times \frac{f_{d,c}}{\|d\|_{C}}\right) \tag{14}$$

$$RSV_{SUM}(d,q) = RSV_{5G}(d,q) + RSV_{W}(d,q) + RSV_{C}(d,q) \tag{15}$$

9. 5gram is a ngram consists of five characters. We picked out 5grams because they gave the best results using ImageCLEF2010 comparing to the other ngrams.

4.4 Results and Discussion

The following table (see Table 2) contains the obtained results. The first row (Best) is the result of the first ranked run in the ad-hoc image-based retrieval task in the official campaign [10]. We presented the results using four different metrics: 1) MAP: Mean Average Precision, 2) P@10: Precision after 10 documents retrieved, 3) P@20: Precision after 20 documents retrieved, and 4) #rel_ret: total number of relevant documents retrieved over all queries.

TAB. 2 – *The results of ad-hoc image-based retrieval task*

	MAP	P@10	P@20	# rel_ret	Rank
Best	**0.2172**	**0.3467**	0.3017	1471	1
5G* (Formula 9)	0.1123	0.2033	0.1567	1260	
5G (Formula 10)	0.1473	0.2367	0.2017	1290	
W* (Formula 11)	0.1313	0.1900	0.1967	1421	
W (Formula 12)	0.1963	0.3100	0.2850	1501	
C* (Formula 13)	0.1461	0.2333	0.2133	1456	
C (Formula 14)	0.1664	0.2933	0.2633	1463	
5G+T+C (Formula 15)	0.2008	0.3033	**0.3050**	**1544**	8

The first direct deduction from the results (Table 2) is the importance of $\|d \cap q\|$ component. For ngrams, words, and concepts the precision of the system is improved by using $\|d \cap q\|$ component.

– Promoting documents that share more distinct elements with a query, improves system effectiveness.

The other notable thing in the results is the effectiveness of the data fusion technique, especially, in the number of relevant-retrieved documents. Knowing that the formula (SUM formula), which is used for fusion, was very simple.

– Data fusion is effective in retrieving more relevant documents.

The performance of our model was not bad, even with precision degradation by -7.5%, comparing to the best result. Actually, our model is very simple, and even when we used concepts, we did not exploit any relation. Whereas, the best result Vázquez et al. (2011) was obtained by using a type of query expansion, and exploiting the content and relations of the MeSH [11] ontology.

5 Conclusion

We presented in this paper our approach to index and retrieve documents. We used three types of indexing elements (ngrams, words, concepts) for building a multi-facet document representation, and then we used a simple formula based on three hypotheses (the amount of overlap between a document and a query, the descriptive power of an indexing element, and the discriminative power of an indexing element) for retrieving documents, considering all facets

10. To see all results: http://www.imageclef.org/2011/medical
11. http://www.ncbi.nlm.nih.gov/mesh

(elements' types) of documents. We tried, by using three types of elements then merging them together, to solve the concept-mismatch and external resources incompleteness problems.

We obtained good results. The eighth out of 64 runs in the ad-hoc image-based retrieval task. Knowing that, we used a very simple structure for representing documents and queries and also a very simple ranking formula.

Finally, this study still needs some work. We will compare the performance of our model to the performance of some well-known models e.g. DFR Amati and Van Rijsbergen (2002), BM25 Robertson and Walker (1994), etc.

In addition, we verified the effectiveness of our model by using only one corpus ImageCLEF2011. In order to obtain a more reliable and stable deductions, we should check our model using other corpuses.

Concerning data fusion technique, we tried simple formula (SUM formula). We could try other formulas Shaw and Fox (1994).

In section 3.2, we have introduced the notion of the importance of an element itself. Finding effective and expressive measure, according to each type (ngrams, words, and concepts), is not an easy mission and needs a lot of work.

References

Abdulahhad, K., J.-P. Chevallet, and C. Berrut (2011a). Exploiting and Extending a Semantic Resource for Conceptual Indexing. In *Troisième Atelier Recherche d'Information SEmantique (RISE 2011), Avignon, France*.

Abdulahhad, K., J.-P. Chevallet, and C. Berrut (2011b). Solving concept mismatch through bayesian framework by extending umls meta-thesaurus. In *CORIA 2011*, pp. 311–326.

Amati, G. and C. J. Van Rijsbergen (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst. 20*, 357–389.

Aronson, A. R. (2006). Metamap: Mapping text to the umls metathesaurus.

Aronson, A. R. and T. C. Rindflesch (1997). Query expansion using the umls metathesaurus. *Proceedings of the AMIA Annual Fall Symposium*, 485–489.
français

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

Belkin, N. J., C. Cool, W. B. Croft, and J. P. Callan (1993). The effect of multiple query representations on information retrieval system performance. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, New York, NY, USA, pp. 339–346. ACM.

Bodenreider, O., A. Burgun, G. Botti, M. Fieschi, P. L. Beux, and F. Kohler (1998). Evaluation of the unified medical language system as a medical knowledge source. *Journal of the American Medical Informatics Association 5*(1), 76–87.

Bodenreider, O., A. Burgun, and T. C. Rindflesch (2001). Lexically-suggested hyponymic relations among medical terms and their representation. In *in the UMLS, in Proceedings of TIA 2001*.

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

Chevallet, J.-P., J. H. Lim, and T. H. D. Le (2007). Domain knowledge conceptual inter-media indexing, application to multilingual multimedia medical reports. In *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal*.

Crestani, F. (2000). Exploiting the similarity of non-matching terms at retrievaltime. *Inf. Retr. 2*, 27–47.

Croft, W. B. (1981). Incorporating different search models into one document retrieval system. *SIGIR Forum 16*, 40–45.

Croft, W. B. (2000). *Combining Approaches to Information Retrieval*, Volume 7. Springer.

Das-Gupta, P. and J. Katzer (1983). A study of the overlap among document representations. *SIGIR Forum 17*, 106–114.

Deerwester, S. (1988). Improving information retrieval with latent semantic indexing. In C. L. Borgman and E. Y. H. Pai (Eds.), *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, Volume 25, Atlanta, Georgia. American Society for Information Science.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 41*(6), 391–407.

Dozier, C., R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. S. Guo (2007). Fast tagging of medical terms in legal text. In *ICAIL*, pp. 253–260.

Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Systems and Technology (ARIST) 31*, 121–187.

Fang, H., T. Tao, and C. Zhai (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, New York, NY, USA, pp. 49–56. ACM.

Frakes, W. B. (1992). *Stemming algorithms*, pp. 131–160. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Harter, S. P. (1975a). A probabilistic approach to automatic keyword indexing. part i: On the distribution of specialty words in a technical literature. *Journal of the American Society for Informaiton Science 26*(4), 197–206.

Harter, S. P. (1975b). A probabilistic approach to automatic keyword indexing. part ii: An algorithm for probabilistc indexing. *Journal of the American Society for Informaiton Science 26*(4), 280–289.

Kalpathy-Cramer, J., H. Müller, S. Bedrick, I. Eggel, A. G. S. de Herrera, and T. Tsikrika (2011). The clef 2011 medical image retrieval and classification tasks.

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

Lee, J. H. (1995). Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, New York, NY, USA, pp. 180–

188. ACM.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev. 2*, 159–165. Français

Maisonnasse, L. (2008). *Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale.* Ph. D. thesis, Université Joseph-Fourier - Grenoble I.

Radhouani, S. (2008). *Un modèle de recherche d'information orienté précision fondé sur les dimensions de domaine*. Ph. D. thesis, Co-tutelle Université Joseph Fourier Grenoble, Université de Genève (Suisse).

Robertson, S. E. and S. Walker (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, New York, NY, USA, pp. 232–241. Springer-Verlag New York, Inc.

Shaw, J. A. and E. A. Fox (1994). Combination of multiple searches. In *Text REtrieval Conference*.

Vázquez, J. M., M. Crespo, and M. J. M. López (2011). Laberinto at imageclef 2011 medical image retrieval task. In V. Petras, P. Forner, and P. D. Clough (Eds.), *CLEF (Notebook Papers/Labs/Workshop)*.

Woods, W. A. (1997). Conceptual indexing: A better way to organize knowledge. Technical report, Mountain View, CA, USA.

Résumé