

Knowledge Modeling In Prior Art Search

Erik Graf, Ingo Frommholz, Mounia Lalmas, and Keith van Rijsbergen

University of Glasgow,
{graf, frommholz, mounia, keith,
}@dcs.gla.ac.uk
<http://ir.dcs.gla.ac.uk/>

Abstract. This study explores the benefits of integrating knowledge representations in prior art patent retrieval. Key to the introduced approach is the utilization of human judgment available in the form of classifications assigned to patent documents. The paper first outlines in detail how a methodology for the extraction of knowledge from such a hierarchical classification system can be established. Further potential ways of integrating this knowledge with existing Information Retrieval paradigms in a scalable and flexible manner are investigated. Finally based on these integration strategies the effectiveness in terms of recall and precision is evaluated in the context of a prior art search task for European patents. As a result of this evaluation it can be established that in general the proposed knowledge expansion techniques are particularly beneficial to recall and, with respect to optimizing field retrieval settings, further result in significant precision gains.

1 Introduction

Identifying relevant prior art, i.e. trying to retrieve documents in patent and non-patent literature that are closely related to the matter described in a patent document, is probably the most commonly executed task in the patent domain. These searches form an essential part of the process of determining the patentability of a specific invention [2]. In order for an invention to be viable for patenting, no prior record of a similar or identical product or process may exist (See Section B IV 1/1 in [2] for a more detailed description). A prior art search therefore aims at clarifying whether any such records exist in patent and non-patent literature that have been published prior to the filing of the patent application in question. Since the erroneous granting of a patent can result in later litigation costs of hundreds of million Euros, extensive effort is invested into retrieving every relevant document. In this context the search for prior art constitutes a good example of a recall-focused task.

In this study we explore in what ways knowledge modeling and the integration of knowledge representations into the prior art retrieval task can be beneficial in light of these requirements. Modeling and representing knowledge has been widely researched in Cognitive Psychology [20] and Artificial Intelligence (AI) [12] as part of their quests to understand and replicate aspects of

human cognition. In the context of Information Retrieval (IR) the integration of knowledge has been explored as part of the Intelligent Information Retrieval (IIR) [10] and Associative Information Retrieval (AIR) initiatives [27] as a means of building more effective systems. With respect to recall-oriented tasks, mitigating vocabulary mismatch [14], i.e. to allow for the detection of semantic relatedness of documents where it is not reflected through the mutual occurrence of terms, represented a central aim of these approaches. While initial results obtained in these subdomains of IR have been promising, widespread adoption has been limited by a variety of factors. Of these the prohibitively high costs of manual knowledge representation creation, and lack of success concerning the automation of the process, proved to be most severe. As a result knowledge modeling related research in IR relies on the utilization of available knowledge artifacts such as thesauri [19], ontologies, and citations [27].

In the patent domain, a structure that can be interpreted in this sense is given by the International Patent Classification (IPC) system¹. In this system every new patent application is, with respect to technological aspects of the described invention, assigned to one or more classes within a hierarchy consisting of more than 70,000 different elements. These assignments are conducted by patent examiners based on their in depth knowledge of the respective technologies.

As a consequence this structure represents a highly precise hierarchical mapping of technological concepts, and provides an excellent basis for the extraction of knowledge. In light of this, the patent domain can be interpreted as an excellent new testbed for revisiting IIR and AIR related concepts.

The focus of this study is placed on evaluating the potential benefit of integrating knowledge representations into the prior art patent retrieval task. More specifically, inspired by research from Cognitive Psychology with respect to hierarchical aspects of memory, we propose to model knowledge in the form of a hierarchical conceptual structure extracted from available IPC information. In our chosen representation each element of the IPC hierarchy is comprised of a set of terms reflecting its characteristic vocabulary. To this cause a method aimed at extracting representative vocabularies for the technological aspects covered by specific IPC elements is developed. Based on this representation, strategies concerning the integration of the extracted knowledge into the retrieval task, with respect to the underlying aim of enabling the identification of similarity between a query and a document even in the absence of mutually shared terms, are investigated. Finally the potential benefit of these techniques is evaluated based on the prior art search task of the CLEFIP 09 collection.

The remainder of this paper is structured in the following way. Upon giving an overview of relevant previous research from AIR, IIR, and Patent Retrieval in Section 2, we explore the process of knowledge structure extraction based on the IPC in Section 3. In Section 4 an overview of our strategy concerning the integration of knowledge representations into the retrieval process is provided. Section 5 details the experimental setup chosen for the evaluation of our approach with respect to the prior art search task. In Section 6 we report and discuss the

¹ <http://www.wipo.int/classifications/ipc/en/>

obtained results. In the final section we present our conclusions and provide an outlook of future extensions to this work.

2 Related Work

In the following we will provide an overview of relevant research concerning Patent Retrieval and previous approaches of knowledge integration in IR.

The majority of relevant retrieval research in the patent domain has been pioneered by the NTCIR series of evaluation workshops and further diversified and expanded by the CLEF [1] Intellectual Property (IP) Track 2009 [26]. A task related to the prior art search task is presented by the invalidity search run at NTCIR 5 [13], and 6 [18]). Invalidity searches are exercised in order to render specific claims of a patent, or the complete patent itself, invalid by identifying relevant prior art published before the filing date of the patent in question. As such, this kind of search, that can be utilized as a means of defense upon being charged with infringement, is related to prior art search. Likewise to the prior art search task of CLEF IP 09, the starting point of the task is given by a patent document, and a viable corpus consists of a collection of patent documents.

The initial challenge with both tasks consists of the automatic formulation of a query w.r.t. a topic document. Commonly applied techniques [26] are based on the analysis of term frequency distributions as a means of identifying effective query term candidates. In addition to these techniques, the usage of bibliographical data associated with a patent document has been applied both for filtering and re-ranking of retrieved documents. Particularly the usage of the hierarchical structure of the IPC classes and applicant identities have been shown to be highly effective [13].

Concerning the integration of bibliographical data such as the IPC into retrieval, our work differs through its utilization of the IPC solely as a source for extracting term relation knowledge that is applied to mitigate the effect of vocabulary mismatch. As outlined in detail in Section 3 its aim lies in improving the query document matching process and it is not envisioned as a potential replacement, or exclusive of application of IPC based filtering or re-ranking methods. In a retrieval context where such information (i.e. IPC classification of the topic) is available these techniques could be applied on result listings returned by a retrieval setup as described in Section 5.

As pointed out before, the majority of relevant research in IR stems from the subdomains of Associative IR (AIR) and Intelligent IR (IIR). In AIR the most commonly applied scheme of applying knowledge consists of the construction of conceptual graphs and the application of spreading activation algorithms [9] as a means of expanding user queries. In IIR, which is focusing on the exploration of the 'overlap of AI and IR' [10], knowledge representations have been utilized in form of semantic networks [7] and hierarchies of retrieval subtopics [22]. A recent study [4] undertaken to evaluate the benefit of query (QE) and document expansion (DE) in the context of ad hoc-retrieval introduced two novel DE

methods based on adding terms to documents in a process that is analogous to QE, and on regarding each term in the vocabulary as a query. The study concluded state of the art QE to be generally beneficial and corpus-based DE in its applied form not to be promising.

Our work differs from these previously described approaches through its inherent focus on the patent domain, and by aiming at expanding documents with respect to their 'aboutness' [17] as expressed through their classification with specific IPC codes, and the resulting 'grouping' with related documents. In view of this, it can be interpreted as a form of meta-data based document expansion.

3 Constructing Knowledge Representations

This section initially recapitulates on the underlying motivation for the integration of knowledge modeling into patent retrieval. This will be followed by an overview of the knowledge representation utilized in this study. Finally key aspects of the IPC based knowledge extraction process are outlined.

As noted before, our proposed integration of knowledge representations into patent retrieval is primarily aimed at increasing retrieval effectiveness in terms of recall. A common approach to reach this goal consists of conceiving strategies to mitigate problems associated with vocabulary mismatch. Furnas et al. coined this concept based on 'the fundamental observation', 'that people use a surprisingly great variety of words to refer to the same thing' [14]. Attempting to limit potential negative impact can be interpreted as aiming at detecting semantic relatedness of textual artifacts where it is not reflected through the mutual occurrence of terms. In its most trivial form this could be achieved through the expansion of queries or documents with available synonyms. In a more complex form this can be interpreted as the attempt of mimicking the human ability to infer what documents 'are about', and to base decisions concerning their relatedness on this 'aboutness' [17]. This notion can best be illustrated through a basic example. Table 1 depicts a sample document collection consisting of the five documents *A* to *E*. In this example each document consists of only four terms.

In the following, given the hypothetical query q 'matrix collagenase-3 arthritis', we will illustrate potential rankings with respect to conventional best matching retrieval strategies and our envisioned knowledge expansion strategy.

A result list returned by a best match retrieval function such as TF/IDF or BM25 would consist of the documents *A*, *B*, *D*, and *E*. Document *A* would be ranked at 1 as it matches two of the query terms. Rank 2 to 4 would fall to documents *B*, *D*, and *E*. As is evident through the chosen sample texts of the documents and the explicit relevancy statements in Table 1 such a ranking does not represent an optimal outcome.

An optimal ranking would take the form A,B,C,D,E. The first three documents are related to 'arthritis', and in this example deemed relevant to the query. A human expert in possession of the knowledge that 'rheumatoid' and 'arthritis' are both terms describing medical problems affecting joints and connective

Table 1. Sample document collection. Matching query terms with respect to query 'matrix collagenase-3 arthritis' are highlighted in bold.

Doc.					Rel.
A	arthritis	cartilage	collagenase-3	specimen	Yes
B	antibody	matrix	metalloproteinase	osteoarthritis	Yes
C	immunohistochemical	nonfibrillated	rheumatoid	metalloproteinase	Yes
D	chondrosarcoma	matrix	metalloproteinase	vitro	No
E	machine	matrix	turing	zuse	No

tissue, and that 'collagenase-3' is a 'matrix metalloprotease', might deduce such an optimal ranking based on the following argumentation.

- *A* should be ranked first as it contains 'collagenase-3' and 'arthritis'.
- *B* should be ranked second as 'osteoarthritis' is a subtype of 'arthritis' and 'collagenase-3' is a 'matrix metalloprotease'.
- *C* should be ranked third since the term 'rheumatoid' is closely related to 'arthritis' and 'collagenase-3' is a 'matrix metalloprotease'.
- *D* should be ranked fourth since 'collagenase-3' is a 'matrix metalloprotease' and *D* therefore relates to one aspect of the query. While it is, as evidenced through the term 'chondrosarcoma' relating to another medical condition, it should still be ranked higher than *E* that is referring to a completely different topic.

Research from cognitive psychology indicates that such reasoning with respect to text is enabled through knowledge based term associating during the process of reading [28].

In order to enable an IR system to, in analogy to this, retrieve document *C* although it does not contain any of the query terms, two things would be necessary: Firstly the extraction of term relation knowledge (e.g. in this case mapping the relation between 'rheumatoid' and 'arthritis'), and secondly the integration of such a knowledge representation with the documents in the collection (i.e. to allow for consideration of this relationship during the retrieval phase).

With regard to the first point, concerning the question of choosing a suitable knowledge representation to benefit patent retrieval, research from cognitive psychology can provide additional guidance. Specifically the described hierarchical aspects [21] of human memory with respect to natural kinds (e.g. collagenase-3 is a matrix metalloprotease, a matrix metalloprotease is a protease, and proteases are enzymes) and artifacts (i.e. artificial concepts such as hard disk drive and tape being magnetic storage devices) seem relevant in consideration of the technical nature of the patent domain. This is further underlined by the reported role of categorical and hierarchical aspects of memory for inductive inference [8] and higher level extraction of meaning [23]. In view of this the choice of a hierarchical structure for storing knowledge seems sensible in regard to inference of relatedness. Such a structure is also well suited considering automatic knowledge extraction based on the available IPC information, as the IPC itself exhibits a

hierarchical order. Following this notion, extracted knowledge from the IPC will be modeled in a hierarchical manner that represents specific elements via sets of descriptive terms. By choosing to represent elements in a bag-of-words fashion, as outlined in detail in Section 4, this representation allows to expand documents in a flexible way that enables direct integration with existing retrieval models in a way that is scalable to realistic collection sizes. Subsequently a methodology for the extraction of such an above outlined representation of technological term relatedness based on IPC classification code assignments to patent documents will be introduced.

3.1 IPC Based Knowledge Representation Extraction

This section is focused on providing an overview of the proposed knowledge representation extraction process. As part of this we will also describe in detail the process of generating representative term sets for specific IPC elements.

The IPC is a hierarchical classification system comprised of roughly seventy thousand elements. These elements are spread over five main layers that are exemplary depicted in Figure 2 together with a sample IPC classification code. Each patent document is assigned to one or more elements within this hierarchy with respect to its described technical invention. To extract this knowledge, and allow for its representation in form of a hierarchical structure, a methodology comprised of 4 distinct steps is applied. Based on utilization of IPC codes found on patent documents, the methodology aims at representing elements of the IPC hierarchy by the most descriptive terms w.r.t the technological aspects covered by the documents filed to a common classification element. As will be subsequently described in detail the extraction process is based on the statistical analysis of two observed events: A pair of documents belonging to the same IPC element, and a pair of documents sharing a specific term. An overview of this process is depicted in Figure 1, and its four steps are listed below.

1. **Document Pair Formation:** For a given element E of the IPC hierarchy a representative set of N document pairs is formed out of the set of all documents assigned to this element.
2. **Mutual Term Extraction:** On completion of this process for each chosen pair of element E the set of all mutual terms is extracted. Requiring a term to occur in at least two documents belonging to E in order to be considered, represents the first selection within the task of extracting a set of terms representative for all documents of E .
3. **LL Ratio Computation:** To identify the mutual terms that are most representative of E we then apply the Log Likelihood (LL) ratio test on the basis of the extracted mutual terms of the N document pairs.
4. **Term Selection:** Finally a representative set of terms for the element E is selected by including all terms exhibiting a LL ratio score higher than a chosen threshold t .

The above described procedure is then repeated until a representative set of terms has been generated for each element of the IPC. The LL ratio test

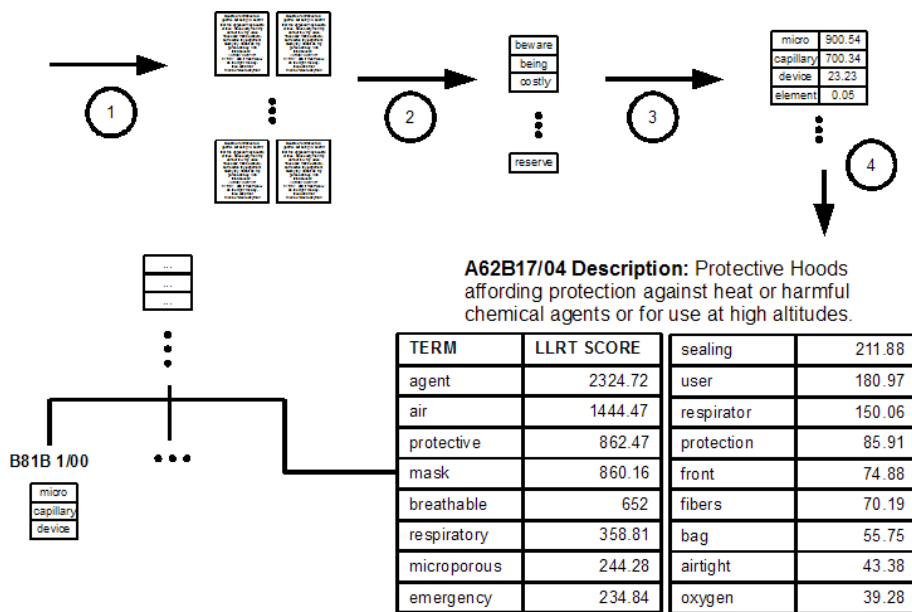


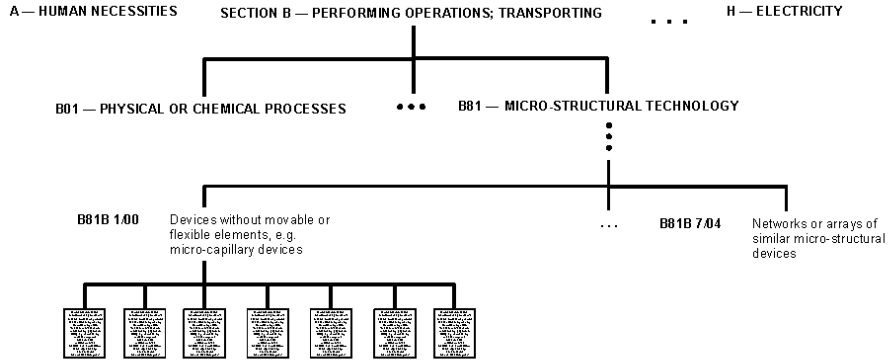
Fig. 1. Overview of the knowledge extraction process and a depiction of the representative vocabulary extracted for A62B17/04. The listed IPC description represents the complete descriptive text of the specific element.

performed in step (3) of our methodology is described in detail in the following subsection.

3.2 Extracting Representative Term Sets

The basic idea of our extraction approach is based on the identification of significant diversion in the statistical distribution of term occurrence frequencies within documents belonging to the same IPC element and documents in the rest of the corpus. In the following we will describe the applied process of estimating how much more likely the occurrence of a mutually shared term within a document pair belonging to the same element is in contrast to its occurrence in a pair in the rest of the collection. To this cause we devise a Log Likelihood Ratio test. Such tests have been widely applied to the task of collocation analysis due to good performance with respect to sparse distributions [6]. One advantage using likelihood ratios lies in their clear intuitive interpretation. For example, in our model a likelihood ratio of 900 expresses that the occurrence of a term within a pair of documents belonging to the same IPC element is 900 times more likely than its base rate of occurrence in document pairs in the rest of the collection would suggest. In the following we outline in detail how the Log Likelihood ratio test is applied w.r.t. our aim of extracting representative term sets.

In the space $W = \{d_1d_2, \dots, d_jd_k\}$ we observe two possible events t and e :



Section	Class	Subclass	Maingroup	Subgroup
B	81	B	7	04

Fig. 2. Exemplary overview of the IPC system and mapping of IPC code B81B7/04 to hierarchical level denominations

- **Event t :** A pair of documents $d_j d_k$ mutually contains t_i .
- **Event e :** A pair of documents $d_j d_k$ belongs to the same element.

Based on this two hypothesis can be formulated:

- **Hypothesis 1 (Independence):** H1: $P(t|e) = p = P(t|\neg e)$
- **Hypothesis 2 (Dependence):** H2: $P(t|e) = p_1 \neq p_2 = P(t|\neg e)$

With the log-likelihood ratio defined as:

$\log \lambda = \frac{L(H1)}{L(H2)}$ it can be computed as the fraction of two binomials b :

$$\log \lambda = \log \frac{b(c_{te}, c_e, p) * b(c_t - c_{te}, n - c_e, p)}{b(c_{te}, c_e, p_1) * b(c_t - c_{te}, n - c_e, p_2)}$$

Where c designates the observed counts of the events t , e , and te obtained in the third step (3) of our extraction methodology, and n represents all possible document pairs that can be formed from all N documents contained in a collection: $n = \frac{N!}{(N-2)! * (2)!}$

With respect to our goal of identifying representative terms, if the mutual occurrence of a term t in a pair of documents belonging to the same element e results in a large LL Ratio, we deem this term to be representative of the class.

4 Integration Strategy

For the integration of the generated knowledge representations described in the previous section we utilize the concept of document fields. In this approach both

the text of internal structural elements of a document such as the title, abstract, or passages, as well as external meta-data can be represented in form of a distinct field. This enables the computation of a separate score for each respective field during retrieval. These scores can then be aggregated under consideration of assigned weights with respect to the benefit of distinct fields to a retrieval task. In this form the approach has been successfully applied to a variety of different domains using document internal (e.g. title and body of an e-mail) and external fields (e.g. anchor text for web documents [29]). Basing our integration strategies on this technique allows us to build up on the large amount of research concerning combination techniques of scores [11] for individual fields and allows us to utilize existing retrieval system functionality.

In a similar fashion as HTML documents have been expanded with anchor text in previous research, we propose to expand each patent document in the corpus with representative sets of terms according to its classification within the IPC hierarchy. As exemplary outlined in Figure 3 for each document a field representing the text of the document itself, and a field for each hierarchy level can be created with respect to the assigned IPC code. As a consequence, indexing of the collection results in the creation of an index for each level of the IPC. This approach requires additional computational effort during the indexing phase, but enables the application of field-adapted retrieval models such as BM25F. BM25F [25] is a variant of the BM25 Okapi retrieval algorithm, that allows for the combination of scores from several fields in a way, that does not break the non-linear saturation of term frequency in the BM25 function. This form of aggregation has been shown to deliver strong results in field based experimentation in the news, web, and e-mail domain.







Field 0	Field 1	Field 2	Field 3	Field 4	Field 5
					
Patent Doc.	Subgroup Terms	Group Terms	Subclass Terms	Class Terms	Section Terms

Fig. 3. Integration through mapping extracted knowledge structure hierarchy levels via field indices

In the following section we will now describe how this strategy has been implemented in our experimental setup. Further an overview of the test collection and its main task will be provided.

5 Experimental Setup

This section introduces the experimental setup that was applied to evaluate the integration of our knowledge representation. We first provide details concerning the corpus and the associated task of the CLEF IP 09 [26] test collection that was used in our experimentation. Following this we will outline the applied indexing process and provide details of the retrieval models that were applied.

5.1 Test Collection

For the evaluation of our approach we used the CLEF-IP 09 collection that formed part of the CLEF evaluation workshop. The collection focuses on a patent retrieval task, and features thousands of topics that were created based on a methodology of inferring relevance assessments from the references found on patent documents [15]. The corpus of the collection consists of 1.9 million patent documents published by the European Patent Office (EPO). This corresponds to approximately 1 million individual patents filed between 1985 and 2000. As a consequence of the statutes of the EPO, the documents of the collection are written in English, French and German.

The main task of the CLEF-IP 09 track test collection consists of the search for Prior Art. Performed both, by applicants and the examiners at patent offices, it is one of the most common search types in the patent domain. Three sets of topics labeled as S, M, and XL consisting of 500, 1000, and 10000 topics are provided for this task. Each topic consists of an information need in the form of a patent application document and a set of qrels specifying relevant documents for the application. Based on the text of the patent application, participants of the track were required to infer a query in order to retrieve a ranked list of relevant prior art documents. The inference of effective queries formed the main challenge of the task. In this study we will utilize a methodology applied by participants of the 2009 track that is based on identifying effective query terms based on document frequency [16].

5.2 Retrieval Setup

Indexing of the collection is performed using the MG4J retrieval system [5]. For our document expansion strategy each document is indexed in the form of a set of field indices. While field index '0' represents the text of the document itself, knowledge in the form of representative terms is associated with one field index per hierarchy level. No form of stemming was applied. This decision was based on the fact that the corpus contains a large amount of technical terms (e.g. chemical formulas) and tri-lingual documents. In order to increase indexing efficiency, stop-wording based on a minimum stop-word list was applied.

Based on this setup we apply the following retrieval models. As baseline for our experiments the BM25 model is applied to a full text index of the collection. BM25 has shown strong performance in the CLEFIP 09 prior art track [24].

For the knowledge expanded field indices the BM25F model [25] is applied. Essentially identical to BM25 it has shown very good performance in field based scenarios due to the ability to set a field specific normalization parameter b in addition to applying weighting of each index [29].

For the initial retrieval runs and optimization the large training topic set consisting of 500 query documents has been applied. The performance with optimized parameter sets was evaluated based on the medium sized (1000) topic set. As stated before the main challenge of prior art search initially consists of the automatic extraction of effective query terms based on a topic posed in form of a several pages long patent application. State of the art automatic query formulation methods rely on choosing query terms with respect to the distribution of term features. One such feature that has been applied to the task of automatic query formulation consists of the global document frequency (df) of terms. It has been found, that effective queries can be formulated by including only those terms of a patent application that occur in a low percentage of the documents in the collection. A query selection parameter called percentage threshold is defined as $\frac{df}{N} * 100$, where N denotes the total number of documents in the collection. A percentage threshold of 0.5% therefore denotes, that all terms from a topic document are included in a query that appear in less than 0.5% of the documents in the collection.

In light of this the experimental evaluation of the knowledge integration is divided into two parts:

- **Query Dependency Analysis:** Induced by the lack of one definitive set of queries as for example encountered in Web domain based tracks such as HARD [3], a first step in estimating the benefit of knowledge integration consists of an evaluation of the performance of various knowledge representations with respect to a varied set of generated queries. This step is aimed at generally clarifying if, and in what ways, knowledge based expansion can impact the prior art retrieval task.
- **Parameter Set Optimization:** In order to more precisely estimate the potential benefit of the integration, we propose to conduct parameter set optimization for query generation settings that have shown promising results in the first experimentation phase. The need to optimize the parameter sets of BM25F based retrieval attempts in order to allow for best performance is discussed and outlined in detail in [29].

The results of these investigations are outlined in the subsequent section.

6 Experimental Results and Discussion

In the following the results for both experimentation phases will be outlined.

6.1 Query Dependency Analysis

The results of the query dependent analysis with respect to three generated knowledge representations are depicted in Figure 4 and 5. The graphs outline

performance in terms of Recall and MAP w.r.t. the association of knowledge representations extracted based on LL Ratio thresholds of 15, 300, and 900. A baseline is provided by the BM25 model operating on a full text index. In these initial experiments only the lowest level of the IPC (subgroup) is considered for the expansion.

Evident from Figure 4 is that the expansion with subgroup level fields benefits recall over all query generation parameters. The observation of a positive effect on recall being retained among the full spectrum of query formulations is a very promising result. Since queries created with a df-threshold of 0.25 exhibit an average query term length of 103.838 in contrast to an average length of 259.548 terms for a 3.25 query-threshold, the applied set of queries represents not only a large spectrum with respect to length, but also in regard of the document frequency of the contained query terms. In light of this it seems reasonable to assume, that the observed positive effect and its robustness might also apply to related query generation methods such as TF/IDF, and potentially also to manually created prior art queries. Further it can be deduced from Figure 4 that the LLRT 15 based representation exhibits the highest amount of variance with regard to the observed performance. Responsible for these results may be application of a comparatively low threshold of 15 in order to select knowledge representation terms. Generally this will lead to the inclusion of more general terms within the representative term set of the modeled subgroup level elements. It seems likely that the inclusion of more general knowledge terms raises the probability of topic drift occurrence. The stricter term selection criteria set by LLRT likelihoods of 300 and 900 in contrast exhibit more robust performances.

The negative impact of this observation becomes also evident by studying the MAP related performance shown in Figure 5. While the contained vocabulary of the LLRT 15 based knowledge representation seems still descriptive of the general technological aspects, as expressed in the higher recall with respect to the baseline, substantial noise seems to be introduced, resulting in a clearly visible negative impact on precision. This is not exhibited by the more strict LLRT300 and LLRT 900 knowledge representations, which again show lower variance in their observed performance. Promising with respect to precision is the exhibited strong MAP performance of the LLRT 300 and LLRT 900 based runs for percentage-threshold values of 0.5 and 0.75. In order to estimate their full potential benefit, linear optimization of the BM25 and BM25F parameter sets was performed based on the large training topic set (500) of the Clef-IP 09 collection.

6.2 Parameter Optimization for BM25F

Based on the above reported initial results a complete linear optimization of BM25 and BM25F parameters was performed for a percentage-threshold value of 0.75. Training of the parameters for BM25F followed the strategy of dividing the optimization of $k1$, and index specific b and w parameters into several smaller optimization tasks as outlined in Zaragoza et al. [29]. Table 2 lists the results.

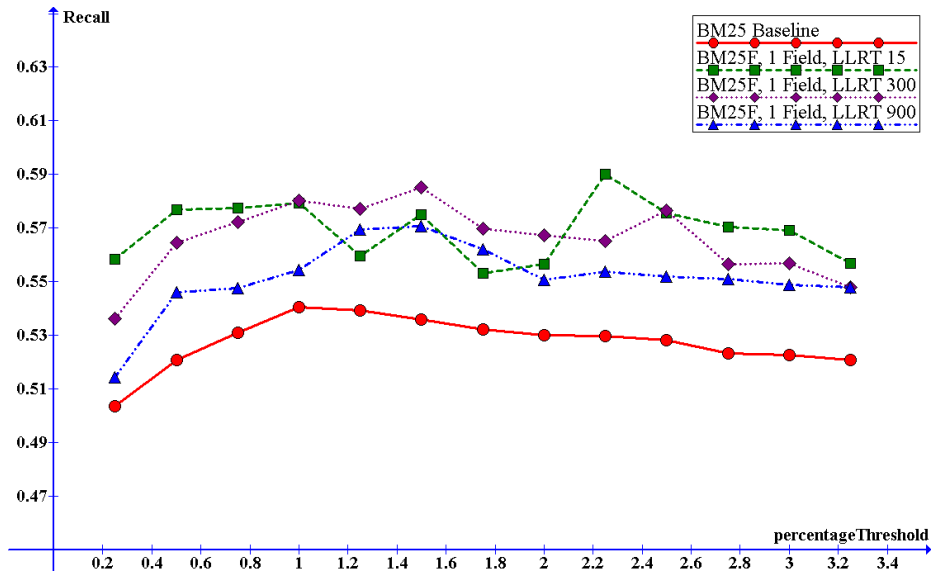


Fig. 4. Recall over various query length and LLRT thresholds for 1 field expansion

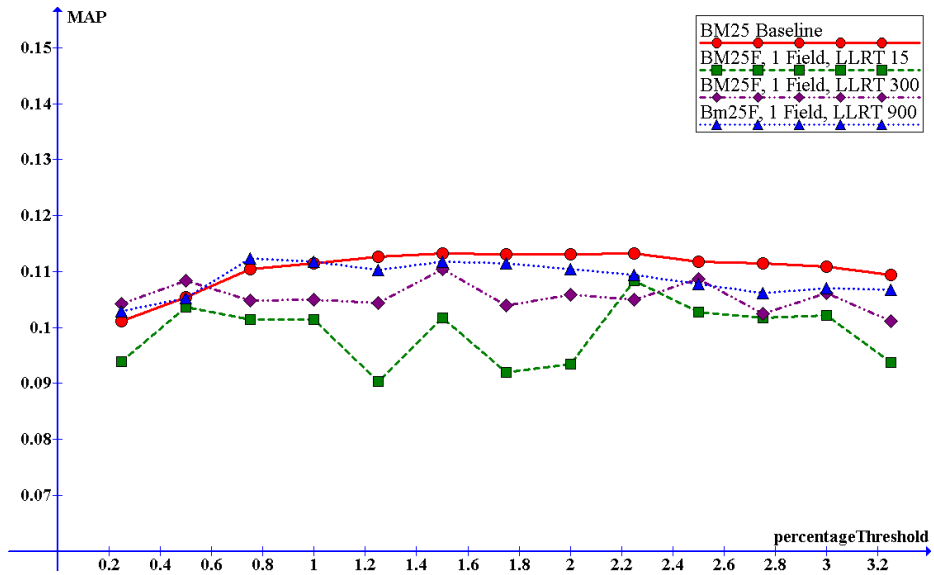


Fig. 5. MAP over various query length and LLRT thresholds for 1 field expansion

Table 2. Optimized results BM25 baseline versus BM25F knowledge expansion. $b1$ and $w1$ constitute the parameter values for the full text index; $b2$ and $w2$ represent parameters for the subgroup based knowledge field (** strong statistical significance)

Query 0.75	MAP	% change	Recall	% change	Bpref	k1	b1	b2	w1	w2
BM25	0.1036	/	0.5377	/	0.5544	1.2	0.4	/	/	/
BM25F,LLRT300	0.1073	3.57**	0.5887	9.48**	0.6083	1.0	0.4	0.65	1.0	0.3
BM25F,LLRT900	0.1091	5.31**	0.5819	8.22**	0.6038	0.2	0.4	0.7	1.0	1.0

As can be seen from the table the optimization results in a much improved performance for the knowledge expanded BM25F runs. MAP as well as recall are substantially and statistically significantly increased for both the LLRT300 and LLRT900 based knowledge representation. This constitutes an especially promising result in light of the sparse exploration with respect to optimization of the LLRT threshold space.

7 Conclusion and Future Outlook

A first study of modeling knowledge within the task of prior art patent retrieval was presented. Initial results are very promising as it is shown that the proposed knowledge association is beneficial in terms of recall, and very robust with regard to query variation. Given prior optimization of the BM25F parameter space the knowledge association results in significant improvement of both recall and precision in comparison to an in the same manner optimized BM25 baseline.

This is specifically encouraging, since in the introduced work only the lowest level of the extracted knowledge representation, corresponding to the subgroup IPC hierarchy level, has been utilized. Integration of higher hierarchy levels constitutes a logical next step and could potentially result in further improvements. Moreover a more fine-grained exploration of the term extraction parameters, and the application of varying methods for representative term selection, merit extensive additional investigation. Further an exploration of n-gram representations, proximity, and document structure exploitation within the extraction and retrieval process should be considered. Finally an evaluation of the potential benefit of the introduced knowledge representations with respect to other tasks such as classification and clustering and the feasibility of applying the introduced techniques to other domains form interesting long term aspects of this research.

8 Acknowledgments

The authors would like to thank Matrixware Information Services² and the Information Retrieval Facility³ (IRF) for their support of this work. Mounia Lalmas is currently funded by Microsoft Research/Royal Academy of Engineering.

² <http://www.matrixware.com>

³ <http://www.ir-facility.org/>

References

1. The Cross-Language Evaluation Forum (CLEF).
2. *Guidelines for Examination in the European Patent Office*, December 2007.
3. ALLAN, J. HARD track overview in TREC 2004: High accuracy retrieval from documents. In *In Proceedings of the thirteenth Text REtrieval Conference (TREC 2004)* (2004), no. Ldc, NIST, pp. 1–11.
4. BILLERBECK, B., AND ZOBEL, J. Document expansion versus query expansion for ad-hoc retrieval. *Proceedings of the 10th Australasian Document Computing Symposium* (2005).
5. BOLDI, P., AND VIGNA, S. MG4J at TREC 2005. In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, number SP* (2005), vol. 500, Citeseer, p. 266.
6. BORDAG, S. *Elements of Knowledge-free and Unsupervised lexical acquisition*. PhD thesis, 2007.
7. COHEN, P. Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management* 23, 4 (1987), 255–268.
8. COLEY, J. Knowledge, expectations, and inductive reasoning within conceptual hierarchies. *Cognition* 90, 3 (2004), 217–253.
9. CRESTANI, F. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review* 11, 6 (1997), 453482.
10. CROFT, W. Approaches to intelligent information retrieval. *Information Processing & Management* 23, 4 (1987), 249254.
11. CROFT, W. Combining approaches to information retrieval. *Advances in information retrieval* 7 (2000), 136.
12. DAVIS, R., SHROBE, H., AND SZOLOVITS, P. What is a knowledge representation? *AI magazine* 14, 1 (1993), 17.
13. FUJII, A., IWAYAMA, M., AND KANDO, N. Overview of Patent Retrieval Task at NTCIR-5. In *Proceedings of NTCIR-5 Workshop Meeting* (2005).
14. FURNAS, G., LANDAUER, T., GOMEZ, L., AND DUMAIS, S. The vocabulary problem in human-system communication. *Communications of the ACM* 30, 11 (1987), 971.
15. GRAF, E., AND AZZOPARDI, L. A methodology for building a patent test collection for prior art search. In *Proceedings of the Second International Workshop on Evaluating Information Access (EVIA)* (2008).
16. GRAF, E., AZZOPARDI, L., AND VAN RIJSBERGEN, K. Automatically Generating Queries for Prior Art Search.
17. HUTCHINS, W. The concept of aboutness in subject indexing. *Aslib Proceedings* 30, 5 (1978), 172181.
18. IWAYAMA, M., FUJII, A., AND KANDO, N. Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task. In *Proceedings of NTCIR-6 Workshop Meeting* (2007), pp. 366–372.
19. JING, Y., AND CROFT, W. An association thesaurus for information retrieval. In *Proceedings of RIAO* (1994), vol. 94, Citeseer, p. 146160.
20. KINTSCH, W. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review* 95, 2 (April 1988), 163–82.
21. MARKMAN, E., AND CALLANAN, M. *An analysis of hierarchical classification*. Hillsdale, NJ: Erlbaum, 1980, pp. 325–366.
22. McCUNE, B., TONG, R., DEAN, J., AND SHAPIRO, D. RUBRIC: a system for rule-based information retrieval. *Readings in information retrieval*, 9 (1997), 445.

23. MEDIN, D. L., AND RIPS, L. J. *Concepts and categories: Memory, meaning, and metaphysics*. Cambridge Univ Press, 2005.
24. PIROI, F., RODA, G., AND ZENZ, V. CLEF-IP 2009 Evaluation Summary, 2009.
25. ROBERTSON, S., ZARAGOZA, H., AND TAYLOR, M. Simple BM25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management* (New York, NY, USA, 2004), ACM Press, pp. 42–49.
26. RODA, G., TAIT, J., PIROI, F., AND ZENZ, V. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. *CLEF working notes 2009* (2009).
27. SALTON, G. Associative document retrieval techniques using bibliographic information. *Journal of the ACM (JACM)* 10, 4 (1963), 440457.
28. WHARTON, C., AND KINTSCH, W. An overview of construction-integration model. *ACM SIGART Bulletin* 2, 4 (1991), 169173.
29. ZARAGOZA, H., CRASWELL, N., TAYLOR, M., SARIA, S., AND ROBERTSON, S. Microsoft Cambridge at TREC-13: Web and HARD tracks. In *Proceedings of TREC 2004* (2004), Citeseer.