Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies

Duy Dinh, Lynda Tamine, Fatiha Boubekeur

Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies

Duy Dinh [a,*], Lynda Tamine [a], Fatiha Boubekeur [b]

[a] Institut de Recherche en Informatique de Toulouse, Paul Sabatier University, 31062 Toulouse, France
[b] Department of Computer Science, Mouloud Mammeri University, 15000 Tizi Ouzou, Algeria

A B S T R A C T

Objective: The aim of this work is to evaluate a set of indexing and retrieval strategies based on the integration of several biomedical terminologies on the available TREC Genomics collections for an *ad hoc* information retrieval (IR) task.
Materials and methods: We propose a multi-terminology based concept extraction approach to selecting best concepts from free text by means of voting techniques. We instantiate this general approach on four terminologies (MeSH, SNOMED, ICD-10 and GO). We particularly focus on the effect of integrating terminologies into a biomedical IR process, and the utility of using voting techniques for combining the extracted concepts from each document in order to provide a list of unique concepts.
Results: Experimental studies conducted on the TREC Genomics collections show that our multi-terminology IR approach based on voting techniques are statistically significant compared to the baseline. For example, tested on the 2005 TREC Genomics collection, our multi-terminology based IR approach provides an improvement rate of +6.98% in terms of MAP (mean average precision) ($p < 0.05$) compared to the baseline. In addition, our experimental results show that document expansion using preferred terms in combination with query expansion using terms from top ranked expanded documents improve the biomedical IR effectiveness.
Conclusion: We have evaluated several voting models for combining concepts issued from multiple terminologies. Through this study, we presented many factors affecting the effectiveness of biomedical IR system including term weighting, query expansion, and document expansion models. The appropriate combination of those factors could be useful to improve the IR performance.

1. Introduction

Over the past few decades of the last century and the first decade of the present one, the growing amount of scientific literature in genomics and related biomedical disciplines has led to an increase in research and development of effective methods for searching and accessing biomedical literature and full-text journal articles [1–6]. A lot of research in the field of information retrieval (IR) aims at improving the quality of search results. However, traditional IR systems are facing new challenges due to the large amount of information and the special language usage in biomedical literature, such as the frequent occurrences of synonyms, acronyms, abbreviations, gene symbols, . . . in citations or full-text articles.

It is well known that in domain-specific IR, concepts predefined in ontologies constitute a relevant source of knowledge for indexing documents by alleviating the term mismatch problem faced by IR systems [7,8,6]. The indexing task can be based on one terminology (*mono-terminology indexing*) or several terminologies (*multi-terminology indexing*). While the former is usually based on MeSH, the latter is based on different terminologies (e.g., ICD-10, SNOMED-CT, GO [8,6,9], CCAM, TUV [10,11]). In general, indexing can be done manually or automatically. Manual indexing is undertaken by human experts with specialized knowledge of terminologies and many years of experience. Automatic indexing is less likely to be expensive in terms of costs and time and thus could be an alternative for helping the manual task. For example, MTI (Medical Text Indexer) is a program for producing MeSH indexing recommendations. It is the major product of NLM's Indexing Initiative and has been used in both semi-automated and fully automated concept-based indexing [2]. Concept-based indexing and retrieval are mainly supported by the task of concept extraction, which is one of the important techniques in natural language processing for identifying concepts in controlled terminologies [12].

Although the MeSH thesaurus has been widely accepted as the main controlled vocabulary to index biomedical documents,

* Corresponding author.
E-mail addresses: Duy.Dinh@irit.fr (D. Dinh), Lynda.Tamine@irit.fr (L. Tamine), amirouchefatiha@mail.ummto.dz (F. Boubekeur).

it is so far unable to cover all medical terms in all the domains of medicine [13]. For instance, a growing interest has arisen in the use of SNOMED as a standard for the electronic health record and content coverage [14,9]. This problem has been faced so far by adopting two main approaches. The first one aims at building (semi)-automatic mappings between terminologies [15,16]. The second one, most related to our work, attempts to extract concepts from different terminologies in order to better cover the subject matter(s) of the document and so to improve the recall of concept extraction [17,10,11,9].

In this paper, we consider a different and novel multi-terminology based concept extraction approach to indexing and retrieving biomedical information. More specifically, our approach is inspired by the principle of poly-representation in IR [18] making simultaneous combination of evidences that are cognitively different in order to increase the information value of documents via concept extraction using several terminologies. In particular, we consider concept extraction as a voting problem taking into account both the scores and ranks of identified concepts predefined in different terminologies (viewed as sources of evidence). More specifically, using the list of concepts extracted from each document when applying an approximate concept extraction on each terminology, we propose first to weight concepts using a particular term weighting model. Second, we propose to merge the candidate concept lists by means of a voting process [19]. We focus in this paper, on the study of the impact of several sources of evidence, such as concept weighting schema, document relevance model on the retrieval effectiveness.

The remainder of the paper is organized as follows: Section 2 introduces keyword and concept based indexing in IR general domain. Section 3 provides an overview of related work in biomedical IR dealing with concept extraction and concept-based indexing and retrieval. Section 4 gives a brief overview of our contribution to evaluate the utility of multi-terminology-based indexing and retrieval. Section 5 presents our concept-based IR framework taking into account several domain knowledge sources or terminologies. Sections 6 and 7 describe our experimental methodology and results. We then discuss several aspects of our multi-terminology IR approach in Section 8. Finally, we conclude the paper and outline research directions for future work.

2. Word based *vs.* semantic based document indexing and retrieval

Information retrieval (IR) is concerned with selecting, from a collection of documents, those that are likely to be relevant to a user information need expressed using a query [20]. Three basic tasks are carried out in an IR system: document representation, query representation and matching of these representations. Document representation is usually called indexing. The main objective of indexing is to assign to each document a descriptor represented with a set of features automatically or manually derived from the document content. Representing the query involves a one step or a multi-step query formulation by means of prior terms expressed by the user and/or additive information driven by iterative query improvements like relevance feedback [21]. The main goal of document–query matching is to estimate the relevance of documents to the given query. Most of IR models handle during this step, an approximate matching process using the frequency distribution of query terms over the documents to compute a relevance score called Relevance Status Value (RSV). This latter is used as a criterion to rank the list of documents returned to the user in response to his query. Regarding information representation, we mainly distinguish between word-based and concept-based IR approaches.

2.1. Word based document indexing and retrieval

Traditional IR systems are based on the *bag of words* expressing the fact that both documents and queries are represented using basic *words* commonly called keywords or terms. A keyword may be a simple word (as in "*computer*") or a multi-word (as in "*computer science*"). Weights assigned to document and/or query terms express their importance in the considered information unit. The weighting model is generally based on the well-known TF*IDF schema [20].

A key characteristic of traditional IR systems is that the degree of document–query matching depends on the number of the shared keywords. Indeed, in such IR systems, a relevant document will not be retrieved in response to a query if the document and query representations do not share at least one word. However, it is well known that the query is usually an incomplete and vague description of the user information need and authors of documents use a very wide vocabulary to express the same concepts. This leads to the critical issue called *keyword-barrier* [22] mainly due to term ambiguity [23] and term disparity [24].

- *Ambiguity* refers to the existence of multiple interpretations of the specific meaning that a word assumes in context at the *syntactic* or *semantic* levels [25]. Syntactic ambiguity refers to differences in syntactic categories (e.g., noun, verb, adjective, etc.). Semantic ambiguity refers to *homonymy* or *polysemy*. Homonymy traduces the fact that some terms, represented by the same lexical word, have different meanings within different contexts, e.g., "*France*" in "*Anatole France*" *vs.* "*France*" in *politics*. Polysemy traduces the fact that a word could have different meanings. *Opening the door* versus *opening of a theatre piece* is an example of polysemy. In traditional IR systems, ambiguity induces noisy results. For instance, a document on *Anatole France*, nevertheless not relevant for a query on *politics in France*, will be retrieved because of the (homonym) shared word *France*.
- *Disparity* refers to term mismatch [24]. Term mismatch means that document terms do not match query terms, even if the document is relevant to the query. This problem is due to linguistic variations that could be either morphological, lexical, syntactic or semantic between terms used to express the query, and those used in the documents. Synonymy is one of the main origins of term mismatch. For instance, a document on "*universe*", nevertheless relevant for a query on "*cosmos*", will not be retrieved if the word "*cosmos*" does not occur in this document.

The keyword barrier problem has serious drawbacks at both document representation and query formulation levels, which are likely to dramatically decrease the IR effectiveness [22]. To cope with this problem, several works have focused on enhancing both query and document representations by means of *word sense* notion. This latter is the in-context usage meaning used to tag the index as outlined below.

2.2. Semantic based indexing and retrieval

Attempts in document representation improvements are related to the use of semantics in both indexing and retrieval steps. It is well known that word senses can not be easily discretized, that is, reduced to finite senses [26]. The main challenge in semantic indexing is therefore to identify the sense *ground truth* in order to assign the accurate meaning within the accurate context in the document. Regarding this issue, semantic based indexing and retrieval approaches are ranged into corpus-based and knowledge based ones.

2.2.1. Corpus-based semantic indexing and retrieval

According to this approach, senses are extracted from the semantic structure of the documents in the corpus as a whole or learned using prior patterns. We distinguish statistical based approach to machine learning approach. The statistical approach generally relies on co-occurrence analysis and linear algebra theory in order to discover and extract implicit senses from document contents. The most popular statistical approach for semantic indexing is Latent Semantic Indexing (LSI) [27] that computes semantic aggregated dimensions by semantically clustering close words via Singular Value Decomposition (SVD) based dimensional reduction of the term-document matrix. Each term cluster represents a sense within the document context. Machine learning (ML) [28] uses manually labelled corpus for training classifiers, which basically try to learn several features for binding terms from text to predefined classes. Each class represents a semantic entity (the implicit sense) assigned to words within the document.

2.2.2. Knowledge-based semantic indexing and retrieval

Knowledge-based indexing makes use of external resources in order to identify explicit senses [29,30]. A description of all the resources used for semantic indexing is out of the scope of this paper. Here, we focus on structured linguistic resources such as machine readable dictionaries, thesauri and ontologies. Such resources are likely to provide almost a wide range of meanings assigned to words in different usage contexts as well as a set of semantic relations between them. A specific meaning covers an entry of the resource through the basic notions of *concept*, *synset*, *term* (or multi-term)... In general domain, a very popular used resource is *WordNet* lexicon [31] that encodes concepts in terms of sets of synonyms (called *synsets*). For each synset, WordNet provides several lexical and semantic relations (e.g. Antonymy, Hypernymy). In the biomedical domain, several terminologies such as MeSH, UMLS and SNOMED have been provided as a description of domain knowledge and used for assigning senses to flat textual descriptions. Each concept in these knowledge sources is represented by one "**concept identifier**" to which are linked one preferred term and several non-preferred terms.

Knowledge based indexing task runs in two main steps: concept extraction and concept weighting.

- *Concept extraction*. The objective of this step is to link terms in documents to accurate entries (or concepts) in the resource. For this aim, representative terms are first identified in the document using classical indexing techniques (tokenization, lemmatization) and then mapped to the accurate resource entries. Ambiguity occurs also in knowledge-based semantic approaches, when a polysemic term is mapped to several entries (senses) of the resource. Word sense disambiguation techniques generally exploit local contexts and definitions from the resource [32,33]. The underlying idea is to estimate the *semantic relatedness* between each sense associated with the target word to other senses from its local context. Formally, the disambiguation process relies on the computation of a score assigned to each concept (entry)considering the semantic distance to other concepts (entries) in the document context. The best scored concept is selected retained as the accurate sense of the term within the document.
- *Concept weighting*. Concept weighting aims at evaluating the importance of concepts in a document content. The importance score can either be estimated statistically, through the estimation of its frequency distribution within a document using an extended version of the classical TF*IDF schema [34] or a more elaborated variant one based on structure based importance [35].

Our work is in this context, and mainly consists in evaluating multiple knowledge-based (also called terminology-based) semantic indexing strategies on textual documents from the biomedical literature. An overview of biomedical literature knowledge-based indexing and retrieval is introduced in the following section.

3. Terminology-based biomedical indexing and retrieval

In a specific domain, e.g., biomedical IR, the idea of using concepts from controlled vocabularies comes from the fact that they are able to cover different instances related to a given idea in a domain such as synonyms, abbreviations, etc. Indeed, the great majority of journals reporting significant research work in biomedicine are selected for indexing in MEDLINE[1] after careful review based on several issues such as the journal's scope and coverage as well as the quality of its scientific content and editorial work. A distinctive feature of MEDLINE is that documents are indexed with MeSH terms denoting domain concepts. The purpose of performing a MEDLINE search (retrieval) through the NLM's PubMed portal is to identify relevant articles for the question of interest (the user query). PubMed translates the user's initial query and automatically adds field names, relevant MeSH terms, synonyms, Boolean operators, and 'nests' the resulting terms appropriately to construct a boolean query. However, most boolean IR systems represent a major disadvantage that there is no weighting of index or query terms, which constitutes a ranking strategy of an IR system.

The limitations of the boolean search systems have led to the research and development of several IR ranking strategies by exploiting the collection statistics such as within-document term frequency, document frequency... leading to a substantial improvement in retrieval quality [20]. Early IR approaches incorporating document and query features into a ranking function include Salton's SMART system [36] and the NLM's IRX Project [37]. As stated above, traditional IR systems use indexing and retrieval based only on individual keywords, causing the term mismatch problem in IR since biomedical information is organized at the level of concepts and not in separated words [6,38,39].

To cope with the limitations of word-based indexing and retrieval approaches, within a specific domain, e.g., biomedical IR, works on conceptual indexing have been extensively studied in literature [40–42,2,3,43–45]. One of the first attempts to automatically index and retrieve biomedical information at the level of full medical concepts was undertaken in the SAPHIRE system [40,41]. Similar to the manual indexing of MEDLINE database undertaken by human indexers, the principle goal of the system is to automatically assign MeSH terms to MEDLINE citations by reducing or completely replacing human efforts in terms of costs and time to manually index documents. Furthermore, the output of retrieved documents in response to a user query is ranked by its relevance.

Conceptual indexing and retrieval approaches, i.e. the use of concepts in controlled vocabularies for IR, relies on the key component of concept extraction or concept mapping to find relevant concepts from documents and/or users' queries. State-of-the-art concept extraction methods can be subdivided into three categories: (1) *rule-based* methods, (2) *dictionary-based* methods and (3) *statistical* methods. In what follows, we describe the dictionary-based method, which will be used as the baseline method in our experiments. For a more detailed categorization of these methods, we refer to the work in [42].

[1] The U.S. National Library of Medicine's (NLM) premier bibliographic database that contains over 20 million references to journal articles in life sciences with a concentration on biomedicine.

One of the first work on concept extraction from biomedical text by exploiting external terminological resources viewed as a dictionary of terms denoting concepts was the concept finding algorithm implemented in the SAPHIRE system [46]. The concept finding algorithm tries to look up all synonyms of each word in a text and map all possible combinations of synonymous words to create a multi-word term that must be compared with entries in a dictionary of terms. Inspired by this idea, MetaMap is a program developed at the NLM to map biomedical text to concepts in the UMLS Metathesaurus [47]. For each phrase (group of consecutive words), variants are generated using the knowledge in the SPECIALIST lexicon in UMLS and a supplementary database of synonyms. A variant of a word phrase consists of its acronyms, abbreviations, synonyms, derivational variants, meaningful combinations of these, and finally inflectional and spelling variants [47]. Candidate terms denoting UMLS concepts are retrieved if they contain at least one of the generated variants. MetaMap becomes later the main component of the MTI tool [2], which consists of several methods for creating a list of recommended indexing terms: MetaMap Indexing, Trigrams and PubMed Related Citations. MTI provides in the first stage several lists of UMLS concepts and then restricts to MeSH concepts using the mappings between UMLS and MeSH. The work in [42] suggested a method based on an approximate string matching to recognize gene and protein names. In their approach, both protein dictionaries and target text are encoded using the nucleotide code (A, C, G, T). Then, the alignment techniques of DNA and protein sequences in databases are applied to the converted text in order to identify character sequences that are similar to existing gene and protein names. Authors in [48] proposed an approximate dictionary lookup, namely **MaxMatcher**, to cope with term variations. The basic idea of their approach is to capture the significant words instead of all words of a particular concept. For example, the token "gyrb" is obviously important to the concept "gyrb protein"; MaxMatcher is able to recognize it as a concept name even if the token "protein" is not present. In a comparative study, their approximate extraction method reached a 71.60% precision and a 75.18% recall while exact matching only reached a 54.97% precision and a 57.73% recall.

4. Our contribution within terminology-based indexing and retrieval for biomedical IR

The main contribution of our work relies on the following key points:

- First, we evaluate the utility of integrating terminologies as domain knowledge sources into the process of biomedical indexing and retrieval. We compare the IR performance when using only one terminology (mono-terminology indexing) to the baseline without using any terminology. Then, we evaluate and discuss the IR performance of our conceptual indexing and retrieval method when using multiple terminologies (multi-terminology indexing) compared to the one which is solely based on one terminology. The mono-terminology indexing is based on a dictionary-based concept extraction while the multi-terminology indexing is based on a multi-terminology concept extraction, which employs several voting techniques for fusing concepts issued from several terminologies into a single list of unique concepts. In addition, we propose to weight concepts using a particular weighting model.
- Second, we conduct a series of indexing and retrieval scenarios to evaluate the utility of taking into account the following sources of evidence: (1) the knowledge about the document, i.e., extracted concepts that are predefined in domain knowledge sources, which we refer to as "global context", (2) the knowledge about the user query, i.e., extracted terms from top-ranked documents, which we refer to as "local context". By "context", we mean the source of information where terms are extracted to expand the document content or the original query. "Global context" is independent from document/query while "local context" is dependant on the document/query.

For the first contribution, to the best of our knowledge, there are two main categories of works in the biomedical domain dealing with multiple terminologies: the one focuses on the use of multiple terminologies only for indexing biomedical documents [10,11], while the other only focuses on expanding the user query via a process of termino-ontological query expansion [49,6,38]. However, there is so far no work investigating the **evaluation of the multi-terminology document indexing** for *biomedical information retrieval*.

Concerning the multi-terminology indexing, the work in [10] proposed a multi-terminology indexing approach based on the bag-of-words concept extraction. In their approach, each sentence in the document is represented as multiple bags of words independently of the word order correlation between words in the sentence and the ones in concept names. Similarly, authors in [11] presented a multi-terminology approach to indexing documents in the CISMeF portal but the concept extraction between free text and terminologies is based on the simple bag-of-words representation. Indeed, in their concept extraction method, the "bag of words" representing each candidate term is matched independently of the word order against all the MeSH, ICD10, SNOMED, CCAM and TUV terms that have been processed in the same way. Such a method could not give a high precision since it is not able to capture the order or words in both terms in documents and candidate concept names. In the context of data mining, the work in [9] presented a multi-view approach based on multiple terminologies (GO, MeSH, eVOC, OMIM, LDDB, KO, MPO, SNOMED CT, and UniprotKB) to investigate the effect of using multi-source algorithms (kernel fusion for clustering) [50] to undertake two fundamental computational disease gene identification tasks: gene prioritization and gene clustering. They concluded that in practice research the relevance of specific vocabulary pertaining to these tasks is usually unknown. In such case, multi-view text mining is a superior and promising strategy for text-based disease gene identification.

Concerning works dealing only with query expansion using multiple domain knowledge sources, the work in [49] identified unique concepts issued from several terminologies such as MeSH, Entrez Gene and ADAM for expanding the original query with the synonyms, hypernyms, hyponyms, lexical variants and implicitly related concepts. MeSH terms denoting concepts are identified using the Pubmed's Automatic Term Mapping (ATM) service, which basically maps untagged terms from the user query to lists of pre-indexed terms in Pubmed's translation tables (MeSH, journal and author). Gene symbols are identified using a set of manually handcrafted heuristics relating to the lexical variation of gene names [51] or a statistical learning algorithm logistic regression to score abbreviations based on their resemblance to previously identified ones [52]. Similarly, authors in [6] exploited several medical knowledge sources such MeSH, Entrez gene, SNOMED, UMLS, etc. for **query expansion** by expanding the original user query with synonyms, abbreviations and hierarchically related terms identified using PubMed. For this task, they have submitted the natural language query to PubMed and used information presented in the *details* tab, to obtain a parsed version of the query which they used to extract concepts to expand the original query. Furthermore, they defined several (hand-coded) rules for filtering the candidate terms according to each knowledge source.

Motivated by recent work on concept extraction and information retrieval, especially in the biomedical domain, in this paper, we

study the impact of several factors on biomedical multi terminology based document retrieval effectiveness. In particular, in order to index documents with conceptual information, we propose a novel multi-terminology based concept extraction from each document: first, we use an approximate concept extraction method to identify concepts in each document using a mono terminology. Candidate concepts are weighted to measure their relevance to the document using a particular term weighting model (e.g., probabilistic model); second, we apply the concept extraction algorithm on several terminologies and combine several concept lists using several voting techniques. We see each concept identified from each document using multiple terminologies as an implicit vote for the document. Therefore, the multi-terminology based concept extraction can be modelled as a voting problem.

The second point of our contribution concerns the evaluation of IR performance using domain knowledge sources for document expansion and query expansion. Here, we study the impact of using several different terminologies or a combination of multiple terminologies within a biomedical IR framework integrating several state-of-the-art term weighting models and query expansion models. Unlike previous work which focused only on query expansion using different knowledge sources, we aim to point out that the combination of the document's global context (domain knowledge sources or termino-ontological resources) and the query's local context (the elite set of top-ranked documents of the query) may be a source evidence to improve the biomedical IR effectiveness. The difference between the global context and local context is that the former is independent from the query while the latter is dependent on the query.

5. A biomedical IR framework based on multiple terminologies

Our general indexing and retrieval framework is made up of two main components detailed below: (1) *Multi-terminology Indexing* and (2) *Document Retrieval* (*cf.* Fig. 1). We integrate them into a conceptual IR process as the combination of the global and local semantic contexts for improving the biomedical IR effectiveness. The **global context** is referred to as domain knowledge sources (multiple biomedical terminologies) and the **local context** is referred to as the top-ranked documents returned from the previous retrieval stage.

During the indexing stage, documents are first **expanded** with all words as part of the preferred terms denoting concepts that are extracted using termino-ontological resources. For concept extraction, we adopt MaxMatcher [43], which is an approximate dictionary-based lookup for concept extraction [43]. Given a document, MaxMatcher splits the document into sentences. For each sentence, a set of terms or phrases denoting concepts as well as their corresponding concept unique identifiers are extracted for this document. However, MaxMatcher does not measure the importance of each concept for describing the semantics of the document. To achieve this, we use the BM25 term weighting model [53] to measure the degree of description of a concept c_j in terminology T_i:

$$w_{ji}^{D} = \max_{v_p \in c_j} \sum_{k=1}^{\ell} tf(t_k) * \frac{log(N - n_k + 0.5/n_k + 0.5)}{k_1 * ((1-b) + b * (dl/avg\_dl)) + tf(t_k)} \quad (1)$$

where v_p is the pth term variant, also called as entry term for concept c_j; t_k is the kth word constituent[2] of entry term e_p of length ℓ in terminology T_i; $tf(t_k)$ is the number of occurrences of word t_k in document D; N is the total number of documents in the collection; n_k is the number of documents containing word t_k; dl is the document length, avg_dl is the average document length; k_1, and b are the tuning parameters of the BM25 model.

Let $R(D, T_i)$ be the set of concepts extracted from document D using terminology T_i. Then the list of concepts extracted from D using several terminologies $T = \{T_1, T_2, \ldots, T_n\}$, can be defined as: $R(D, T) = \cup_{i=1}^{n} R(D, T_i)$, where n is the number of terminologies used for concept extraction. For each document, a list of a limited number of candidate concepts are extracted using each terminology separately. Afterwards, they are merged together thanks to their Concept Unique Identifier (CUI) from UMLS.

During the retrieval, the query is splitted into single word terms delimited by white spaces and punctuations for matching document terms. We first submit the query to obtain the top-ranked documents. Afterwards, the best candidate terms extracted from top-ranked expanded documents returned by the first retrieval stage are used to expand the original query. We then detail our IR framework of two components: (1) *Multi-terminology indexing* and (2) *Context Sensitive IR*.

5.1. Multi-terminology document expansion (DE): a voting problem

Suppose we have n terminologies used for indexing biomedical documents. We first extract concepts from each document D using a particular terminology T_i, i.e., we will obtain n lists of concepts for document D. We need to fuse n concept lists to obtain a final list of unique concepts of the document. Inspired by the principle of poly-representation in IR [18] making simultaneous combination of evidences that are cognitively different in order to increase the information value of documents, our concept extraction method is typically based on well known voting-based data fusion techniques (e.g., CombMAX, CombMIN, CombSUM, etc.) that have been used to combine data from different information sources [19]. Our purpose here is to select the best multi-terminological concepts as a fusion of mono-terminological concepts by means of voting scores assigned to candidate concepts. For this purpose, we propose to combine rankings of the extracted concepts for each document using their scores and/or their ranks from the extraction stage. Intuitively speaking, the concept fusion can be seen as the voting problem described as follows. We compute the combined score of the candidate concept c_j voting for document D, given its score w_{ji}^{D} and rank r_{ji}^{D} when using terminology T_i, as the aggregation of votes of all identified concepts. We consider two sources of evidence when aggregating the votes to each candidate concept: (E1) Scores of the identified concept voting for each document; (E2) Ranks of the identified concept voting for each document.

We evaluate 8 voting techniques based on known data fusion methods [19], which aggregate the votes from several rankings of concepts into a single ranking, using both the ranks and/or scores of candidate concepts. The lists of extracted concepts from each document using several terminologies are merged together to obtain a final single concept list representing the document's subject matter(s). An appropriate number of extracted concepts will be chosen in order to expand the document content. Table 1 depicts all the voting techniques that we use and evaluate in this work. They are grouped into two categories according to the source of evidence used. The $\|.\|$ operator indicates the number of concepts having non-zero score in the described set; r_{ji}^{D} is the rank of concept c_j defined in terminology T_i and extracted from document D; and w_{ji}^{D} is the score of concept c_j, defined in T_i and extracted from document D, computed using the probabilistic BM25 scheme [53].

[2] A constituent is a token forming a part of concept names, e.g., 'back', 'pain' are constituents of "back pain".

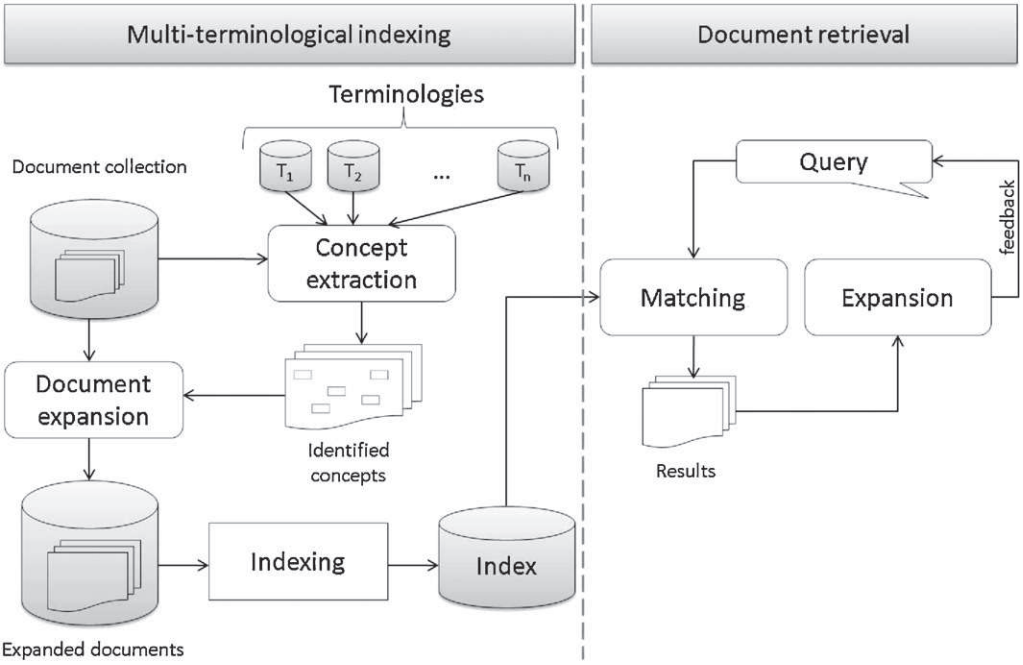

Fig. 1. Multi-terminology indexing and retrieval framework.

5.2. Biomedical document retrieval based on a multi-terminology based indexing approach

The document retrieval component aims at matching the user's query representation to the documents' representation by taking into account the subject matters of both queries and documents in the collection. As stated in Section 5.1, the subject matters of documents can be detected via concepts extracted from documents using one or many terminologies (namely document's global context). For gathering more information of the user query, we use a *local context* query expansion method for expanding the original user query with relevant terms extracted from the top-ranked expanded documents (namely query's local context). The top-ranked expanded documents are obtained by matching the original user query to expanded documents in the collection. We detail in what follows the process of *document–query matching* (first retrieval stage) and *query expansion* (used for the second retrieval stage). The second retrieval stage is similar to the first retrieval stage except that in the second stage, the user query is expanded with more terms.

5.2.1. Document–query matching

The aim of an IR system is to retrieve documents that respond to a particular user query. To reach this goal, each candidate document is assigned a Relevance Status Value (RSV). Afterwards, documents are ranked using a particular term weighting model, which we describe below, in a descending order according to this value. In IR, the fact that each document is assigned by an RSV value is referred to as document weighting. In order to have a general vision of our IR approach, we use three different state-of-the-art weighting scheme: BM25 [53] and two other DFR models namely In_expB2 [54] and LGD [55].

The BM25 weighting schema: in this model, the RSV of a document D for a given query Q is computed as follows:

$$\mathrm{RSV}(D,Q)=\sum_{t\in Q}\frac{(k_1+1)*tfn}{K+tfn}*\frac{(k_3+1)*qtf}{k_3+qtf}*w^{(1)} \tag{2}$$

where tfn is the normalized within-document term frequency given by:

$$tfn=\frac{tf}{(1+b)+b*(dl/avg\_dl)}, \tag{3}$$

where tf is the within-document term frequency, dl and avg_dl are respectively the document length and average document length,

- k_1, k_3 and b are tuning parameters,
- K is $k_1*((1-b)+b*dl/avg\_dl)$,

Table 1
Voting techniques used for a multi-terminology based concept extraction.

Category	Technique	w_{ji}^D=score(c_j, T_i, D)	Description
Rank-based	CombRank	$\sum_{i=1}^{n}(\|R(D,T_i)\|-r_{ji}^D)$	Sum of concept ranks
	CombRCP	$\sum_{i=1}1/r_{ji}^D$	Sum of inverse concept ranks
Score-based	CombSUM	$\sum_{i=1}^{n}w_{ji}^D$	Sum of concept scores
	CombMIN	$\min\{w_{ji}^D, i=\overline{1\ldots n}\}$	Minimum concept sores
	CombMAX	$\max\{w_{ji}^D, i=\overline{1\ldots n}\}$	Maximum concept scores
	CombMED	$\mathrm{median}\{w_{ji}^D, i=\overline{1\ldots n}\}$	Median of concept scores
	CombANZ	$\sum_{i=1}^{n}w_{ji}^D \div \|\{c_j\in R(D,T_i)\}\|$	$CombSUM \div \|\{c_j\in R(D,T_i)\}\|$
	CombMNZ	$\sum_{i=1}w_{ji}^D \times \|\{c_j\in R(D,T_i)\}\|$	$CombSUM \times \|\{c_j\in R(D,T_i)\}\|$

- *qtf* is the within-query term frequency,
- $w^{(1)}$ is the *idf* (inverse document frequency) factor computed as:

$$w^{(1)} = log_2 \frac{N - N_t + 0.5}{N_t + 0.5} \quad (4)$$

where N is the total number of documents (or cardinality) in the collection, and N_t is the number of documents containing term t (also called as document frequency).

The In_expB2 weighting schema: in this model, query terms are weighted using the Inverse Expected Document Frequency model with Bernoulli after-effect and term frequency normalisation [54]. Formally, the RSV of a document D for a query Q is:

$$\text{RSV}(D, Q) = \sum_{t \in Q} \frac{qtf \times (tf + 1) \times tfn_2}{N_t \times (tfn_2 + 1) \times ln2} \times log_2 \frac{N + 1}{N \times (1 - e^{-tf/N}) + 0.5} \quad (5)$$

where t is a query term occurring in document D, N_t is the document frequency, N is the total number of documents in the collection, qtf is the query term frequency, tf is the within-document term frequency, tfn_2 is the normalised within-document term frequency, given by:

$$tfn_2 = \frac{tf}{ln2} \times log_2 \left[1 + c \times \frac{avg\_dl}{dl}\right] \quad (6)$$

The LGD weighting model: in this model, query terms are weighted using the log logistic distribution [55]. Formally, the RSV of a document D for a query Q is:

$$\text{RSV}(D, Q) = \sum_{t \in Q} qtf \times \left[log_2 \left(\frac{N_t}{N} + tf_n\right) - log_2 \left(\frac{N_t}{N}\right)\right] \quad (7)$$

where t is a query term occurring in document D; N_t is the document frequency (i.e., number of documents containing term t); N is the total number of documents in the collection; qtf is the query term frequency; tf_n is the normalised within-document term frequency, given by:

$$tf_n = tf \times log_2 \left(1 + c \times \frac{avg\_dl}{dl}\right) \quad (8)$$

where avg_dl is the average document length (in tokens), dl is the document length (in tokens) and c is a multiplying factor or tuning parameter.

5.2.2. Query expansion

Query expansion (QE) aims at enriching the original user's query in order to retrieve more relevant documents or modify the document rankings so to optimize the IR performance. Two main QE approaches can be distinguished: **global context** *vs.* **local context** analysis approach.

The global context QE tries to extract concepts from multiple terminologies (global context) and automatically expand to the original user query with the most appropriate terms denoting concepts. The local context QE applies a blind feedback technique to select the best terms from the top-ranked expanded documents in the first retrieval stage. In this expansion process, terms in the top-returned documents are weighted using a particular Divergence From Randomness (DFR) term weighting model [54]. In our work, two statistical measures namely Bose–Einstein (Bo) and Kullback–Leibler (KL) statistics [54] are used to weight terms in the expanded query Q^e derived from the original query Q. The DFR framework employs a QE mechanism that is a generalization of Rocchio's method [21]: terms in the top-ranked documents retrieved in the first stage are weighted using a particular DFR term weighting model. In general, the weight of a term of the expanded query Q^* derived from the original query Q is obtained as follows:

$$weight(t \in Q^*) = qtfn + \beta * \frac{Info_{DFR}}{MaxInfo} \quad (9)$$

$qtfn$ is the normalised within-query term frequency, β is a decay factor for which the default value is 0.4; $Info_{DFR}$ is the term frequency in the expanded query induced by using a DFR model:

$$Info_{DFR} = -log_2 Prob(Freq(w|K)|Freq(w|C)) \quad (10)$$

where *Prob* is the probability of obtaining a given within-query term frequency from the top-ranked documents retrieved in the first stage; and $Freq(w|K)$ (resp. $Freq(w|C)$) is the term frequency within the top-ranked documents (resp. the collection).

- $MaxInfo = arg_{t \in Q^*} \max Info_{DFR}$ is the value that maximizes $Info_{DFR}$ [56].

In the DFR framework, several measures are used to compute this probability such as: Bose–Einstein (Bo) and Kullback–Leibler (KL) measures [54]. Each of the measures is described below.

The Bose–Einstein measure gives the following term frequency normalisation computed by a *geometric distribution* with probability $p = 1/(1 + \lambda)$, that is:

$$Info_{Bo} = -log_2 \left(\frac{1}{1 + \lambda_{Bo}}\right) - Freq(w|K) * log_2 \left(\frac{\lambda_{Bo}}{1 + \lambda_{Bo}}\right) \quad (11)$$

where

$$\lambda_{Bo1} = \frac{Freq(w|C)}{N} \quad and \quad \lambda_{Bo2} = \frac{TotalFreq(K) * Freq(w|C)}{TotalFreq(C)} \quad (12)$$

The Kullback–Leibler measure gives the following term frequency normalisation, which is non-symmetric measure of the difference between two probability distributions, i.e., the set of topmost documents satisfying the query (also known as the elite set of the query) and the whole collection, computed as follows:

$$Info_{KL} = \frac{Freq(w|K)}{TotalFreq(K)} * log_2 \frac{Freq(w|K) * TotalFreq(C)}{Freq(w|C) * TotalFreq(K)} \quad (13)$$

6. Experimental evaluation

6.1. Evaluation objectives

Within the context of biomedical IR, the TREC Genomics Track aims at creating test collections for evaluating biomedical IR systems and related tasks in the Genomics domain. The Genomics Track differs from other TREC tracks in that it is focused on retrieval in a specific domain as opposed to general retrieval tasks, such as Web searching or question answering. The TREC Genomics collections have changed over years according to the task defined by TREC organizers. For example, TREC Genomics 2004 and 2005 focused on *ad hoc* document retrieval and text categorization. A copy of a small subset of the MEDLINE database has been used for creating the test collections. Each document in the collection consists of a title and/or an abstract which constitute the main content of the document. Since 2006, the focus of TREC Genomics Track was the development of test collections for evaluating IR methods that return passages (from part to sentence or paragraphs in length). Therefore, the *ad hoc* document retrieval task is only supported in TREC Genomics 2004 and 2005.

In this large context, we conducted several series of experiments to achieve the following objectives:

Table 2
Examples of TREC Genomics queries.

<ID>2</ID>
<TITLE>Generating transgenic mice</TITLE>
<NEED>Find protocols for generating transgenic mice.</NEED>
<ID>15</ID>
<TITLE>ATPase and apoptosis</TITLE>
<NEED>Find information on role of ATPases in apoptosis</NEED>

- Evaluate the IR effectiveness of the combination of document and query expansion for improving biomedical IR effectiveness.
- Study the impact of using domain knowledge sources (terminologies) for biomedical IR, especially for Genomics *ad hoc* retrieval. We compare our terminology based IR approach to state-of-the-art IR models.

6.2. Data sets

By focusing on an *ad hoc* IR task at the level of **document** (which is mainly constituted of *title* and/or *abstract*), we validate our multi-terminology based IR approach using two collections: TREC Genomics 2004 and 2005, which are the subset of about 4.6 millions MEDLINE citations from 1994 to 2003, under the IRToolkit platform.[3] A MEDLINE document contains six fields: title (.T), abstract (.W), MeSH indexing terms (.M), author (.A), source (.S), and publication type (.P). The MeSH field (.M) corresponds to MeSH main headings/subheadings assigned by human experts to each document. We only used title and abstract for indexing documents. According to the conventional pooling method of TREC, human relevance judgments were merely made to a relative small pool, which were built from the top-precedence run from each of the participants. Since extracting concepts for each document in the huge TREC Genomics collection is quite expensive (e.g., human MeSH-based indexing has been done through many years), it is practically impossible to train and test a series of experiments that integrate document expansion and query expansion. Therefore, like authors in [48], our prototype IR system only indexes and searches all human relevance judged documents, i.e. the union of 50 single pools that contains total 48,753 citations in TREC 2004 and 41,018 ones in TREC 2005, without using manually assigned MeSH tags.

There are 50 queries in the TREC Genomics 2004 and 49 queries in TREC Genomics 2005 (see Table 2 for some examples of TREC Genomics queries). The topics for the *ad hoc* retrieval task were developed from the information needs of real biologists and modified as little as possible to create needs statements with a reasonable estimated amount of relevant articles [57,58]. The relevance judges were instructed according to several criteria, e.g., the relevant article must describe how to conduct, adjust, or improve a standard, a new method, or a protocol for doing some sort of experiment or procedure.

6.3. Evaluation protocol

In order to achieve the objectives mentioned above, we designed an evaluation protocol with three series of experiments:

- the first one concerns the classical indexing and retrieval of MEDLINE citations without using neither domain knowledge sources nor MeSH terms which are manually assigned by human indexers: at the indexing stage, stop-words are removed from documents and queries; terms are stemmed before indexing and searching. At the retrieval stage, documents are ranked by using the three term weighting models described in Section 5.2.1;
- the second series of experiments concerns a mono-terminology indexing and retrieval based on four different terminologies: MeSH, SNOMED, ICD-10 and GO. Here, we used a modified version of MaxMatcher[4] (MM) [43], which is basically a term/concept extractor from biomedical documents. In order to quantify the extracted concepts from each document, we extract preferred terms denoting concepts which are weighted by the BM25 model. Finally, extracted terms are used for document expansion (DE). Since our previous experiments presented in [59] have demonstrated the utility of combining document and query expansion using a limited number of terms from the top-ranked expanded documents returned from the first retrieval stage, we then applied this technique on all of our terminology-based IR runs;
- the third series of experiments concerns a multi-terminology based indexing and retrieval. Similar to the second experiment, we applied DE and QE but with the exception that the extracted concepts from each document are combined together using several voting techniques to obtain a final single list of unique concepts for document expansion. Finally, like the previous scenario, we combine QE and DE.

Since our terminology-based IR approaches are based on QE and DE, we need to learn a ranking function incorporating three main parameters: (1) the number of terms extracted from (2) top k documents for QE and (3) the number of extracted terms denoting domain concepts for DE. We train the ranking function on TREC Genomics 2004 and apply the best configurations on TREC Genomics 2005 for testing.

For measuring the IR effectiveness, we used the *MAP* metrics representing the *Mean Average Precision* calculated over all queries. The average precision of a query is computed by averaging the precision values computed for each relevant retrieved document of rank $x \in (1 \ldots 1000)$. Our MAP results are generated by the *trec_eval* tool,[5] which has been widely used by the TREC community for evaluating *ad hoc* retrieval runs.

7. Experimental results

The results presented in what follows are related to the three series of experiments described earlier. Firstly, we compare the performance of query expansion models for IR to the classical IR models. We then investigate the IR effectiveness of our terminology-based IR using **document expansion** (DE) based on four different terminologies and **query expansion** (QE) based on the collection statistics related to the original query. We further demonstrate the utility of using voting techniques for combining concepts issued from several terminologies to provide a coherent list of concepts representing the semantics of the document.

7.1. Comparing baseline IR models to QE models

The objective of this experiment is to show the stability of the strong baseline model that we use for evaluating our terminology-based IR approach. Fig. 2 shows the MAP results obtained on the TREC Genomics 2004 collection that we use for training QE parameters, i.e., the number of terms extracted (*# terms*) from the top-ranked documents (*# docs*). We can see that among three weighting models, the *In_expB2* model (without QE) performs slightly better than two other weighting models. The MAP value

[3] Information Retrieval Toolkit: http://sourceforge.net/projects/irtoolkit/files/ (accessed 11.07.12).

[4] Downloadable at: http://sourceforge.net/projects/cxtractor/files/ (accessed 11.07.12).

[5] http://trec.nist.gov/trec_eval/ (accessed 11.07.12).

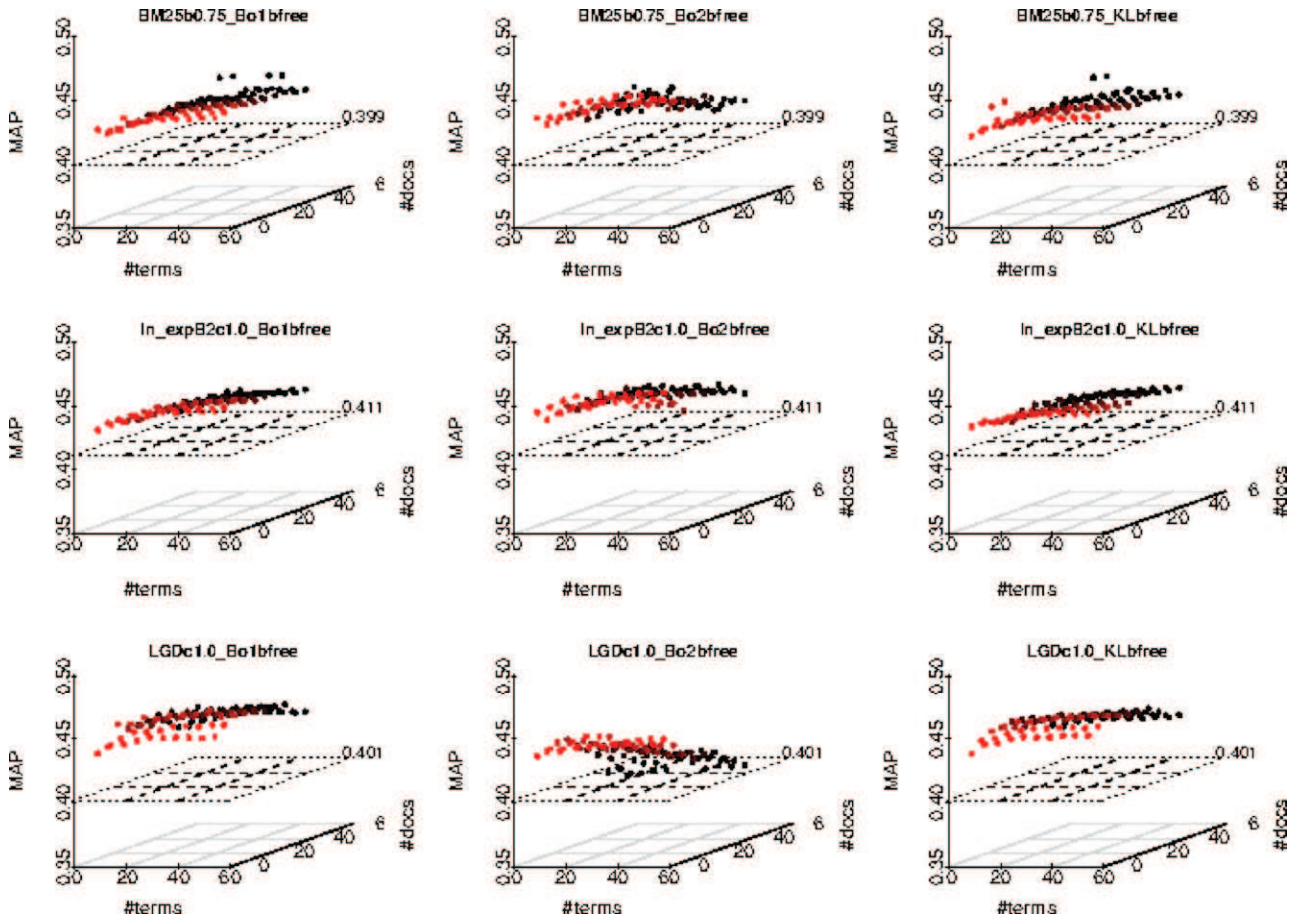

Fig. 2. MAP results obtained by different QE models among the Bo1, Bo2 and KL models in combination to one of the underlying term weighting models among the BM25, LGD and In_expB2 on the TREC Genomics 2004 collection. Dotted planes correspond to the performance of the underlying term weighting model.

obtained by the *In _expB*2 model is **0.4117** compared to 0.3997 and 0.4018 for the *BM*25 and *LGD* models respectively. When applying query expansion on the query set, almost all of the QE models outperform each of the term weighting models. For example, the combination of the *LGD* term weighting model and the Bo_1 QE model (namely LGD_Bo_1) gives the best MAP of 0.4637 by using **40** terms extracted from **15** top-ranked documents, which yields an improvement up to +15.41% compared to the *LGD* model. The combination of *LGD* model with the Bo_2 QE model (namely LGD_Bo_2) results in a MAP of 0.4464 (# *terms* = 20, # *docs* = 10) in the best case and a MAP of 0.3937 (# *terms* = 5, # *docs* = 50) in the worst case. This is probably because the top 5 extracted terms from the top 50 documents are not relevant to the query causing the document–query term mismatch.

Fig. 3 depicts the distribution of the kernel density, which is a non-parametric method to estimate the probability density function of a random variable [60], related to MAP values obtained by each retrieval model combined with all QE models and vice-versa. We can see that the LGD model (with an average MAP of 0.4411) performs better than the two other term weighting models. Concerning QE models, although the three QE models have a very competitive performance in terms of MAP, the average MAP of the Bo1 QE model (0.4389) has a strongest confidence than the two other models.

Consequently, all subsequent runs reported in the next experiments use the **LGD model** for *term scoring* and the **Bo1 model** for *query expansion* using **40** terms extracted from the top **15** documents (our *strong baseline*). In the next experiments, we aim to study the impact of our mono *vs.* multi terminology based IR approaches on the TREC Genomics 2005 collection using this configuration.

7.2. *Effectiveness of the mono-terminology indexing and retrieval*

In this subsection, we consider expanding original documents (titles and/or abstracts) with preferred terms denoting concepts extracted from each document using a mono-terminology (MESH, SNOMED, ICD-10 and GO). Our DE method is inspired by the work of human indexers providing a dozen of MeSH terms in MEDLINE citations [5]. In this experiment, we extract a limited number of terms denoting concepts (from 5 to 50) in each terminology separately to expand the document contents using MaxMatcher++, implemented in the **cxtractor** software.[6] The difference between the original version of MaxMatcher and our modified version concerns the term scoring function that we introduced to take into account the importance of the extracted concepts.

Table 3 depicts the MAP results of our runs based on the mono-terminology document expansion for biomedical IR on the TREC Genomics 2005 test set as well as the results obtained by the selected baseline run (described in Section 7.1). As we can see, most of the MAP results obtained by our mono-terminology IR approach outperform the baseline with an improvement rate from +0.90 % to +5.26 %. We also notice that expanding the document content with a number of 15 concepts yields the best results for the

[6] http://sourceforge.net/projects/cxtractor/files/ (accessed 11.07.12).

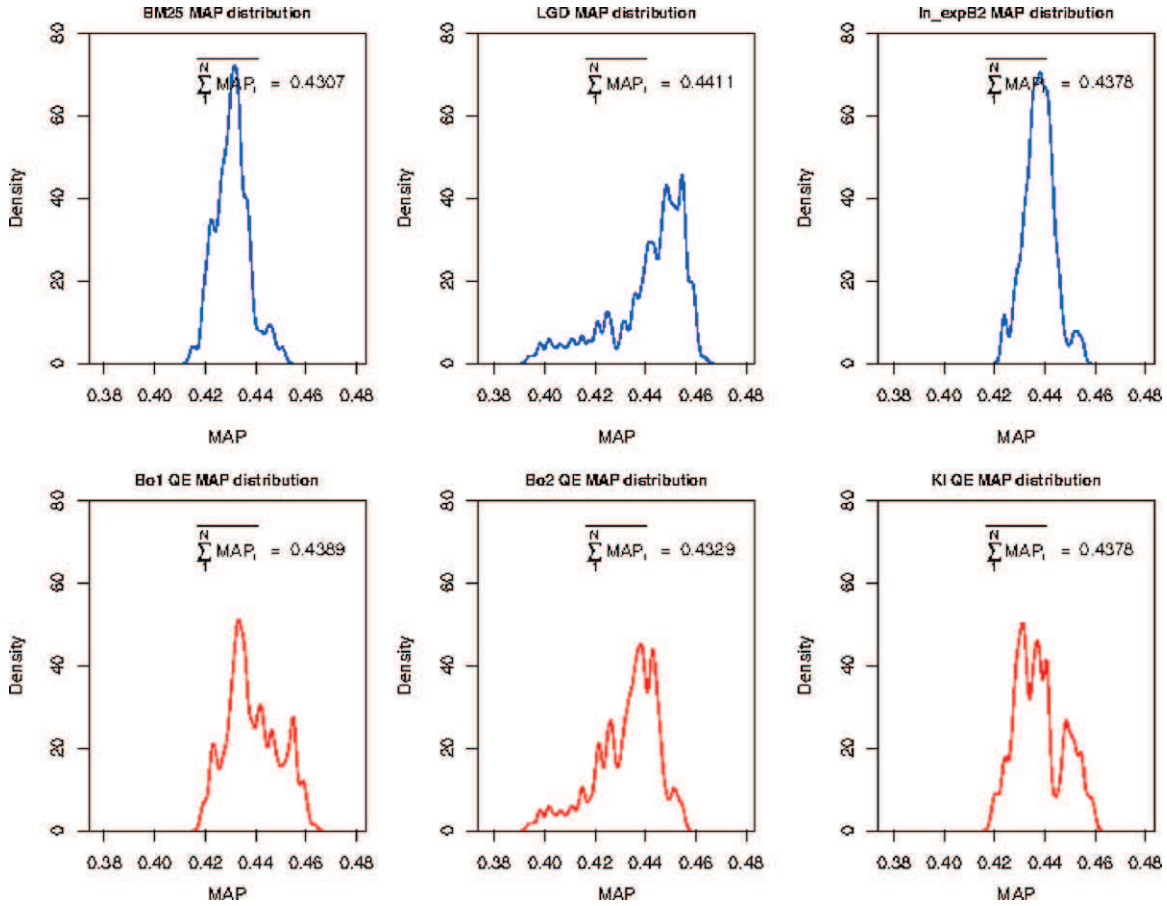

Fig. 3. Kernel density distribution of $N = 300$ individuals ($10 \times 10 \times 3$ MAP values) obtained by each weighting model combined with three QE models and vice-versa.

mono-terminology IR. This number is equivalent to the number of MeSH indexing terms selected by human indexers. In general, MeSH terms are relevant for indexing documents because they allow to optimize the MAP results over all terminologies for most of the numbers of extracted concepts. However, the choice of terminology used for concept extraction and document expansion for IR among the first three terminologies (MeSH, SNOMED and ICD-10) is not important since the MAP values obtained by using each terminology are not significantly different. For GO, we see that document expansion using top 5 concepts yields an improvement rate of +1.99 % in terms of MAP. However, the performance tends to decrease when the number of expanded concepts is greater than or equal to 10. This could be probably due to the fact that the concept extractor could not find best GO terms for each document or the extracted terms are useless for representing the subject matters of the document. We conclude that document expansion using a dozen of terms denoting concepts issued from domain knowledge sources could improve the retrieval performance. However, the expanded terms must represent well the subject matters of the document, which help to find best results in response to the user query. Although GO terms are not relevant for indexing documents, we aim to integrate them into our multi-terminology IR approach in order to study the influence of using concepts predefined in several terminologies on the biomedical IR performance.

Table 3
Effectiveness of our mono-terminology based IR on TREC Genomics 2005 collection.

N	Terminology			
	MeSH	SNOMED	ICD-10	GO
Baseline		0.2664		
5	0.2724 (+2.25)	0.2697 (+1.24)	0.2709 (+1.69)	**0.2717** (+1.99)
10	0.2763 (+3.72)†	0.2688 (+0.90)	0.2759† (+3.57)	0.2622 (−1.58)
15	**0.2791** (+4.77)	**0.2736** (+2.70)	**0.2804**† (+5.26)	0.2623 (−1.54)
20	0.2778 (+4.28)	0.2732 (+2.55)	0.2776† (+4.20)	0.2623 (−1.54)
25	0.2759 (+3.57)	0.2733 (+2.59)	0.2748 (+3.15)	0.2623 (−1.54)
30	0.2758 (+3.53)	0.2731 (+2.52)	0.2736 (+2.70)	0.2623 (−1.54)
35	0.2756 (+3.45)	0.2731 (+2.52)	0.2734 (+2.63)	0.2623 (−1.54)
40	0.2752 (+3.30)	0.2731 (+2.52)	0.2733 (+2.59)	0.2623 (−1.54)
45	0.2753 (+3.34)	0.2731 (+2.52)	0.2733 (+2.59)	0.2623 (−1.54)
50	0.2754 (+3.38)	0.2731 (+2.52)	0.2731 (+2.52)	0.2623 (−1.54)

N is the number of extracted terms denoting concepts used for document expansion; the bold (resp. underlined) numbers indicate the optimal MAP value within particular terminology (resp. the optimal MAP over all terminologies at *N* concepts); the numbers in parentheses indicate the improvement compared to the baseline.

† Statistically significant improvements represented by the paired-sample *T*-tests (computed with R [61]) at $p \leq 0.05$.

Table 4
Effectiveness of our multi-terminology based IR approach on TREC Genomics 2005.

Voting model	#docs				
	10	12	15	18	20
Baseline			0.2664		
CombANZ	0.2814	0.2790	0.2744	0.2684	0.2623
	(+5.63)	(+4.73)	(+3.00)	(+0.75)	(−1.54)
CombMAX	**0.2831**	0.2849	0.2750	0.2695	0.2648
	(+6.27)	(+6.94)	(+3.23)	(+1.16)	(−0.60)
CombMED	0.2805	0.2846	0.2782	0.2699	0.2652
	(+5.29)	(+6.83)	(+4.43)	(+1.31)	(−0.45)
CombMIN	0.2812	0.2823	0.2776†	0.2678	0.2651
	(+5.56)	(+5.97)	(+4.20)	(+0.53)	(−0.49)
CombMNZ	0.2812	0.2849	0.2777	0.2675	0.2650
	(+5.56)	(+6.94)	(+4.24)	(+0.41)	(−0.53)
CombRank	0.2800	**0.2859**	0.2776	0.2670	0.2637
	(+5.11)	(+7.32)	(+4.20)	(+0.23)	(−1.01)
CombRCP	0.2810	0.2848†	0.2779	0.2700	0.2648
	(+5.48)	(+6.91)	(+4.32)	(+1.35)	(−0.60)
CombSUM	0.2806	0.2850†	**0.2785**	**0.2704**	**0.2666**
	(+5.33)	(+6.98)	4.54)	(+1.50)	(+0.075)

The bold numbers indicate the MAP optimal values obtained by the best voting model using 40 terms extracted from the top #docs expanded documents; the underlined numbers are the optimal MAP value obtained by each voting model; the numbers in parentheses indicate the improvement rates in terms of MAP compared to the results obtained by the strong baseline LGD_Bo1 model.

† Statistically significant improvements at $p \leq 0.05$.

7.3. Effectiveness of our multi-terminology based indexing and retrieval

In this subsection, we evaluate the effect of using multiple terminologies for biomedical IR. We discuss the results obtained by our multi-terminology based indexing and retrieval by comparing to the baseline performance as well as the mono-terminology approach. Here, biomedical concepts predefined in four terminologies (MeSH, SNOMED, ICD-10 and GO) are extracted from the document content using a modified version of MaxMatcher, for which we introduce the concept scoring using the BM25 weighting schema (*cf.* formula 1). The extracted concepts obtained in several rankings of concepts are combined together into a single pool of unique concepts for each document using eight voting techniques (described in Section 5.1). All words as part of the preferred terms issued from the top 15 concepts are retained for document expansion. As shown in the previous experiments, query expansion using 40 terms from the top-ranked documents helps to improve the IR effectiveness. In order to assess the quality of the top 40 extracted terms from the top ranked documents for our multi-terminology IR approach, we calculate the MAP results for each voting model used for combining the extracted concepts from each document using multiple terminologies. The number of top documents involved in the selection of terms for query expansion is among the following values {10, 12, 15, 18, 20}.

Table 4 shows the retrieval performance in terms of MAP on the TREC Genomics 2005 selected test set obtained by using 8 voting techniques for combining concepts issued from multiple terminologies. First, we compare the MAP results of each voting model to the strong baseline, i.e., the LGD_Bo_1 model). Afterwards, we discuss the added value of our voting models compared to the mono-terminology based scenario.

As we can see in Table 4, most of the voting models outperform the baseline when terms are extracted from no more than 20 top-ranked expanded documents. Globally, the *CombSUM* technique, which selects candidate concepts by the sum of their scores, gives the best performance compared to other voting models. However, the overall best results are only obtained by the *CombRank* method at $\#docs = 12$. This proves that concepts with high scores or ranks voted by each terminology in the list of extracted concepts are important to describe the semantic content or the subject matters of the document. Indeed, such concepts are assigned a TF-IDF weight (i.e., the BM25 score in our concept fusion algorithm), which is a good way to model the relevance of candidate concepts.

In order to show how our multi-terminology IR approach is statistically significant, we computed the paired-sample *T*-tests (using R [61]) between the set of MAP values obtained based on each voting model to the strong baseline. For example, the paired-sample *T*-tests for CombSUM ($p = 0.0303$, $df = 49$, $t = 2.2311$, $M = 0.0115$), CombMIN ($p = 0.0277$, $df = 49$, $t = 2.2697$, $M = 0.0112$) and CombRCP ($p = 0.0144$, $df = 49$, $t = 2.5369$, $M = 0.0026$) show that our multi-terminology IR approach based on voting techniques are statistically significant compared to the baseline.

When comparing the IR performance of the multi-terminology to the mono-terminology IR method, we observe that the MAP results obtained by both of them are slightly different (see Table 5). We further look at the precision values at top 10 and 20 returned documents and the recall obtained by each *single terminology* and the *CombSUM* method, which is the best voting model of concept fusion for biomedical IR. As shown in Table 5, the CombSUM method gives best results in terms of MAP at 12 top-ranked expanded documents used for QE and best performance in terms of P@10 and P@20 at 15 top documents used for QE. This proves that our multi-terminology approach combining concepts from several terminologies provides a more stable performance than using a mono-terminology. Indeed, in general, the mono-terminology based IR approach yields a significant improvement rate from +2.70% to +5.26 %; but in the worst case, the IR performance can be decreased because of irrelevant terms added to the document content. With our multi-terminology IR approach based on voting models, we are able to select more relevant concepts which are the results obtained by several votes. Such mechanism allows to extract relevant terms by removing irrelevant terms from the final

Table 5
Comparison of precision/recall between mono- *vs.* multi-terminology IR.

	MAP	P@10	P@20	Recall
MeSH	0.2791	0.3980	0.3561	**4107/4584**
SNOMED	0.2736	0.4102	0.3439	4053/4584
ICD10	0.2804	0.3959	0.3571	4110/4584
GO	0.2623	0.3837	0.3398	4036/4584
CombSUM (#doc 15)	0.2785	**0.4122**	**0.3633**	4082/4584
CombSUM (#doc 12)	**0.2850**	0.4061	0.3551	4062/4584

extracted concept list. The results obtained by either our mono- and multi-terminology IR methods also show that document expansion with no more than 20 terms denoting concepts in combination with query expansion with about 40 terms selected from no more than 20 top ranked expanded documents provides a potential solution to improve the biomedical IR effectiveness.

8. Discussion

Our research work is mainly related to the exploitation of conceptual information in domain knowledge sources as well as statistical characteristics of the collection in order to provide suitable rankings of documents that are potentially relevant to a given user query. Through the experiments reported in this paper, it was hoped to determine the relevant factors affecting the effectiveness of biomedical IR that are: (1) term weighting using a particular model, (2) QE using an appropriate number of terms extracted from a relevant elite set and (3) DE using a dozen of terms denoting domain concepts issued from either a mono-terminology or multiple terminologies.

We claim that the combination of an appropriate *term weighting model* with a relevant *query expansion* model could help to improve the biomedical IR performance. We have trained the combination of term weighting and query expansion models to learn a ranking function that takes into account two main parameters, i.e., the number of extracted terms and the elite set size. We obtained the best performance of query expansion which constitutes our **strong baseline** for evaluating our **terminology-based IR approach**. We further introduced the third parameter, which is the number of terms denoting concepts extracted from each document. We integrated concepts that are extracted from the document content using either a mono-terminology or multiple terminologies into the biomedical IR process via DE. Our intuition for DE was typically based on the assumption that relevant concepts extracted from the document content are able to capture the semantics of documents so to bring the user query closer to relevant documents in the collection. The results obtained by our mono-terminology IR approach demonstrated that DE in combination with QE is useful for improving the performance in terms of MAP in comparison to the results obtained only by QE.

In addition, we have evaluated several voting models for combining concepts issued from multiple domain knowledge sources. According to the results trained and tested on the TREC Genomics collections, on one hand, our mono-terminology IR approach (which is based on either DE and QE) consistently outperforms the baseline approach (which is solely based on QE). On the other hand, when introducing voting techniques for selecting best concepts extracted from each document using multiple terminologies, the results of our multi-terminology based approach demonstrate the utility to take into account the scores and ranks of candidate concepts extracted from a document. Indeed, it has been shown in Section 7.3 that our multi-terminology IR approach consistently and significantly outperforms the baseline and provides a stable performance compared to the mono-terminology IR approach either in terms of MAP and precision at top 10 and 20 returned documents. We could explain the difference in terms of performance of our multi-terminology as follows: since the distribution of concepts in each terminology is different from one to another, the extracted concepts can have different ranks in each list of candidate concepts and a particular concept can be present in one but not in another as well. On the other hand, if a particular concept appears in N terminologies, its total score will be the sum of all component scores, which is also the product of the BM25 score in the document and the number of terminologies where it appears. So, the combination or voting has an impact on different concepts in several terminologies rather on the same concept in our voting methods for concept extraction.

According to the results in Table 4, since the CombSUM and CombRank voting techniques allow to optimise the MAP results, we conclude that "concepts with high scores or ranks voted by each terminology in the list of extracted concepts are important to describe the semantic content or the subject matters of the document". We also noticed that there is no significant difference in terms of MAP between voting techniques for multi-terminology concept extraction. We should investigate further on the frequency of the words in each terminology, i.e., the number of concepts containing the word. If it is more frequent in the terminology, it is likely that this word is less important than those that appear in a smaller number of concepts.

It is also interesting to notice that many research works in the biomedical domain have repeatedly demonstrated the added value of using MeSH terms which are manually or semi-automatically expanded to the MEDLINE citations [1,2,62,59]. In a completely automatic setting, we demonstrate that automatic concept extraction whether based on a mono-terminology or multiple terminologies could be an effective way to improve the IR performance in comparison to the state-of-the-art IR models.

9. Conclusion

We have presented in this research a novel IR approach to combining the global context DE and the local context QE in order to improve the biomedical IR effectiveness. The global context DE is typically based on the use of voting techniques for combining concepts issued from several terminologies. The local context QE is based on statistical properties of a sub-collection (top-ranked documents returned from the first retrieval stage). The results demonstrate that our multi-terminology based IR approach provides a significant improvement over a state-of-the-art IR baseline approach. We argued that concept extraction using multiple terminologies can be regarded as a voting problem taking into account the score of identified concepts. The extracted concepts are used for DE and QE in an attempt to close the semantic gap between the user's query and documents in the collection. The results demonstrate that our multi-terminology IR approach shows a significant improvement over the baseline and is more stable than the mono-terminology based IR approach by maintaining a significant improvement in terms of MAP and furthermore it allows to improve the search precision at top 10 and 20 returned documents.

Our future work aims at incorporating our multi-terminology IR into a semantic model taking into account several concept features such as the concept centrality and specificity, which we believe to be able to overcome the limits of the bag-of-words based models. In addition, we also plan to combine several state-of-the-art concept extraction methods by leveraging their advantages. We believe that concepts extracted from several methods could be able to enhance the concept extraction accuracy, so to improve the conceptual indexing and retrieval performance. Another point could be interesting concerns the application of concept extraction on the query side by incorporating the extracted concepts from the original query into the ranking function which computes the "Relevant Status Value" between documents and the query. To do this, we must be sure that the extracted concepts fit "perfectly" the semantics of the query and do not cause the query drift problem.

Acknowledgements

We would like to thank people at the IRIT laboratory who develop and maintain the OSIRIM platform, which is an infrastructure of several interconnected computers for undertaking experimental research in Information Retrieval.

References

[1] Srinivasan P. Query expansion and MEDLINE. Information Processing & Management 1996 July;32:431–43.
[2] Aronson AR, Mork JG, Gay C, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Medinfo 2004:268–72.
[3] Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 2006;22(6):658–64.
[4] Lu Z, Kim W, Wilbur WJ. Evaluation of query expansion using MeSH in PubMed. Information Retrieval 2009;12(1):69–80.
[5] Névéol A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR. A recent advance in the automatic indexing of the biomedical literature. Journal of Biomedical Informatics 2009;42(5):814–23.
[6] Stokes N, Li Y, Cavedon L, Zobel J. Exploring criteria for successful query expansion in the genomic domain. Information Retrieval 2009;12(1):17–50.
[7] Custis T, Al-Kofahi K. A new approach for evaluating query expansion: query-document term mismatch. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '07. 2007. p. 575–82.
[8] Zhou W, Yu C, Smalheiser N, Torvik V, Hong J. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: SIGIR'07. 2007. p. 655–62.
[9] Yu S, Tranchevent L-C, Moor BD, Moreau Y. Gene prioritization and clustering by multi-view text mining. BMC Bioinformatics 2010;11:28.
[10] Pereira S, Neveol A, Kerdelhué G, Serrot E, Joubert M, Darmoni SJ. Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue; 2008. pp. 586–590.
[11] Darmoni SJ, Pereira S, Sakji S, Merabti T, Prieur É, Joubert M, et al. Multiple terminologies in a health portal: automatic indexing and information retrieval. In: Proceedings of Artificial Intelligence in MEdicine, AIME'09. 2009. p. 255–9.
[12] Krauthammer M, Nenadic G. Term identification in the biomedical literature. Journal of Biomedical Informatics 2004;37:512–28.
[13] Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JHM. Understanding terminological systems I: terminology and typology. Methods of Information in Medicine 2000:16–21.
[14] Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making 2008;8(Suppl 1).
[15] Nyström M, Vikström A, Nilsson GH, Ahlfeldt H, Orman H. Enriching a primary health care version of ICD-10 using SNOMED CT mapping. Journal of Biomedical Semantics 2010:7–28.
[16] Taboada M, Lalín R, Martínez D. An automated approach to mapping external terminologies to the UMLS. IEEE Transactions of Biomedical Engineering 2009:605–18.
[17] Avillach P, Joubert M, Fieschi M. A model for indexing medical documents combining statistical and symbolic knowledge. In: Proceedings of the AMIA symposium. 2007. p. 31–5.
[18] Ingwersen P. Cognitive perspectives of information retrieval interaction-elements of cognitive theory. Journal of Documentation 1996;52:3–50.
[19] Fox EA, Shaw JA. Combination of multiple searches. In: Proceedings of Text REtrieval Conference (TREC). 1994. p. 243–52.
[20] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. 1st ed. Addison Wesley; 1999.
[21] J. Rocchio, Relevance feedback in information retrieval; 1971. pp. 313-323.
[22] Mauldin CJ. Beyond the keyword barrier: knowledge-based information retrieval. Information Services and Use 1987;7(4–5):103–17.
[23] Krovetz R, Croft WB. Lexical ambiguity and information retrieval. ACM Transactions on Informations and System 1992 Apr;10:115–41.
[24] Wei CP, Hu P, Tai CH, Huang CN, Yang CS. Managing word mismatch problems in information retrieval: a topic-based query expansion approach. Journal of Management and Information Systems 2007 December;24:269–95.
[25] Krovetz R. Homonomy and polysemy in information retrieval. In: Proceedings of the COLING/ACL'97 conference. 1997.
[26] Navigli R. Word sense disambiguation: a survey. ACM Computing Surveys 2009;41:10–79.
[27] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science 1990;41:391–407.
[28] Kazama J, Makino T, Ohta Y. Tuning support vector machines for biomedical named entity recognition. In: Workshop on natural language processing in the biomedical domain. 2002. p. 1–8.
[29] Guarino C, Masolo N, Vetere G. OntoSeek: using large linguistic ontologies for accessing on-line yellow pages and product catalogs. In: National Research Council, LADSEBCNR: Padavo. 1999.
[30] Khan L. Ontology-based information selection. PhD thesis. University of Southern California: Faculty of the Graduate School; 2000.
[31] Miller G. WordNet: a lexical database for English, vol. 38. ACM; 1995. pp. 39–41.
[32] Resnik P. Disambiguating noun groupings with respect to WordNet senses. In: 3th workshop on very large corpora. 1995. p. 54–68.
[33] Boubekeur F, Boughanem M, Tamine L, Daoud M. Using WordNet for concept-based document indexing in information retrieval. In: Proceedings of conference on advances in semantic processing. 2010.
[34] Baziz M, Boughanem M, Aussenac-Gilles N. A conceptual indexing approach based on document content representation. In: Conference on conceptions of libraries and information science. 2005. p. 171–86.
[35] Kang B, Lee S. Document indexing: a concept-based approach to term weight estimation. Journal of Information Processing and Management 2005;41:1065–80.
[36] Salton G, McGill MJ. Introduction to modern information retrieval. New York, NY, USA: McGraw-Hill, Inc.; 1986.
[37] Harman D, Benson D, Fitzpatrick L, Huntzinger R, Goldstein C. IRX: an information retrieval system for experimentation and user applications. SIGIR Forum 1988 May;22:2–10.
[38] Trieschnigg D. Proof of concept: concept-based biomedical information retrieval. PhD thesis. University of Twente; 2010.
[39] Dinh D, Tamine L. Biomedical concept extraction based on combining the content-based and word order similarities. In: Proceedings of the 2011 ACM symposium on applied computing, ACM symposium on applied computing. 2011. p. 1159–63.
[40] Hersh W, Greenes R. SAPHIRE – an information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. Computers and Biomedical Research 1990;23:405–20.
[41] Hersh W, Hickam DH, Haynes RB, McKibbon KA. Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. Proceedings of Annual Symposium on Computer Applications in Medical Care 1991;15:808–12.
[42] Krauthammer M, Rzhetsky A, Morozov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. Gene 2000;259(1–2):245–52.
[43] Zhou X, Zhang X, Hu X. MaxMatcher: biological concept extraction using approximate dictionary lookup. In: Proceedings of the 9th Pacific rim international conference on artificial intelligence, PRICAI'06. 2006. p. 1145–9.
[44] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries 2000;3:115–30.
[45] Hliaoutakis A, Zervanou K, Petrakis EGM. The AMTEx approach in the medical document indexing and retrieval application. Data Knowledge Engineering 2009:380–92.
[46] Hersh WR, Greenes RA. SAPHIRE – an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Computers and Biomedical Research 1990 September;23:410–25.
[47] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA symposium. 2001. p. 17–21.
[48] Zhou X, Zhang X, Hu X. Using concept-based indexing to improve language modeling approach to genomic IR. In: Proceedings of the 28th European conference on advances in information retrieval, ECIR'06. 2006. p. 444–55.
[49] Zhou W, Yu CT, Smalheiser NR, Torvik VI, Hong J. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'07. 2007. p. 655–62.
[50] Lange T, Buhmann J. Fusion of similarity data in clustering. In: Weiss Y, Schölkopf B, Platt J, editors. Advances in neural information processing systems 18. Cambridge, MA: MIT Press; 2006. p. 723–30.
[51] Büttcher S, Clarke CLA, Cormack GV. Domain-specific synonym expansion and validation for biomedical information retrieval. In: Proceedings of the thirteenth Text REtrieval Conference. 2004.
[52] Chang JT, Schütze H, Altman RB. Creating an online dictionary of abbreviations from MEDLINE. Journal of the American Medical Informatics Association 2002;9:612–20.
[53] Robertson SE, Walker S, Hancock-Beaulieu M. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In: Proceedings of Text REtrieval Conference. 1998. p. 199–210.
[54] Amati G. Probabilistic models for information retrieval based on divergence from randomness. PhD thesis. Department of Computing Science, University of Glasgow; 2003.
[55] Clinchant S, Gaussier E. Information-based models for ad hoc IR. In: Proceedings of conference on research and development in information retrieval, SIGIR'10. 2010. p. 234–41.
[56] Amati G, Carpineto C, Romano G, Bordoni FU. Query difficulty, robustness and selective application of query expansion. In: European conference on information retrieval. 2004. p. 127–37.
[57] Hersh WR, Bhuptiraju RT, Ross L, Johnson P, Cohen AM, Kraemer DF. TREC 2004 Genomics track overview. In: Proceedings of the Text REtrieval Conference (TREC). 2004.
[58] Hersh WR, Cohen AM, Yang J, Bhupatiraju RT, Roberts PM, Hearst MA. TREC 2005 Genomics track overview. In: Proceedings of the Text REtrieval Conference (TREC). 2005.
[59] Dinh D, Tamine L. Combining global and local semantic contexts for improving biomedical information retrieval. In: European conference on information retrieval (ECIR). 2011. p. 375–86.
[60] Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society Series B (Methodological) 1991;53(3):683–90.
[61] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN 3-900051-07-0.
[62] Abdou S, Savoy J. Searching in MEDLINE: query expansion and manual indexing evaluation. Information Processing Management 2008;44(2):781–9.