Inf Retrieval J (2015) 18:413–444
DOI 10.1007/s10791-015-9264-0

Mining document, concept, and term associations for effective biomedical retrieval: introducing MeSH-enhanced retrieval models

Jin Mao[1] · Kun Lu[2] · Xiangming Mu[3] · Gang Li[1]

Received: 11 August 2014 / Accepted: 14 August 2015 / Published online: 4 September 2015

Abstract Manually assigned subject terms, such as Medical Subject Headings (MeSH) in the health domain, describe the concepts or topics of a document. Existing information retrieval models do not take full advantage of such information. In this paper, we propose two MeSH-enhanced (ME) retrieval models that integrate the concept layer (i.e. MeSH) into the language modeling framework to improve retrieval performance. The new models quantify associations between documents and their assigned concepts to construct conceptual representations for the documents, and mine associations between concepts and terms to construct generative concept models. The two ME models reconstruct two essential estimation processes of the relevance model (Lavrenko and Croft 2001) by incorporating the document-concept and the concept-term associations. More specifically, in Model 1, language models of the pseudo-feedback documents are enriched by their assigned concepts. In Model 2, concepts that are related to users' queries are first identified, and then used to reweight the pseudo-feedback documents according to the document-concept associations. Experiments carried out on two standard test collections show that the ME models outperformed the query likelihood model, the relevance model (RM3), and an earlier ME model. A detailed case analysis provides insight into how and why the new models improve/worsen retrieval performance. Implications and limitations of the study

✉ Kun Lu
kunlu@ou.edu

Jin Mao
danveno@163.com

Xiangming Mu
mux@uwm.edu

Gang Li
ligang@whu.edu.cn

[1] Center for the Studies of Information Resources, Wuhan University, Bayi Road 299, Wuhan 430072, Hubei, China

[2] School of Library and Information Studies, University of Oklahoma, Norman, OK 73019, USA

[3] School of Information Studies, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA

are discussed. This study provides new ways to formally incorporate semantic annotations, such as subject terms, into retrieval models. The findings of this study suggest that integrating the concept layer into retrieval models can further improve the performance over the current state-of-the-art models.

Keywords Relevance model · Concept · MeSH-enhanced retrieval models · Health information retrieval

1 Introduction

The rapid growth of scientific literature in the health domain calls for more effective retrieval systems. When searching health information, vocabulary problems have been found to be a big hurdle for general users (Zielstorff 2003). The vocabulary mismatch between user queries and documents is well acknowledged as a common failure in information retrieval (Metzler et al. 2007). This problem has been particularly well noted in the health domain (Poikonen and Vakkari 2009; Zeng, et al. 2002; Zhang et al. 2008). User queries have been found to be short, generally consisting of one or two words on average (Zeng et al. 2002), and their terms are significantly different from the ones in professional thesauri (Zhang et al. 2008) and in documents. One reason for that is the inexperience with topics (Guisado-Gámez et al. 2013). If users are not familiar with the topic, they may have difficulties in formulating effective queries (Zeng et al. 2004). From one standpoint, authors may use different terms to express the same meaning in different documents, or even in the same document. These problems pose obstacles to successful health information retrieval.

Many technologies have been developed to address the vocabulary problem. One way is to bridge the terminology gap between information resources (i.e. documents) and information needs (i.e. queries). The use of controlled vocabularies is an attempt to bridge the terminology gap. A controlled vocabulary is a carefully constructed knowledge organization system that can be used to describe the concepts of documents and user queries (Kamps 2004; Plaunt and Norgard 1998). By standardizing different expressions of the same concept, controlled vocabularies aim to solve the vocabulary problem in information retrieval.

In the health domain, Medical Subject Headings (MeSH) is the most popular thesaurus. It plays an important role in bridging the terminology gap. MeSH is used to describe the biomedical literature in the MEDLINE/PubMed database and to help users find information. MeSH terms assigned to a document are topically relevant to the document content according to the judgment of professional indexers. Many resources are invested to assign MeSH terms to documents in the hope of better retrieval. However, the issue of how to effectively use MeSH terms in health information retrieval is still under discussion to date. MeSH has been used to improve retrieval performance in a number of ways, such as query expansion (Stokes et al. 2009), or terminology assistance for users (Zeng et al. 2006). In these methods, MeSH terms that are related to users' information needs are first identified either automatically or manually by users. These related MeSH terms are then added to the original queries to try to solve the vocabulary mismatch problems between user queries and documents. Both positive and negative results have been found in previous research (Hersh and Bhupatiraju 2003; Abdou et al. 2005; Hersh et al. 2003; Guo et al. 2004; Bacchin and Melucci 2005; Lu et al. 2009).

In this paper, we explore a different approach to use MeSH in retrieval by integrating the controlled vocabulary into retrieval models. MeSH terms are considered as explicit conceptual representations of the documents to which they are assigned. The assumption is that a document is represented by several concepts (i.e. MeSH terms), and these concepts are further elaborated by the terms in the document. In this approach, MeSH becomes an integrated layer between documents and terms. Through exploiting associations between documents and concepts (i.e. MeSH terms), and concepts and terms via text mining approaches, MeSH can be seamlessly integrated into the generative process and help retrieve conceptually relevant documents. The difference between this approach and most existing ones is that the controlled vocabulary is a representation layer of the retrieval models in our approach rather than being used as terms or used to find other related terms that are directly added to original queries. We refer to this new approach of using MeSH as a "*MeSH-enhanced retrieval model*" (ME model). The three-layer structure in our models is similar to the documents, topics, and terms structure in the topic model (Wei and Croft 2006; Blei et al. 2003). However, in our models, the concept layer consists of the manually assigned MeSH terms rather than latent topics to be inferred, as in the LDA model (Blei et al. 2003). Of all the existing approaches that use controlled vocabularies in IR, the models proposed by Meij et al. (Meij and de Rijke 2007; Meij et al. 2010) are the only few that employ this concept (i.e. MeSH-enhanced retrieval model). Their models use a pseudo-relevance feedback approach based on relevance models (Lavrenko and Croft 2001). The MeSH terms assigned to literature are regarded as conceptual representations for the documents. Their experiments found significant improvements in the MeSH-enhanced model over the query likelihood model and a relevance model (i.e. *RM2*). While their studies represent early attempts to formally integrate controlled vocabularies into retrieval models, further investigations are needed to improve our understanding on the issue. This study proposes two MeSH-enhanced retrieval models. The two ME models consider MeSH terms as a concept layer between documents and terms, and integrate the concept layer into the generative process of the relevance model (Lavrenko and Croft 2001). The conceptual representations for documents are achieved by mining the associations between the documents and their assigned concepts. However, the two models conceptualize the generative process a bit differently. In the first ME model, the associations between concepts and terms are mined to represent the concept layer. The concept layer (i.e. MeSH terms) is added to the process of estimating the document language models of the pseudo-feedback documents to enrich the document language models with concept level characteristics. In the second ME model, the relevance model is constructed by considering a document as a distribution of relevant concepts. Before estimating the relevance model, the pseudo-feedback documents from the first run are reweighted according to these relevant concepts. Our models can be considered as a further development of Meij's model. First, we propose two different ways of integrating the concept layer into retrieval models. Second, our models use a three-layer structure: documents, concepts, and terms, while Meij's work only formally models the concept layer and the term layer. In a sense, Meij's model is similar to our second ME model without the document layer.

Experiments are carried out on two standard test collections, the Ohsumed and the TREC Genomics 2006, to compare the proposed retrieval models with three baseline models. The study aims to address the following research questions:

1. How effective are the two newly proposed ME models?
   This research question addresses whether these MeSH-enhanced retrieval models are more effective than the state-of-the-art models.
2. How do the parameters of the models influence the retrieval performance?
   Parameters need to be examined for the proposed models to evaluate how they impact retrieval results. In particular, we focus on the impact of those parameters that are adopted in the estimation processes of the proposed ME models.
3. In which cases do the ME models improve/worsen the retrieval performance? And why?
   This research question focuses on the specific cases where ME models improve/worsen the performance and aims to address the question of why the ME models improve/worsen the performance.

An understanding of the answers to these research questions contributes to the effective use of MeSH terms in health information retrieval. It should be noted that although the experiments in this study use MeSH in the health domain, the same idea can be applied to other controlled vocabularies and other domains. In Sect. 2, we briefly introduce the related work. We elaborate on the generative processes of the two ME models in Sect. 3 and the probability estimation in detail in Sect. 4. The experimental setups and evaluations are described in Sect. 5, and the experimental results and discussions are provided in Sect. 6. Conclusions and future directions of our research work are provided in Sect. 7.

2 Related work

The related work of this study can be grouped into the following areas: query expansion, methods of optimizing query models in language modeling, and information retrieval using MeSH.

2.1 Query expansion

Query expansion is a popular technique to address the vocabulary gap between queries and documents. The general approach is to add and/or reweight terms given the users' initial queries to improve the quality of the search results (Voorhees 1994). Early studies on query expansion attempted to extract relevant terms from thesauri to formulate a better query (Gauch and Smith 1991). Besides using terms from thesauri, discovering term relationships based upon their co-occurrences in documents or lexical co-occurrences is also an effective approach to picking up candidate terms for expansion (Gauch and Smith 1991; Vechtomova et al. 2003). Term relationships can be considered in the global or the local context (Vechtomova et al. 2003; Xu and Croft 1996). In the global context, statistics about the collection or external knowledge resources are used to identify candidate terms that can be added into original queries. In the local context, the co-occurrence of terms is counted in the context of a specific query. For example, information about a user and her/his queries is often considered as the local context, such as the user's history or profile (Korfhage 1984), or the appearance of the query terms in documents (Finkelstein 2002). Relevance feedback is another popular local technique to find relevant terms for query expansion by analyzing the relevance feedback documents from the initial retrieval runs. Due to the high cost of obtaining relevance judgment from real users, pseudo-feedback (or pseudo relevance feedback) is often used as a substitute, in which the top-ranked documents are regarded as relevant (Manning et al. 2008).

2.2 Methods of optimizing query models in language modeling

Sharing similar ideas as query expansion is another approach that adds related terms to or reweights terms for query language models to improve the effectiveness of information retrieval. For example, corpus language models were used to smooth query models and document models in Zhai and Lafferty (2001a). However, this method only considers the global features of the corpus and ignores the local characteristics of the query. Pseudo-feedback documents represent local characteristics of the query and are often used as the source to estimate the query model (Zhai and Lafferty 2001b; Lavrenko and Croft 2001; Lafferty and Zhai 2003). The underlying assumption of pseudo-feedback documents is that top-ranked documents are relevant to users' queries. This may not always be true and can potentially bring noise to the estimation of a query model. Many efforts have been made to refine pseudo-feedback documents so that relevant terms for a given information need can be singled out to optimize the query model.

One approach is to bring in more relevant documents as pseudo-feedback documents. Offline document clusters have been used to find additional relevant documents in the corpus (Kurland and Lee 2004; Liu and Croft 2004). The hypothesis is that "closely associated documents tend to be relevant to the same requests" (van Rijsbergen 1979). Therefore, document models of pseudo-feedback documents can be smoothed by the clusters to which they belong.

Moreover, pseudo-feedback documents can be refined by picking up or putting more weights on relevant documents (He and Ounis 2009; Lv et al. 2011). Since relevant documents are more likely to belong to the same cluster (or a few clusters), isolated documents in pseudo-feedback documents can be deemed as irrelevant. Under this hypothesis, a more accurate query model can be estimated from refined pseudo-feedback documents. For example, Lee et al. (2008) proposed a resampling method by applying overlapping clusters to select dominant documents, which are connected with many sub-topic clusters and have several highly similar documents. This resampling approach showed higher relevance density and better retrieval accuracy in their experiments. Kurland (2008, 2009) further differentiated the query-specific clusters according to the presumed percentage of relevant documents they contain. The ranked clusters were then used to produce document ranks, yielding an improvement of the top cutoff metrics over the initial ranking.

The aforementioned research aims to improve the accuracy of the query model by refining pseudo-feedback documents. From the perspective of finding related terms, explicit or implicit term relationships can also be exploited (Bai et al. 2005). Large external knowledge resources, such as WordNet (Gonzalo et al. 1998) and Wikipedia (Gabrilovich and Markovitch 2007), provide explicit semantic relationships between terms and concepts. Implicit term relationships can be inferred according to their co-occurrences or using more sophisticated methods (Bai et al. 2005).

2.3 Information retrieval using MeSH

Literature in databases (for example, PubMed) has been annotated with MeSH terms by professionals (Gault et al. 2002; Mata et al. 2012; Shin and Han 2004). The use of MeSH in information retrieval has not always been successful. Some inconsistent findings have been reported (Lu et al. 2009). Positive results were found by some researchers (Abdou et al. 2005; Srinivasan 1996), while negative results were reported by some others (Bacchin and Melucci 2005; Hersh 2008).

MeSH terms are used in information retrieval in different ways. Query expansion is the first venue of applying MeSH in information retrieval. Terms in free texts, either in users' queries or in pseudo-feedback documents, can be mapped to MeSH terms, which are then added to the original queries directly (Griffon et al. 2012; Mata et al. 2012). Synonyms or hyponyms of the terms, can also be drawn through the semantic structure of the thesauri (e.g. MeSH) to perform further query expansion (Díaz-Galiano et al. 2008). Additionally, different query expansion strategies have been discussed. Srinivasan (1996) investigated three query expansion strategies using MeSH terms: expansion via an inter-field statistical thesaurus, expansion via relevance feedback, and expansion using a combined approach. Lu et al. (2009) used MeSH in their automatic query expansion process. Stokes et al. (2009) examined genomic query expansion with different knowledge sources including MeSH. Their results suggest that query expansion with domain-specific knowledge sources is preferable. The essence of these methods is to expand the given query with candidate terms through the semantic relationships embedded in MeSH.

Another way to use MeSH in retrieval is to implicitly integrate MeSH into retrieval models in which MeSH terms that are assigned to documents are used to formulate conceptual representations of documents or queries. With the conceptual representation (i.e. MeSH), such retrieval models are more likely to identify relevant concepts and solve the terminology mismatch problem. Very few studies were found to have adopted this approach. Trieschnigg (2010) proposed a framework for concept-based retrieval by applying conceptual representations to complement textual representations. In his model, a concept-based query model is obtained by translating query words into concepts and is used to determine the coverage of the words in the original word-based query model. Terms that are well covered by the concept-based representations will receive lower weights in the textual query model. The final query model consists of the concept-based query model and the updated word-based query model. Essentially, this model acknowledged that conceptual representations of queries can cover some aspects of information need that are not reflected in the textual representations. Similar to the relevance model, Meij et al. (2010) developed a concept-based retrieval model based on pseudo feedback technique to directly model the relevance via conceptual representations. A two-step translation is applied: translating a query to a conceptual representation, and then translating the conceptual query back into a textual query model using concept language models. In this model, concepts serve as a pivot language between the relevance and terms. In addition, concepts in conceptual representations may also come from automatically annotated MeSH concepts (Trieschnigg et al. 2009).

In this study, we propose two ME retrieval models that integrate MeSH into retrieval processes. Differing from treating concepts as a pivot language as in Meij et al. (2010), our models enhance the relevance model by mining the associations between concepts and other elements, namely, terms, documents, and user's queries.

3 MeSH-enhanced relevance models

The essential idea of the ME models is to treat MeSH terms as a concept layer between documents and terms. The concept layer is then included in the generative process.

3.1 The relevance model

As one of the state-of-art retrieval models, the relevance model (specifically *RM3* used in this paper) has shown effective and robust performance (Lv and Zhai 2009). The theory of the relevance model is to estimate a language model for a user's information need, deemed as the "relevance model (*R*)", using the pseudo-feedback documents (i.e. top-ranked documents from the first retrieval pass) for a given query (Lavrenko and Croft 2001).

The generative process of the *RM3* is that the relevance model (*R*) first generates relevant documents (approximated by pseudo relevant documents), and then each document generates terms (Fig. 1). The *RM3* is estimated through the Eq. (1):

$$P(w|R) \approx \sum_{d \in \Theta} P(w|\mathrm{d}) P(\mathrm{d}) P(d|R) \approx \sum_{d \in \Theta} P(w|\mathrm{d}) P(d|R) \tag{1}$$

where R denotes the relevance model, Θ is the set of pseudo-feedback documents, d is a document in Θ, $P(d)$ is the prior probability of the document language model which is often assumed to be uniform, and $P(w|d)$ can be estimated by the Maximum Likelihood Estimator (*MLE*), often smoothed with the corpus language model. The probability $P(d|R)$ is actually the query likelihood score for the document d, which can be obtained from the first retrieval pass. The relevance model estimated from Eq. (1) is then interpolated with the original query model to obtain the final query model of *RM3*. Essentially, this relevance model is a combination of the weighted document language models in the pseudo relevant feedback set. The probabilities of terms in the relevance model are determined by two parts: the document language models and the weight (determined by the query likelihood score) of each individual pseudo-feedback document. In the two MeSH-enhanced retrieval models, the two parts are enhanced with additional information from the concept layer.

3.2 ME model 1

In Lavrenko and Croft (2001), the relevance model is estimated from document language models of the pseudo relevant documents. The essential idea of our first method is to add a concept layer to the document language model estimation process to capture the concept level characteristics. In ME model 1, document language models of the pseudo relevant documents are enriched by the terms associated with the concepts assigned to the documents. We view this alternative relevance model generative process as follows (Fig. 2):

The generative process of ME model 1 is that the relevance model (*R*) first generates the relevant documents (still approximated by the pseudo relevant documents), each document then generates a number of concepts (represented by MeSH descriptors), and then each concept generates terms.

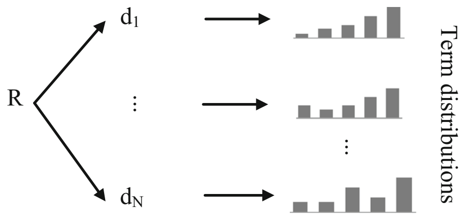
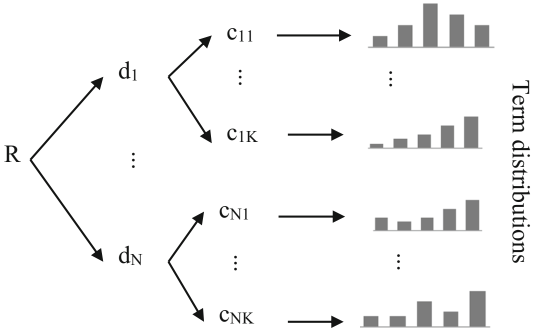


Fig. 1 Generative process of *RM3*

Fig. 2 Generative process of ME model 1 (c_{ij} denotes the *j*th concept (i.e. MeSH term) that is generated by d_i)



ME model 1 employs a different approach to estimate $P(w|d)$. Instead of using the MLE directly, we add a concept layer between the document model and the terms. As each of the documents is assigned with a number of MeSH terms, the document model can be estimated through the concepts that connect documents and terms. Then, the final document model is obtained via a linear interpolation of the document language model estimated from the above process with the original document model $P(w|d')$ by the Jelinek-Mercer smoothing method (Jelinek and Mercer 1980), shown in Eq. (2).

$$P(w|d) = \lambda_{m1} \sum_{c \in \Gamma_d} P(w|\mathrm{c})P(c|d) + (1 - \lambda_{m1})P(w|d') \tag{2}$$

where Γ_d is the set of concepts that are assigned to d, $P(c|d)$ is the probability of the concept c given the document d, λ_{m1} is the interpolating parameter to control the portion of the original document model in the final document model, and $P(w|c)$ is the probability that the concept c generates the term w, which will be estimated from the associations between concepts and terms. With this method, we can integrate the concepts into the retrieval model.

In ME model 1, each concept is represented as a distribution of terms using a multinomial unigram language model (Meij et al. 2010). We use $P(w|c)$ when referring to this multinomial unigram model for a concept (i.e. generative concept model). Likewise, we define $P(c|d)$, the conceptual document model, such that each document can be represented as a multinomial distribution over the concepts that are assigned to the document. We will further elaborate on our estimation methods in Sect. 4. The final equation of model 1 can be obtained by substituting Eq. (2) for $P(w|d)$ in Eq. (1):

$$P(w|R) \approx \sum_{d \in \Theta} \left(\lambda_{m1} \sum_{c \in \Gamma_d} P(w|\mathrm{c})P(c|d) + (1 - \lambda_{m1})P(w|d') \right) P(\mathrm{d}|R) \tag{3}$$

The assumption is that the additional concept layer can potentially enrich the document models by uncovering the concept-term and the document-concept associations.

3.3 ME model 2

In the second model, we conceptualize the generative process differently. Instead of using the concept layer to enrich the document model, we put forward a new method to generate the probability $P(d|R)$. This generative probability plays an important role in the robustness of the relevance model (Lv and Zhai 2009). However, the scores of pseudo-feedback documents in *RM3* do not consider any concept level features. We propose an alternative way to re-rank pseudo-feedback documents by adding the concept layer between the relevance model and documents. The assumption is that the essence of a relevance model consists of some relevant concepts that users are looking for. Therefore, a relevance model first generates relevant concepts, and then these concepts are expressed in documents (Fig. 3). In this method, we approximate the conditional probability $P(\mathrm{d}|R)$ through relevant concepts rather than directly using query likelihood score.

The generative process is as follows:

$$P(d|R) \approx \sum_{\mathrm{c} \in \Theta} P(d|\mathrm{c}) P(\mathrm{c}|R) \tag{4}$$

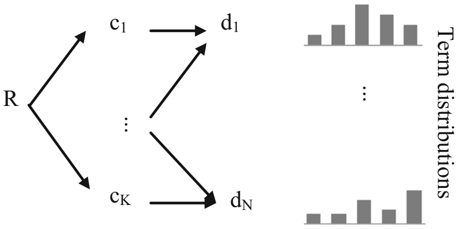
where $c \in \Theta$ denotes all the concepts in the pseudo-feedback documents, $P(c|R)$ is the conceptual relevance model estimated by the *MLE*, $P(d|c)$ can be obtained through the conceptual document model, and $P(c|d)$ by the Bayesian rule. The ultimate equation of model 2 can be obtained by substituting Eq. (4) into Eq. (1):

$$P(w|R) \approx \sum_{d \in \Theta} P(w|d) \sum_{\mathrm{c} \in \Theta} P(d|\mathrm{c}) P(\mathrm{c}|R) \tag{5}$$

4 Mining the document, concept and term associations

Section 3 outlined the generative process of the ME Model 1 and Model 2, where the concept layer is integrated into the retrieval process. In the final equations (Eq. 3 and Eq. 5), the generative concept model $P(w|c)$ and the conceptual document model $P(c|d)$ need to be estimated. This section further elaborates the estimation of these two probabilities. It should be noted that instead of using some more popular probability estimation methods such as MLE, this paper employs TF-IDF and weighted mutual information, which appear to be more heuristic. This is due to considering two aspects. From the practical standpoint, the MLE method does not adequately distinguish the importance of term occurrences. This has also been noted in Meij et al. (2010) where EM estimation is used. We have noted that the EM method achieves a similar effect as the TF-IDF method does but is more computationally expensive. This is particularly true in our three-layer structure. From the theoretical standpoint, at a more fundamental level (explained by Fang and Zhai (2005) as the “axiomatic framework”), all retrieval models, including vector space model, language model, and classic probabilistic model, need to satisfy certain constraints that directly relate to the concept of relevance. The heuristics of TF-IDF method can be explained by these retrieval constraints (Fang and Zhai 2005). In fact, existing models, including language models, all contain some TF-IDF like components.

Fig. 3 Generative process of ME model 2



4.1 The association between concepts and terms

The generative concept model *P(w|c)* captures the association between concepts and terms. As mentioned earlier, the MLE method does not adequately distinguish the importance of term occurrences. General terms that occur frequently in the collection usually receive higher weights. In our models, the importance of terms in the generative concept model is calculated by the *TF-IDF* weighting method that not only considers term occurrences but also the importance of the terms. Terms from the documents that are assigned with the concept are used to represent the concept (the generative concept model). The set of documents that are assigned with the concept are treated as a sub-collection for the concept. Then, *TF-IDF* weighting is applied to the terms in this sub-collection to calculate the importance of terms in representing the concept:

$$\mathrm{tfidf}_{w,c} = (0.5 + \sum_{d \in \Gamma_{\mathrm{c}}} tf_{w,d}) * \log\left(\frac{N + 0.5}{df_w + 0.5}\right) \tag{6}$$

where $tf_{w,d}$ is the term frequency for the term w in document d, Γ_c is the set of documents assigned with the concept c, df_w denotes the document frequency (the number of documents in the entire document set containing the term w), and N is the number of documents in the entire document set.

Then, we obtain the generative concept model by normalizing all the term weightings of the concept:

$$\mathrm{P}(w|c) = \frac{\mathrm{tfidf}_{w,c}}{\sum_{w \in V} \mathrm{tfidf}_{w,c}} \tag{7}$$

4.2 The association between documents and concepts

In our models, we quantify the associations between documents and their assigned concepts, and regard a document as a probability distribution over the concepts, namely, the conceptual document model. The conditional probability $P(c|d)$ is the probability of the concept c given the document d. *MLE* and *EM* have been applied to estimate this probability in previous studies (Meij et al. 2010). This study employs weighted mutual information to quantify the semantic associations between documents and their assigned concepts, which has been found to be effective in a previous study (Lu and Mao 2013).

To measure the associations between documents and the assigned concepts, we calculate the weighted mutual information between a document d and the assigned concept c. The formula is represented as:

$$\mathrm{I}(d; \mathrm{c}) = \sum_{t \in d} w(t, c) P(t, c) \log \frac{P(t, c)}{P(t) P(c)} \tag{8}$$

where $w(t, c)$ is the weight of the pair $\langle t, c \rangle$, t represents a term in the document and c is a concept associated with the document. *TF-IDF* is adopted to calculate the weights:

$$\mathrm{w(t, c) = w(t) * w(c) = (tf + 0.5)} * log \frac{N + 0.5}{df_t + 0.5} * \frac{N + 0.5}{df_c + 0.5} \tag{9}$$

where N is the total number of documents in the collection, df_t is the document frequency of term t, df_c is the document frequency of concept c. The logarithm of the concept IDF is intentionally removed to place more emphasis on specific concepts. With respect to $P(t, c), P(t)$ and $P(c)$, maximum likelihood estimation can be applied. For multinomial distributions, if the document frequency of the object ι the corpus is $\#(\iota)$, the probability can be calculated as:

$$\mathrm{P}(\iota) = \frac{\#(\iota)}{N} \tag{10}$$

Finally, we obtain the probability for concept c in document d by normalizing the weighted mutual information of all concepts of the document.

$$P(c|d) = \frac{\mathrm{I}(d; \mathrm{c})}{\sum_{c' \in \Gamma_d} \mathrm{I}(d; c')} \tag{11}$$

5 Experimental setup

To address the research questions, experiments are conducted on standard IR test collections in the health domain. This section introduces the experimental settings in detail.

5.1 Test collections

Two standard IR test collections are used in the experiments: the Ohsumed (Hersh et al. 1994) and the TREC Genomics Track 2006 (Hersh et al. 2006). The Ohsumed collection is a subset of MEDLINE containing 348,566 references (without full-text) over a five-year period (1987–1991) with MeSH descriptors. The TREC Genomics Track 2006 is a full-text collection with 162,259 documents. The original Genomics collection does not include MeSH descriptors of the documents. A program was developed to fetch MeSH descriptors of each document from the database system at the National Center for Biotechnology Information (NCBI) using the E-Utilities tool.[1] Out of the 162,259 documents, 2215 were not found in the current version of NCBI databases. We added the MeSH descriptors to the rest of documents, and then indexed the expanded documents. We removed 1 test topic (topic 8) from the Ohsumed and 2 topics (topic 173 and 180) from the Genomics as there are no relevant documents for them in the relevance judgments.

In our experiments, we use the main MeSH headings as the concepts (i.e. the qualifiers are not considered). Therefore, a main heading with different qualifiers is considered as the same concept. For example, "*Wound Infection/PC*" and "*Wound Infection/MI*" were both transformed into the main heading "*Wound Infection*", and thus were regarded as the same concept. The language model toolkit, Lemur,[2] was used to index the two collections. The Ohsumed collection was indexed by the following fields: title, MeSH, author, publication type, abstract, and source. Documents in the Genomics collection were indexed as a whole, not by fields. Index terms were stemmed by the Krovetz stemmer. The InQuery's standard stoplist with 418 stop words was used. When constructing generative concept models, single-character terms and numbers (e.g. 2001) were removed. Table 1 lists the statistics of the Ohsumed and Genomics collections.

[1] http://eutils.ncbi.nlm.nih.gov/entrez/eutils/.

[2] http://www.lemurproject.org/.

Table 1 Statistics of the document collections used in the experiments

Collections	Documents	Terms	Avg. terms per doc.	Unique concepts	Concept occurrences	Avg. concepts per doc.
Ohsumed	348,566	58,431,800	167.63	14,638	3,696,239	10.60
Genomics	162,259	1,076,766,373	6636.09	24,622	2,483,152	15.30

5.2 Baselines and evaluation measures

Two general retrieval models are included as the baseline. One is the query likelihood model (*QLH*), and the other is the *RM3*, a state-of-the-art retrieval model using a pseudo-feedback technique (Lv and Zhai 2009). In addition, an earlier MeSH-enhanced retrieval model proposed by Meij et al. (2010) is also included to compare the new models with an earlier model that formally incorporates MeSH. In Meij et al. (2010), the authors proposed a conceptual language model with an EM algorithm (abbreviated *GC*) together with a maximum likelihood based conceptual language model (named *MLGC*). Since the *GC* model is very time consuming and the results of the *GC* model are only insignificantly improved in terms of recall and MAP over the *MLGC,* we chose the *MLGC* model as our baseline.

After estimating the relevance model, we interpolated it with the original query language model:

$$\mathrm{P}\left(\mathrm{w}|\theta'_Q\right) = \lambda P(\mathrm{w}|\mathrm{R}) + (1-\lambda)\mathrm{P}(\mathrm{w}|\theta_Q) \tag{12}$$

where $\mathrm{P}(\mathrm{w}|\theta_Q)$ is the original query language model, $\mathrm{P}(\mathrm{w}|\mathrm{R})$ is the relevance model generated by our models, the *RM1* model, or the *MLGC* model, λ is the parameter to control the proportion of contribution from the original query model and the estimated relevance model in the final query model.Then, the KL-divergence language modeling retrieval framework (Lafferty and Zhai 2001) is used to rank the retrieval results. The KL-divergence between a query model θ_Q and a document model θ_D is calculated as (Zhai 2002; Zhai and Lafferty 2001b):

$$\mathrm{D}(\theta_Q;\theta_D) = \sum_w P(w|\theta_Q) log\frac{P(w|\theta_Q)}{P(w|\theta_D)} \approx -\sum_w P(w|\theta_Q) logP(w|\theta_D) \tag{13}$$

The Dirichlet prior smoothing method (Zhai and Lafferty 2004) was used in all our retrieval models and baseline models.

$$\mathrm{P}(\mathrm{w}|\mathrm{d}) = \frac{c(w;d) + \mu P(w|C)}{|d| + \mu} \tag{14}$$

where $c(w;d)$ is the frequency of a term w in a document d, *p(w|C)* is the language model for the whole corpus, μ is the hyperparameter, and |*d*| is the length of document *d*. Some studies applied machine learning method to predict the optimal value of the hyperparameter μ (Zhai and Lafferty 2002). However, in more recent experiments, this parameter is found not to be a vital concern (Meij et al. 2010). In our experiments, the parameter μ was set to 500 for the Ohsumed and 2000 for the Genomics.

In terms of performance evaluation, the top 500 documents returned from all retrieval runs were compared in terms of mean average precisions (MAP), the eleven point precision-recall curve, and top cut-off metrics (P@5 and P@10) (Manning et al. 2008). A randomization test with 100,000 samples, with a significance level of 0.05, was used for all statistical tests as is suggested by Smucker et al. (2007).

5.3 Parameter tuning

In the estimation process for the MeSH-enhanced models, some controlled parameters need to be tuned (see Table 2). The number of pseudo-feedback documents N was set to vary from 1 to 10. In Model 1, the number of terms used to construct the generative concept model, $|V_c|$, swept from 10 to 300 with an increment of 10 at each step. The parameter λ_{m1}, which is used to control the interpolation between the original document model and the expanded document model, was set to 0, 0.1, 0.2, …, 1.0 respectively. The parameter λ that is used to control the proportion of the original query model in the final query model was set to 0.5 for RM3, MLGC, and our models. The parameter N varied in the same fashion for these models as well. For MLGC, the number of concepts for the conceptual query representation ($|c|$) was adjusted in the range [1, 10] with increments of 1, as in Meij et al. (2010).

In the KL-divergence retrieval process, we used 100 terms to represent the query model for RM3, MLGC, and our models, which was set heuristically by examining the performance of a few trial runs for the RM3. We determined the optimal parameter settings according to the MAP scores as in previous studies, such as Meij et al.(2010), Metzler and Croft(2005), Lafferty and Zhai (2001), and Zhai and Lafferty (2004).

6 Results and discussion

This section summarizes our experimental results. First, the results of the inferred concept-term and document-concept associations are presented. Then, a comparison of different retrieval models (three baseline models and two new models) is outlined. A further discussion on the influence of parameters is also provided.

6.1 The concept-term association and the document-concept association

The accuracy of the inferred concept-term and document-concept associations is crucial to the two MeSH-enhanced retrieval models, as the proposed models are based on the associations between documents and concepts, and concepts and terms.

6.1.1 Concept-term associations

The concept-term associations in the models are obtained through estimating the generative concept models. The generative concept model of a MeSH heading consists of the terms in the documents that are assigned with the concept. The *TF-IDF* of the terms in the generative concept model is used to measure the strength of the associations between the concept and the terms. Table 3 lists the top 10 terms from the generative concept models of two MeSH headings ("*Extracellular Space*" and "*Myocardial Diseases*") in the Ohsumed collection. It shows that the terms that are in the MeSH headings are assigned with high

Table 2 Parameters in ME models

Parameter	Model	Description
N	Model 1, Model 2	The number of pseudo-feedback documents
$\|V_c\|$	Model 1	The number of terms for generative concept model
λ_{m1}	Model 1	Interpolation between original document model and expanded document model
$\|c\|$	Model 2	The number of concepts for the relevance model
λ	Model 1, Model 2	Interpolation between original query model and the estimated relevance model

Table 3 Examples of generative concept models

MeSH concepts			
Extracellular space		Myocardial diseases	
Extracellular	0.0270	Myocardial	0.0231
Space	0.0140	Heart	0.0187
ca2	0.0107	Cardiomyopathy	0.0154
Fluid	0.0094	Ventricular	0.0121
Cell	0.0085	Cardiac	0.0119
Calcium	0.0077	Disease	0.0112
Volume	0.0076	Coronary	0.0083
Rat	0.0066	Hamster	0.0082
Intracellular	0.0064	Patient	0.0080
ph	0.0063	Left	0.0075

"Extracellular Space" and "Myocardial Diseases" are MeSH terms(concepts) in the Ohsumed collection

probabilities, such as "extracellular" and "space" for "*Extracellular Space*", and "myocardial" for "*Myocardial Diseases*". This suggests that the terms in the MeSH headings have strong semantic connections with the concept. Some other terms that are ranked high in the generative concept model appear to be related to the concept, such as "cardiomyopathy" and "ventricular" for the MeSH concept "*Myocardial Diseases*". On the other hand, those general terms that occur frequently in the collection were assigned with lower probabilities. For instance, the index term "human" that nearly occurs in every document annotated with the MeSH term "*Myocardial Diseases*" in the Ohsumed collection was assigned a relatively low probability of 0.0018. Therefore, it is not one of the top ranked terms in Table 3. In the Genomics collection, where full-text documents are available, it is also observed that terms related to the concepts are given higher probabilities than general terms. For example, for the concept "*Air Pollutants*" in the Genomics, the terms "inflammation" and "achy" were assigned higher probabilities (0.0139 and 0.0096 respectively) than the general term "adult" (0.0020).

The above examples demonstrate that the *TF-IDF* approach assigns appropriate weights to terms in generative concept models according to their semantic relationships with the MeSH concepts. It appears that the *TF-IDF* approach yields reasonable representations for the concept-term associations.

6.1.2 Document-concept associations

To determine the associations between documents and their assigned concepts, weighted mutual information is used (Eq. 8). Table 4 lists the results of all the MeSH terms and their weights for the document (id: 91052608) in the Ohsumed, titled "*Neurologic complications of cocaine abuse*". The major MeSH terms are marked with asterisks (*) to reflect the major points of the article. According to Table 4, the major MeSH descriptor "*Cocaine/**" has the highest weight (0.3897), indicating this MeSH term is the major point of the document. The MeSH term "*Human*" was assigned a much lower weight (0.0078), meaning that this MeSH term is not the major point of the document. Similar examples in which higher weights were assigned to major MeSH terms than to non-major MeSH terms can be easily found in the Genomics collection as well. It should be noted that mutual information does not always assign higher weights to major MeSH descriptors than to non-major ones. For example, the major MeSH descriptor "*Nervous System Diseases/*CI/PP*" was assigned a relatively low weight (0.0935). However, weighted mutual information is able to assign significantly higher weights to major MeSH descriptors in general. This has been validated in a previous study (Lu and Mao 2013).

In summary, an examination on the inferred document-concept and concept-term associations in the two test collections suggests that our method yields reasonable results. Terms that are related to the concepts are ranked relatively higher in the generative concept models (i.e. concept-term associations), and the major MeSH descriptors of the documents are generally assigned higher weights for document-concept associations (Tables 3 and 4).

6.2 Performance of different retrieval models

The essential idea of the two MeSH-enhanced retrieval models is to uncover the associations between documents and concepts, and concepts and terms, and formally integrate the concept layer into the retrieval process. The previous section indicates our methods yield reasonable results in inferring the associations. In this section, the focus is to empirically compare the retrieval performance of different models on the two test collections: Ohsumed and Genomics 2006. The retrieval performance reported in this section is based on the optimal parameter settings using the parameter tuning method introduced in Sect. 5.3. Therefore, the results of this section reflect the upper bound effectiveness of different models. This is consistent with the comparison methods in previous studies, such as Meij et al.(2010), Metzler and Croft(2005), Lafferty and Zhai (2001), and Zhai and Lafferty (2004).

Table 4 Examples of the association between documents and concepts

Document title: neurologic complications of cocaine abuse(#91052608 in Ohsumed)	
Cocaine/*	0.3897
Diazepam/TU	0.1886
Substance abuse/*CO	0.1865
Seizures/CI/DT	0.1339
Nervous system diseases/*CI/PP	0.0935
Human	0.0078

Table 5 MAP and precision at top cutoffs of different retrieval models (bold numbers are the best performance for each metrics)

Collection	Metrics	QLH	RM3	Model 1	Model 2
Ohsumed	MAP	0.2487	0.3058	**0.3269**$^{*,+}$	0.3168$^{*,+}$
	P@5	0.4095	0.4590	**0.5086**$^{*,+}$	0.4610*
	P@10	0.3657	0.4295	**0.4629**$^{*,+}$	0.4429*
Genomics	MAP	0.3527	0.3857	**0.4277**$^{*,+}$	0.3997*
	P@5	0.5385	0.5693	**0.6154**$^{*,+}$	0.5923$^{*,+}$
	P@10	0.4808	0.4961	**0.5115**	0.5077

* means statistically significant differences from the query likelihood model with a two-tailed randomization test at 0.05 level

$^{+}$ means statistically significant differences from the relevance model with a two-tailed randomization test at 0.05 level

Three baseline models are included in the study: query likelihood model (Ponte and Croft 1998), *RM3* (Lavrenko and Croft 2001), and an earlier MeSH-enhanced model proposed by Meij et al. (2010). Mean average precision, precision at top cutoffs (P@5 and P@10), and 11-point precision and recall charts are used to evaluate the retrieval performance (Manning et al. 2008).

The optimal parameter settings of different models are provided in Table 5. It can be observed that the best parameter settings depend on the collections. In the Ohsumed, Model 1 has the best performance when N (# of pseudo-feedback documents) equals 6, $|\mathrm{V_c}|$ (# of terms in generative concept model) equals 70, and λ_{m1} (the proportion of document model estimated through the concept layer) equals 1.0. Model 2 obtains its best performance when $N = 7$ and $|c| = 25$ (# of concepts used to represent each query). In the Genomics, Model 1 has its best performance when $N = 4$, $|V_c| = 250$, and $\lambda_{m1} = 0.5$, and Model 2 achieves the best performance when $N = 4$ and $|c| = 35$. Optimal parameters for RM3 and MLGC are also listed.

6.2.1 Comparing with general retrieval models

Table 5 lists the performance of the new MeSH-enhanced retrieval models and two baseline models that do not use MeSH terms, as measured by MAP and precision at top cutoffs.

According to Table 5, the performance of RM3 is better than that of the QLH in terms of almost all evaluation measures in both collections. This is consistent with previous findings that RM3 is superior to QLH. In terms of the new models, in the Ohsumed collection both of our proposed models showed significant improvements over the QLH model in all measures. When compared with the RM3, a very strong baseline as is shown in previous studies (Lv and Zhai 2009), general improvements are found from the new models. The performance of our Model 1 is significantly better than that of RM3 in terms of all the metrics. Also, the MAP of Model 2 is significantly higher than that of the RM3 in the Ohsumed collection. In the Genomics collection, where full-text documents are available, Model 1 shows significantly better performance over the QLH model in terms of MAP and P@5. The MAP and P@5 of Model 2 are significantly improved over that of the QLH model. As for comparing with the RM3, Model 1 and Model 2 had improved results over the RM3 in all metrics, and the improvements are statistically significant for Model 1

Table 6 Results of different MeSH-enhanced retrieval models

Collection	Metrics	MLGC	Model 1		Model 2	
Ohsumed	MAP	0.2895	**0.3269**	**13 %**	**0.3168**	**9 %**
	P@5	0.4705	**0.5086**	**8 %**	0.4610	−2 %
	P@10	0.4229	**0.4629**	**9 %**	0.4429	5 %
Genomics	MAP	0.3568	**0.4277**	**20 %**	0.3997	12 %
	P@5	0.5231	**0.6154**	**18 %**	**0.5923**	**13 %**
	P@10	0.4654	**0.5115**	**10 %**	0.5077	9 %

The bold numbers indicate significant improvements over the MLGC model with a two-tailed randomization test at 0.05 level

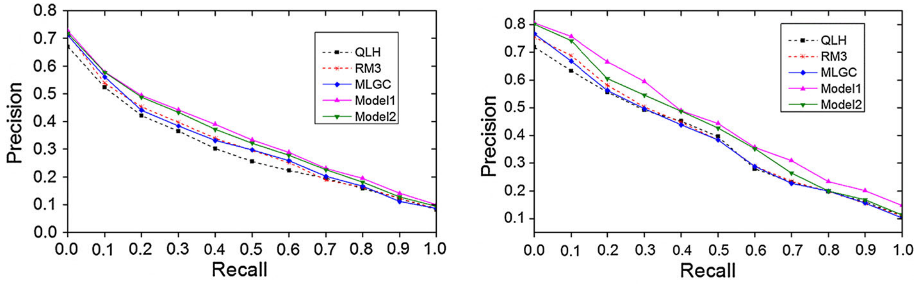
in MAP and P@5 and for Model 2 in P@5. In addition, Table 6 also suggests that the performance of Model 1 is slightly better than that of Model 2, although not significantly except for P@5 in the Ohsumed collection.

Using 11-point precision-recall charts can provide more insight into the performance along different recall values. According to Fig. 4, RM3 outperforms QLH in both collections. This is not surprising as many studies have confirmed the superiority of RM3 to QLH. Nearly all the curves of our models are above the curves of the QLH model and the RM3 model in both Ohsumed and Genomics collections. This indicates the superior performance of the new models to QLH and RM3 with respect to the ranked retrieval results. The precision-recall curve of Model 1 is on top of the curve of Model 2 in both Ohsumed and Genomics collection, which indicates that the performance of Model 1 is better than that of Model 2. The advantage of Model 1 is more obvious in the Genomics than in the Ohsumed, which is consistent with the results from MAP and precision at top cutoffs.

With all the evidence above, it appears that the proposed MeSH-enhanced models achieve better performance than the two robust state-of-the-art retrieval models: the QLH model and the RM3 model, which do not use MeSH terms in the two test collections. This means that the proposed ME models are effective in reconstructing the relevance model by mining the document, concept, and term associations.

6.2.2 Comparing with MeSH-enhanced retrieval model

With the findings that the new models outperform general baseline models that do not use MeSH terms, it is interesting to know how the new models perform when compared to the existing retrieval models that use MeSH. This section compares the results of different retrieval models that integrate MeSH terms: MLGC, Model 1, and Model 2. The MLGC model is among the earliest to formally integrate MeSH into retrieval models. The model estimates the relevance model by exploiting the MeSH terms to update original textual query models, but with a different estimation process from ours. The performance of the MLGC model is better than the QLH and worse than the RM3 as can be seen in Tables 5 and 6. In Table 6, it is shown that almost all the performance measures of our models are higher than those of the MLGC model in both collections except for P@5 of Model 2 in the Ohsumed. In the Ohsumed collection, the results of Model 1 are significantly improved over the MLGC model in all measures, and Model 2 is significantly better than the MLGC model in terms of MAP. As for the Genomics collection, Model 1 shows significant improvements with respect to the MLGC in all metrics. Model 2 has improvements over the MLGC, but only P@5 is significantly improved in Model 2. In addition, in Fig. 4, we



(a) Ohsumed **(b)** Genomics

Fig. 4 Precision-recall curves for Ohsumed and Genomics

can observe that nearly all the precision-recall curves of our models in both collections are higher than the curves of the MLGC. The curve of the MLGC is between RM3 and QLH in the Ohsumed collection (Fig. 4), which indicates the performance of the MLGC is between that of the RM3 and the QLH. In the Genomics, the curve of the MLGC is always below that of the RM3 and is below the curve of the QLH when the recall position is beyond 0.3.

In summary, among the performance of the three models that use MeSH, our models outperform the MLGC model in both collections. The use of MeSH in information retrieval has been discussed for decades (Srinivasan 1996; Shiri 2012). Theoretically, MeSH descriptors can standardize the terms from authors and users, and help to achieve concept level matching. However, empirically, not every attempt has been successful (Bacchin and Melucci 2005; Hersh 2008). The MLGC model is one of the early attempts to formally integrate MeSH into retrieval models and achieved improved performance. It appears that our models further improve the effectiveness of MeSH-enhanced retrieval models.

6.2.3 Robustness of the MeSH-enhanced retrieval models

This section investigates the robustness of the new models. A robust retrieval model is expected to have positive impact on most queries (Wang et al. 2012). To examine the robustness of the MeSH-enhanced retrieval models, we calculate how many queries have improved or decreased average precision in our models when compared with RM3.

Figures 5 and 6 summarize the statistical information of the number of queries that have improved/decreased performance in the MeSH-enhanced models compared with RM3 in the Ohsumed and Genomics respectively. The x-axis in these figures shows how much increase/decrease in the average precision of a specific query in the MeSH-enhanced retrieval models when compared with RM3. Those bars to the left of [0, 25 %] represent the queries that have lower performance in the MeSH-enhanced models than in RM3, and the bars to the right (including [0, 25 %]) show the queries whose performance is improved by the MeSH-enhanced retrieval models. According to Figs. 5 and 6, for both Model 1 and Model 2, more queries have seen an increase rather than a decrease in average precision when compared with RM3. Table 7 lists the number of queries with improved/worsened performance in the MeSH-enhanced models compared with RM3. In the Ohsumed, Model 1 has 61 queries with AP increase, and 44 with AP decrease over RM3. Model 2 has 68 queries with AP increase, and 37 with decrease. In the Genomics, Model 1 has 18 queries with AP increase, and 8 with decrease. Model 2 has 19 queries with AP increase, and 7 with decrease over RM3. Of all the queries in both collections, there are about a third or

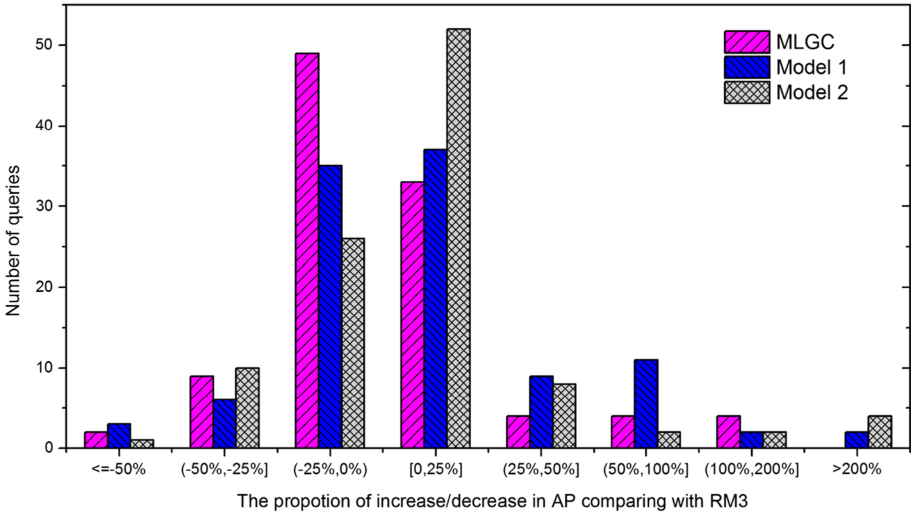


Fig. 5 Statistics of increased/decreased queries of MeSH-enhanced retrieval models compared with RM3 in Ohsumed

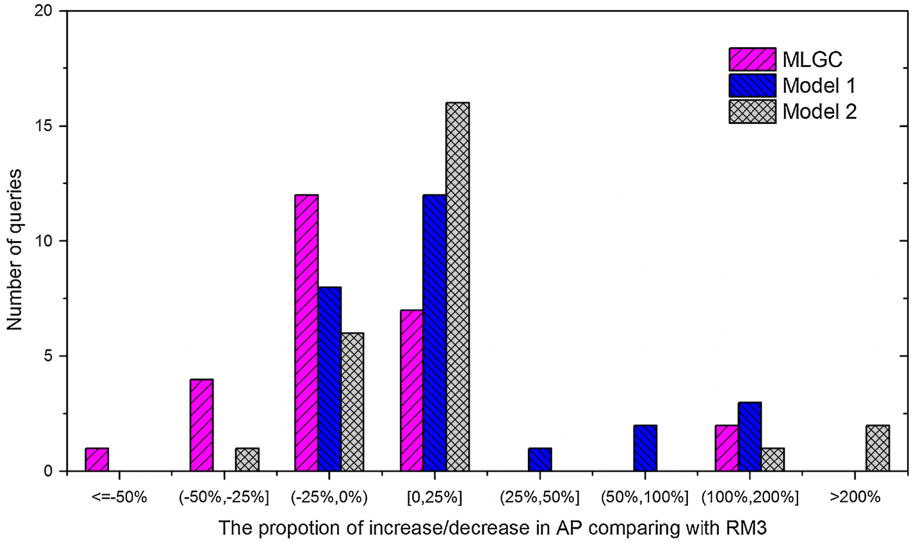


Fig. 6 Statistics of increased/decreased queries of MeSH-enhanced retrieval models compared with RM3 in Genomics

more with a performance improvement by [0, 25 %], and the performance of some queries are improved by more than 100 %. Compared with the state-of-the-art model, the RM3, the results show that our models are more robust and can overall improve the average precision.

Then, we compare our models with the MLGC model using RM3 as a reference. As Table 7 shows, the number of queries whose performance worsens in the MLGC model is

Table 7 The number of increased/decreased queries compared with RM3

Model	# of queries with AP increase		# of queries with AP decrease	
	Ohsumed	Genomics	Ohsumed	Genomics
Model 1	61	18	44	8
Model 2	68	19	37	7
MLGC	45	9	60	17

greater than the number of queries whose performance is improved when compared with RM3. This is consistent with the previous observations that MLGC does not outperform RM3. A total number of 60 queries have lower average precision in MLGC than in RM3, and only 45 are improved over RM3 in the Ohsumed. In addition, 17 queries have lower average precision in MLGC than in RM3, and only 9 queries are improved in the Genomics. It seems that our models improve the performance of more queries and worsen the performance of fewer queries than the MLGC model does. Therefore, it can be said that our models are more robust than the MLGC model.

It is also noted that although Model 2 improves the performance of more queries than Model 1 does (68 vs. 61 in Ohsumed, 19 vs. 18 in Genomics), the level of improvements in Model 1 is greater than that in Model 2. In the Ohsumed, Model 1 improves 24 queries with an increase of more than 25 %, and Model 2 only improves 16 queries with the same level of increase. In the Genomics, Model 1 improves 6 queries with an increase of more than 25 %, and Model 2 has 3 queries with the same level of improvements. On average, the performance of Model 1 is a bit better than that of Model 2 as is shown in Tables 5 and 6.

6.3 Cross validation for different models

The previous section compares the performance of different models under the optimal parameter settings. This provides evidence for the upper bound effectiveness of the retrieval models. However, it is not clear how the results will generalize to new topics or unseen topics in practice. To examine their effectiveness on unseen topics, we carried out ten-fold cross-validation for the RM3 and the three MeSH-enhanced models. The cross-validation method divides the test topics into 10 equal size subsets (in our case the subsets are roughly equal size because the total number of test topics is not divisible by 10). In each run, it uses one subset of topics for testing and trains the parameters on the rest nine subsets. It repeats 10 times with each of 10 subsets as the testing set, and then averages the performance on the testing sets from the 10 runs. The cross-validation method will be able to evaluate the performance of the retrieval models on unseen topics since the parameters are not trained on the test topics. We used the same method as introduced in Sect. 5.3 to tune the parameters on the training set, and then evaluated the retrieval performance on the testing set in each run. The cross-validation was not applied to the QLH as no parameter from this model was tuned in this study.

Table 8 lists the results from the cross-validation for different models. In the Ohsumed collection, Model 1 shows significant improvements over the QLH, the RM3, and the MLGC in all metrics. The performance of Model 2 is significantly better than that of the QLH in all metrics. The MAP and P@10 of Model 2 are greater than those of the RM3 and the MLGC, but only the MAP improvement over the MLGC is statistically significant. Model 2 has a lower value in P@5 than the RM3 and the MLGC, but the differences are

Table 8 Cross-validation results of different models

Collection	Metrics	QLH	RM3	MLGC	Model 1	Model 2
Ohsumed	MAP	0.2487	0.3055	0.2828	**0.3253**$^{*,+,\dagger}$	0.3101*†
	P@5	0.4095	0.4590	0.4610	**0.5105**$^{*,+,\dagger}$	0.4495*
	P@10	0.3657	0.4286	0.4143	**0.4638**$^{*,+,\dagger}$	0.4381*
Genomics	MAP	0.3527	0.3634	0.3209	**0.4177**$^{*,+,\dagger}$	0.3854$^{+\dagger}$
	P@5	0.5385	0.5462	0.5077	**0.6154**$^{*,+,\dagger}$	0.5692†
	P@10	0.4808	0.4808	0.4462	**0.5077**†	0.5000†

Some values for P@5 and P@10 in Table 8 are greater than the corresponding ones reported in Sect. 6.3. This is because the parameter settings in Sect. 6.3 is optimized according to MAP and may not be optimal for P@5 and P@10

Bold numbers are the best performance for each metrics

* means statistically significant differences from the query likelihood model with a two-tailed randomization test at 0.05 level

$^{+}$ means statistically significant differences from the relevance model with a two-tailed randomization test at 0.05 level

† means statistically significant differences from the MLGC model with a two-tailed randomization test at 0.05 level

not significant. In the Genomics collection, Model 1 has significant improvements over the QLH, the RM3, and the MLGC in terms of MAP and P@5. The P@10 of Model 1 is also greater than those of the three baseline models, but only the improvement over the MLGC is statistically significant. In terms of Model 2, it has significant improvements over the MLGC in terms of all metrics. The performance of Model 2 is higher than that of the QLH, although the differences are not significant. When compared with the RM3, the MAP of Model 2 is significantly better.

In general, the cross-validation results are consistent with the results in the previous section and confirm the superior performance of our models. Model 1 significantly improves the performance over the QLH, the RM3, and the MLGC. Model 2 is generally better than the QLH and the RM3, and significantly better than the MLGC. Model 1 also shows slight advantages over Model 2 in cross-validation, which is also consistent with previous results.

6.4 The influence of parameters

Previous sections compare the performance of the retrieval models under the optimal parameter settings and in cross-validation tests. In our models, some controlling parameters can be adjusted when estimating the relevance model (listed in Table 2). This section focuses on how these parameters impact the effectiveness of the retrieval models.

6.4.1 The influence of the number of pseudo-feedback documents

Figure 7 shows the performance of our proposed models in terms of MAP when the number of pseudo-feedback documents varies from 1 to 10. The performance of our models increases as the parameter N increases until reaching their peaks, and then the performance goes down slightly. This pattern holds for both collections. In the Ohsumed collection, Model 1 achieved its best performance when $N = 6$, and Model 2 reached its

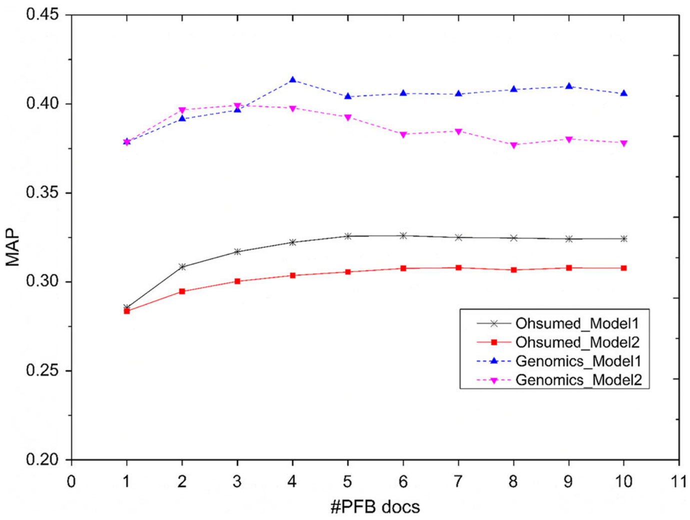


Fig. 7 Performance of different number of pseudo-feedback documents (*N*) (other parameters are kept uniform in different retrieval runs)

best when $N = 7$. In the Genomics, the best results was obtained when $N = 4$ for Model 1 and $N = 3$ for Model 2.

In relevance models, pseudo-feedback documents are often used to approximate a true relevance model. If a relevance model is generated from all relevant documents, it is supposed to achieve optimal performance under the assumption of relevance models (Lavrenko and Croft 2001). However, in reality, pseudo-feedback documents may include irrelevant documents, which may lead to a suboptimal estimation of the relevance model. When only 1 pseudo-feedback document was used, neither Model 1 nor Model 2 performed well in the test collections. This may be due to the fact that if the top ranked document happens to be irrelevant, the estimation of relevance model can be inaccurate. In the Genomics, for example, the average precision of topic 160 in Model 1 is 0.1976 when only one feedback document is used, as it happens to be an irrelevant one. While the AP climbs to 0.6844 when three pseudo-feedback documents are used, and among the three the other two documents are relevant. On the other hand, when too many pseudo-feedback documents were used, irrelevant pseudo-feedback documents can also be included in the estimation process along with relevant ones, which might worsen the performance. Therefore, similar to relevance models, the performance of the ME models varies depending on the number of pseudo-feedback documents. The best practice is to find a balanced number of pseudo-feedback documents, neither too few nor too many. This parameter may be influenced by the number of relevant documents in the collection and the performance of the first retrieval pass. According to Fig. 7, the optimal number of pseudo-feedback documents also depends on test collections and the retrieval models. This is consistent with the discussion on pseudo relevance feedback technique in the literature (Lv and Zhai 2009; Montgomery et al. 2004).

6.4.2 The influence of the number of terms in the generative concept model

Generative concept models capture the associations between concepts and terms. As Eq. (2) shows, document language models in Model 1 are affected by document-concept associations (the conceptual document model, $p(c|d)$) and concept-term associations (the generative concept model, $p(w|c)$). In Model 1, terms in the relevance model originate from the generative concept model. Thus, the number of terms for generative concept models (the parameter $|V_c|$) may impact the probabilities of terms in the final estimated relevance model, and in turn influence the retrieval performance. The results are shown in Fig. 8. In the Ohsumed, the best performance was obtained when $|V_c|$ was set to 70. After $|V_c|$ reached 70, the value of MAP declined gradually. In the Genomics, the optimal performance was obtained when $|V_c|$ was set to 250. It seems that the optimal setting of this parameter is dependent on the collection. A large value for this parameter may introduce noisy terms, while, a small value may be inadequate and miss some important terms in the relevance model. Both situations may cause a negative impact that leads the relevance model to drift away from the true relevance model.

6.4.3 The impact of the expanded document model

In Model 1, the parameter λ_{m1} controls the proportion of the document language model estimated through the concept layer in the final document language model. When λ_{m1} is set to 0, only the original document language model is used. When λ_{m1} is set to 1, only the document language model estimated through the concept layer is used. Examining the impact of λ_{m1} helps to understand how the concept layer contributes to the performance.

According to Fig. 9, the best performance in the Ohsumed was obtained when only the document language model estimated from the generative concept models was used. In the Genomics, the best performance was achieved when λ_{m1} equals 0.5. It is also noted that the performance of using the document model estimated through the concept layer alone is

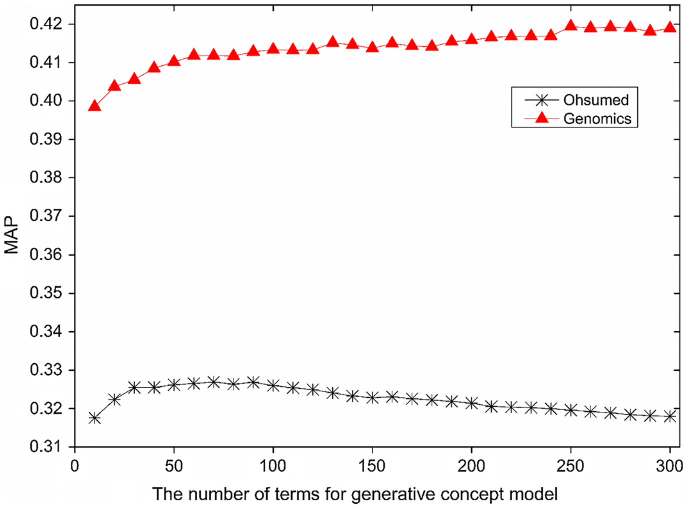


Fig. 8 The results of using different number of terms in the generative concept model ($|V_c|$). (Experiment runs were conducted when the number of pseudo-feedback documents was optimal.)

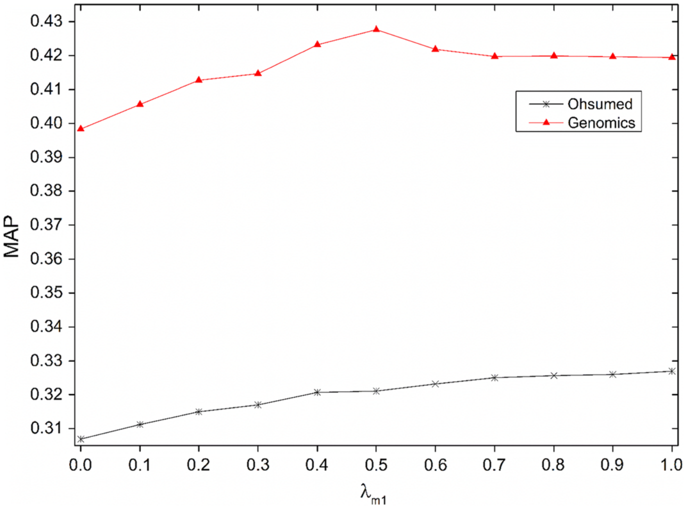


Fig. 9 Performance of experiments when the parameter λ_{m1} varied from 0 to 1.0 (The parameter N and the parameter $|V_c|$ were optimal.)

better than the performance of using the original document language model in both collections. This provides evidence that expanding the original document model with the document language model estimated through the concept layer can improve the retrieval performance. The results confirm our hypothesis that the document models estimated through the concept layer (or MeSH-enhanced models) do help to improve retrieval performance.

6.4.4 The influence of refining concepts

The hypothesis of Model 2 is that a user query can be represented as a set of relevant concepts. In the process of subject indexing, documents are often assigned with some general subject terms. For example, a large number of documents in the test collections are assigned with “Human”, “Child”, and “Animal”. These general concepts may not help to distinguish the relevant documents from irrelevant ones since most of the search topics are much more specific. Selecting the most important concepts to represent user queries is imperative for Model 2. Figure 10 lists the MAP values for three groups of experimental runs in each collection where the number of MeSH terms to represent user queries (the parameter $|c|$*)* varied from 5 to 50 with a step of 5. The number of pseudo-feedback documents (the parameter N) in each group was 4, 7, and 10 respectively. According to Fig. 10, the best performance in the Ohsumed for each group was obtained when $|c| = 15$ for $N = 4$, $|c| = 25$ for $N = 7$ and $N = 10$. While the optimal performance in the Genomics for each group was obtained when $|c| = 35$ for $N = 4$, $|c| = 30$ for $N = 7$, and $|c| = 5$ for $N = 10$.

It appears that the optimal value of parameter $|c|$ varies depending on the number of pseudo feedback documents N and the test collection. According to our experiments, setting $|c|$ in the range from 10 to 30 seems to produce reasonable results.

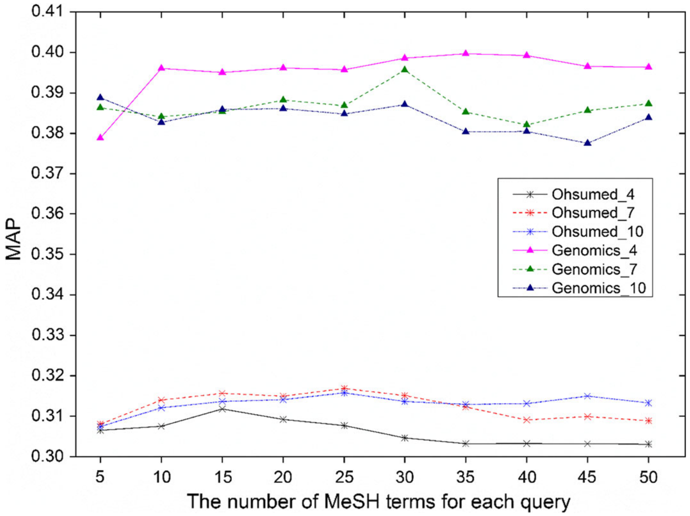


Fig. 10 Performance of using different number of MeSH terms (the parameter |c|) for each query in the Model 2 (experimental runs using 4, 7 and 10 pseudo-feedback documents respectively)

6.5 Case analysis

To understand how the new ME models work, we present some query cases from the Ohsumed collection with improved/worsened performance. Similar cases can be also found from the Genomics collection. The optimal parameter settings are used for the results reported in this section (Table 9).

Case 1 Query 13 in the Ohsumed improved in Model 1—Relevant terms are further emphasized through the concept layer.

Query 13 in the Ohsumed is "*lactase deficiency therapy options*". For this query, Model 1 achieved an 81.43 % improvement in average precision over the RM3 (0.5878 v.s. 0.3240). When examining the terms in the relevance models estimated from Model 1 and the RM3 (Table 10), Model 1 assigns noticeably higher weights to three topically relevant terms, "*milk*", "*galactosidase*", and "*yogurt*", which are not in the top 10 terms in the relevance model from the RM3. A further content analysis on the top 5 retrieved documents from the two models suggests that in the top 5 documents retrieved by Model 1 (all are relevant), at least one of the three terms occurs; In the top 5 documents retrieved by the RM3, where two are relevant and three are irrelevant, the three terms only occur in the relevant documents and not in the irrelevant ones. It is these three terms that bring more relevant documents to the top of the result list and improve the performance of Model 1.

Table 9 The optimal parameter settings for all models

Collection	RM3	MLGC		Model 1			Model 2	
	N	N	$\|c\|$	N	$\|V_c\|$	λ_{m1}	N	$\|c\|$
Ohsumed	10	10	4	6	70	1.0	7	25
Genomics	10	3	7	4	250	0.5	4	35

Table 10 Top 10 terms in the relevance models of the query 13

RM3		Model 1	
Lactase	0.0940	Lactase	0.0845
Deficient	0.0912	Deficient	0.0766
Therapy	0.0806	Therapy	0.0723
Option	0.0798	Option	0.0714
Lactose	0.0105	Lactose	0.0397
Patient	0.0071	Milk	0.0192
Subject	0.0071	Galactosidase	0.0167
Human	0.0058	Yogurt	0.0142
Study	0.0051	Bone	0.0103
Calcium	0.0048	Beta	0.0088

The underline words indicate higher weights are assigned to the three relevant terms by Model 1 than RM3

To understand how Model 1 uncovers the three relevant query terms, we examined the concept-term associations and the document-concept associations. The three relevant terms are found to be strongly associated with a number of major concepts in the pseudo feedback documents (Table 11). In fact, the term "*milk*" is associated with eighteen concepts in the pseudo feedback documents, the term "*galactosidase*" is associated with twelve, and the term "*yogurt*" is associated with ten. The concept layer of Model 1 helps to place additional emphasis on the three relevant terms and boost their weights in the relevance model. In this case, the additional concept layer in Model 1 contributes to the retrieval performance.

Case 2 Query 53 in the Ohsumed worsened in Model 1- Performance decrease is caused by "Topic Drift".

In the Ohsumed, the performance of query 53 in Model 1 is lower than that in the RM3 (AP: 0.2894 v.s. 0.3997). The query is "*lupus nephritis, diagnosis and management*". A comparison of the relevance models estimated from the two models indicates that the term *"erythematosus"* is in the top 10 ranks in Model 1 but not in the RM3. Another term "*sle*"

Table 11 Some examples of the associations between the three relevant terms and major MeSH descriptors in the pseudo feedback documents

Query term	ConceptId	MeSH Descriptors	Document	$P(c\mid d)$
Milk	11885	Milk/*ME	89132386	0.1034
Galactosidase	68277	Galactosidases/*BL	87103870	0.1295
	62433	beta-Galactosidase/*BL	87103870	0.0914
	68277	Galactosidases/*DF	89132386	0.1521
	62433	beta-Galactosidase/*DF	89132386	0.1046
	68277	Galactosidases/*ME	89244632	0.0860
	62433	beta-Galactosidase/AN/*ME	89244632	0.0560
	62433	beta-Galactosidase/BI/*DF	91009843	0.2785
	62433	beta-Galactosidase/*DF	91096782	0.1481
Yogurt	146457	Yogurt/*	89244632	0.2536

has a much higher probability in Model 1 than in the RM3. A content analysis on the relevant and the irrelevant documents in the top 20 retrieved documents from the two models suggests that these two query terms lead to higher rankings of the irrelevant documents in Model 1 than in the RM3. For example, the irrelevant document 90203598, ranked in the 8th position by the RM3, is placed in the 3rd position by Model 1. Another irrelevant document 90204393 is at the 16th position in the RM3 and rises to the 7th in Model 1. Both documents mention "*erythematosus*" and "*sle*" a number of times.

The ranks of the two terms in the relevance model are boosted due to their strong associations with a number of MeSH concepts in the pseudo feedback documents, such as "*Lupus Nephritis*" and "*Lupus Erythematosus, Systemic*". According to the MeSH thesaurus, "*systemic lupus erythematosus"* is an entry term of the concept "*Lupus Erythematosus, Systemic*" which is a broader concept of the query topic "*Lupus Nephritis*". Therefore, the emphasis on "*erythematosus*" and "*sle*" drifts the query to a more general level concept, which leads a lower precision. This is actually a typical "Topic Drift" problem that is well noted in query expansion (Harman and Buckley 2009).

Case 3 Query 12 in the Ohsumed improved in Model 2—Relevant documents are re-ranked higher in pseudo feedback documents.

Model 2 improved the average precision of query 12 over the RM3 in the Ohsumed by 313 % (0.3000 vs. 0.0726). The query is "*descriptions of injuries associated with cult activities*". Model 2 attempts to improve the retrieval performance by re-ranking the pseudo feedback documents according to additional information from the concept layer. Table 12 lists the ranks of pseudo feedback documents in the RM3 (which is essentially the QLH), as well as the ranks and the probabilities of the re-ranked pseudo feedback documents in Model 2. It can be found that Model 2 ranks the relevant document (90225113) higher than the RM3 does. Specifically, Model 2 identifies 25 relevant concepts for query 12 according to the proposed method [Eq. (4)]. Among them, seven MeSH concepts appear in the relevant document (90225113) and are strongly associated with the document. This is why the rank of this relevant document is boosted in Model 2. In addition, it is notable that the differences of the probabilities between the relevant document and irrelevant documents in Model 2 are much larger than the differences of KL divergence values in the QLH. This suggests that Model 2 further separates the relevant document from the irrelevant ones in the pseudo feedback documents through the concept layer. This relevant document in turn brings more relevant query terms in the final relevance model, and helps to improve the retrieval performance.

Table 12 The ranks of pseudo feedback documents for query 12

QLH		Model 2	
DOCID	KL_DIV	DOCID	P(d|R)
91140322	−5.4423	90225113	0.7548
91289992	−5.4727	89257902	0.0968
90225113	−5.5381	90072026	0.0645
90072026	−5.6055	91140322	0.0516
90250337	−5.6232	91289992	0.0129
87280708	−5.6265	87280708	0.0129
89257902	−5.6578	90250337	0.0065

The underlined documents are relevant

Case 4 Query 67 in the Ohsumed worsened in the Model 2—Irrelevant documents are re-ranked higher in pseudo feedback documents.

Model 2 does not always re-rank relevant documents higher in the pseudo feedback document. For example, query 67 has a lower average precision in Model 2 than in the RM3 (0.1865 v.s. 0.3047). The query is "*outpatient management of diabetes, standard management of diabetics and any new management technique*". In this case, the irrelevant documents are ranked higher in the pseudo feedback documents by Model 2. Further investigation reveals that the higher rankings of the irrelevant documents in the pseudo feedback documents are due to their strong associations with the relevant concepts identified by Model 2. In this case, the relevant concepts identified by Model 2 occur more frequently in the irrelevant documents than in the relevant ones in the pseudo feedback documents. The highly ranked irrelevant documents promote their terms in the relevance model. Accordingly, the estimated relevance model of Model 2 become less accurate and the retrieval performance worsens.

7 Conclusion and future work

This study proposed two new MeSH-enhanced retrieval models for health information retrieval by integrating the controlled vocabulary MeSH into the retrieval process. The MeSH terms become a conceptual representation in our models. The two ME models reconstruct the relevance model by employing the generative concept model and the conceptual document model. The generative concept model is formulated by mining the concept-term associations, while the conceptual document model is constructed by inferring the associations between documents and the assigned concepts. Model 1 enriches the document language models of the pseudo-feedback documents using the generative concept models of the assigned concepts. Model 2 re-ranks the pseudo-feedback documents according to the relevant concepts that users are looking for. The document-concept and the concept-term associations are used to reweight the pseudo-feedback documents. Experiments on two test collections in the health domain suggest that our MeSH-enhanced models can further improve retrieval performance over the two state-of-the-art retrieval models, the query likelihood (QLH) model and the RM3 model, which do not use MeSH, as well as one similar model that incorporates MeSH. Model 1 also showed slight advantages over Model 2.

Comparing ME models with the query likelihood model and the RM3 model, our experimental results indicated that ME models significantly improved the performance over the QLH in terms of MAP. To compare with the relevance model, we selected a strong baseline, RM3, and tuned the parameters to achieve optimal performance. It is observed that the performance of Model 1 is significantly higher than that of the RM3 in terms of MAP and P@5 in both collections. Model 2 also show significant improvements over the RM3 in terms of MAP in the Ohsumed collection and in terms of P@5 in the Genomics collection. The 11-point precision and recall charts suggest that our models have advantages over the QLH and the RM3 cross the different recall values. In addition, the robustness analysis also shows that our models are more robust than the RM3 model.

A comparison of our models with an earlier MeSH-enhance model, MLGC, shows that the new ME models improved the retrieval performance significantly over the MLGC and are more robust than the MLGC in both collections. In the MLGC, the query model is generated by combining the conceptual model and the generative concept model. The

MLGC model uses a two-layer structure to estimate the query model (i.e. concepts and terms) and does not formally incorporate the document layer which is an essential and fundamental instrument of the original relevance model. For this reason, it might not be optimal to directly use concept as a pivot language between the relevance model and terms. On the other hand, our models employ a three-layer structure with documents, concepts, and terms. Empirical results from the experiments showed the improvements in our models compared to MLGC.

The cross-validation results further confirm the superiority of our models on unseen topics to the selected baseline models, and thus enhancing the generalizability of the findings.

In summary, in terms of the first research question we proposed, the performance of ME models are superior to the baseline models, including a popular retrieval baseline (QLH), a strong baseline (RM3), and an earlier MeSH-enhanced retrieval model.

In term of the influence of parameters in the models, pseudo-feedback documents play an important role in estimating the relevance model. Similar to relevance models, the performance of the ME models varies according to the number of pseudo-feedback documents. The optimal number of pseudo relevant documents depends on the retrieval models and collections. In addition, choosing an appropriate number of terms for generative concept models is important for Model 1. The generative concept model is found to be an effective smoothing instrument to enhance the document model of pseudo-feedback documents and leads to improved results. In Model 2, the concept selection helps to refine the relevance model. In practice, the optimal parameter settings should be tuned according to the above evidence and the characteristics of the data collection.

A detailed case analysis provides further insight into how and why the new models improve/worsen retrieval performance. In general, the findings support our hypotheses. Model 1 improves the retrieval performance by enriching the document language models of the pseudo feedback documents through the additional concept layer. Model 2 uses the concept layer to re-rank the pseudo feedback documents and improves the accuracy of the relevance model. However, there are cases where Model 1 leads to topic drift by emphasizing related concepts, and Model 2 re-ranks the irrelevant documents higher in the pseudo feedback documents due to the inaccuracy in identifying relevant concepts.

Some limitations of the study need to be noted. First, we carried out our experiments on two standard test collections. This is due to the availability of the data at the time. More and larger data collections can be used to further validate the results. Second, the influence of parameters is analyzed and reported. However, the underlying mechanisms of the impact are not fully understood. Further investigation is still needed to explain the relationships among the involved factors. Third, as is mentioned in the paper, the accurate estimation of document-concept and concept-term associations is crucial for the ME models. This study provides one way to estimate the probabilities. Case study has revealed that the method is not perfect and further discussions on the estimation methods will enhance the ME models.

Controlled vocabularies, such as MeSH, are invented to bridge the gap of vocabulary problem in IR. How to effectively use MeSH in health information retrieval is still an ongoing research question. The ME models proposed in this study use the associations between documents and concepts, and concepts and terms to incorporate MeSH into the retrieval process. Empirical results suggest the superiority of the ME models to the selected baseline models. The influence of parameters is discussed and a detailed case analysis is presented. The findings of this study contribute to the effective use of MeSH in health information retrieval and improve our understanding on this issue. Future work will further optimize the MeSH-enhanced models and apply these models to other domains and environments. First, different domains and digital collections have adopted different controlled vocabularies. The ME models proposed in this study can be easily generalized to a different domain. However, given the different features of different controlled vocabularies, domains, and collections, the parameters of the models may need adjustments accordingly. Future work will investigate the applicability of our methods in different environments and how to modify the models to account for the variations. Second, in this study, we used the TF-IDF method to mine the concept-term associations and the weighted mutual information to infer the associations between documents and assigned concepts, but, some other methods are also available for this purpose. More sophisticated natural language processing techniques may also be adopted to enhance the bag-of-words representations. A future study will compare different methods to infer the associations and investigate how that impacts the performance. Moreover, this study did not differentiate the roles of MeSH terms in the literature, such as main headings and qualifiers. Different qualifiers of the same main heading are not distinguished. Investigating the roles of MeSH terms and how they impact the retrieval models will be addressed in our future research.

Acknowledgments We thank the anonymous reviewers for their constructive comments on an earlier version of the manuscript. Their comments definitely help to improve the paper. We are also grateful for helpful comments from Professor Wei Lu and Associate Professor Lu An. This work was sponsored by the National Natural Science Foundation of China funded projects under grant No. 71273196, No. 71420107026, and No. 71373286.

References

Abdou, S., Ruck, P., & Savoy, J. (2005). Evaluation of stemming, query expansion and manual indexing approaches for the genomic task. In *Proceedings of TREC 2005*.

Bacchin, M., & Melucci, M. (2005). Symbol-based query expansion experiments at TREC 2005 Genomics track. In *Proceedings of TREC 2005*.

Bai, J., Song, D., Bruza, P., Nie, J. Y., & Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In *Proceedings of CIKM 2005* (pp. 688–695). Bremen: ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Díaz-Galiano, M. C., García-Cumbreras, M. A., Martín-Valdivia, M. T., Montejo-Ráez, A., & Urena-López, L. A. (2008). Integrating mesh ontology to improve medical information retrieval. In *Advances in multilingual and multimodal information retrieval* (pp. 601–606). Berlin: Springer.

Fang, H., & Zhai, C. (2005). An exploration of axiomatic approaches to information retrieval. In *Proceedings of SIGIR 2005* (pp. 480–487). Salvador: ACM.

Finkelstein, L. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems, 20*, 116–131.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artifical intelligence* (pp. 1606–1611). Morgan Kaufmann Publishers Inc.

Gauch, S., & Smith, J. B. (1991). Search improvement via automatic query reformulation. *ACM Transactions on Information Systems (TOIS), 9*(3), 249–280.

Gault, L. V., Shultz, M., & Davies, K. J. (2002). Variations in Medical Subject Headings (MeSH) mapping: From the natural language of patron terms to the controlled vocabulary of mapped lists. *Journal of the Medical Library Association, 90*(2), 173.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). *Indexing with WordNet synsets can improve text retrieval*. arXiv preprint cmp-lg/9808002.

Griffon, N., Chebil, W., Rollin, L., Kerdelhue, G., Thirion, B., Gehanno, J. F., & Darmoni, S. J. (2012). Performance evaluation of unified medical language system®'s synonyms expansion to query PubMed. *BMC Medical Informatics and Decision Making, 12*(1), 12.

Guisado-Gámez, J., Dominguez-Sal, D., & Larriba-Pey, J. L. (2013). *Massive query expansion by exploiting graph knowledge bases*. arXiv preprint arXiv:1310.5698.

Guo, Y., Harkema, H., & Gaizauskas, R. (2004). Sheffield university and the TREC 2004 Genomics track: Query expansion using synonymous terms. In *Proceedings of the thirteenth Text REtrieval conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Harman, D., & Buckley, C. (2009). Overview of the reliable information access workshop. *Information Retrieval, 12*(6), 615–641.

He, B., & Ounis, I. (2009). Finding good feedback documents. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 2011–2014). Hong Kong: ACM.

Hersh, W. (2008). *Information retrieval: A health and biomedical perspective*. Berlin: Springer.

Hersh, W., & Bhupatiraju, R. T. (2003). TREC Genomics track overview. In *Proceedings of the twelfth text retrieval conference, TREC 2003* (pp. 14–23). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Bhupatiraju, R. T., & Price, S. (2003). Phrases, boosting, and query expansion using external knowledge resources for genomic information retrieval. In *Proceedings of the twelfth text retrieval conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Buckley, C., Leone, T. J., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of SIGIR 1994* (pp. 192–201). London: Springer.

Hersh, W. R., Cohen, A. M., Roberts, P. M., & Rekapalli, H. K. (2006). TREC 2006 Genomics track overview. In *TREC 2006*.

Jelinek, F., & Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the workshop on pattern recognition in practice*. Amsterdam: North-Holland.

Kamps, J. (2004). Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In *Advances in information retrieval* (pp. 283–295). Berlin: Springer.

Korfhage, R. R. (1984). Query enhancement by user profiles. In *Proceedings of SIGIR 1984* (pp. 111–121). Cambridge: British Computer Society.

Kurland, O. (2008). The opposite of smoothing: A language model approach to ranking query-specific document clusters. In *Proceedings of SIGIR 2008* (pp. 171–178). Singapore: ACM.

Kurland, O. (2009). Re-ranking search results using language models of query-specific clusters. *Information Retrieval, 12*(4), 437–460.

Kurland, O., & Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of SIGIR 2004* (pp. 194–201). Sheffield: ACM.

Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR 2001* (pp. 111–119). New Orleans: ACM.

Lafferty, J., & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In *Language modeling for information retrieval* (pp. 1–10). Netherlands: Springer.

Lavrenko, V., & Croft, W. B. (2001). Relevance based language models. In *Proceedings of SIGIR 2001* (pp. 120–127). New Orleans: ACM.

Lee, K. S., Croft, W. B., & Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of SIGIR 2008* (pp. 235–242). Singapore: ACM.

Liu, X., & Croft, W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of SIGIR 2004* (pp. 186–193). Sheffield: ACM.

Lu, Z., Kim, W., & Wilbur, W. J. (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval, 12*(1), 69–80.

Lu, K., & Mao, J. (2013). Automatically infer subject terms and documents associations through text mining. In *Proceedings of the 76th annual conference of association for information science and technology* (ASIST'2013), Montreal, Canada.

Lv, Y., & Zhai, C. (2009). A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of CIKM 2009* (pp. 1895–1898). Hong Kong: ACM.

Lv, Y., Zhai, C., & Chen, W. (2011). A boosting approach to improving pseudo-relevance feedback. In *Proceedings of SIGIR 2011* (pp. 165–174). Beijing: ACM.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Mata, J., Crespo, M., & Maña, M. J. (2012). Using MeSH to expand queries in medical image retrieval. In *Medical content-based retrieval for clinical decision support* (pp. 36–46). Berlin: Springer.

Meij, E., & De Rijke, M. (2007). Integrating conceptual knowledge into relevance models: A model and estimation method. In *International conference on the theory of information retrieval (ICTIR 2007)*. Budapest: Alma Mater Series.

Meij, E., Trieschnigg, D., De Rijke, M., & Kraaij, W. (2010). Conceptual language models for domain-specific retrieval. *Information Processing and Management, 46*(4), 448–469.

Metzler, D., & Croft, W. B. (2005). A Markov random field model for term dependencies. In *Proceedings of SIGIR 2005* (pp. 472–479). Salvador: ACM.

Metzler, D., Dumais, S., & Meek, C. (2007). Similarity measures for short segments of text. In *Advances in information retrieval* (pp. 16–27). Berlin: Springer.

Montgomery, J., Si, L., Callan, J., & Evans, D. (2004). Effect of varying number of documents in blind feedback: Analysis of the 2003 NRRC RIA workshop "bf_numdocs" experiment suite. In *Proceedings of SIGIR 2004* (pp. 476–477). Sheffield: ACM.

Plaunt, C., & Norgard, B. A. (1998). An association-based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science, 49*(10), 888–902.

Poikonen, T., & Vakkari, P. (2009). Lay persons' and professionals' nutrition-related vocabularies and their matching to a general and a specific thesaurus. *Journal of Information Science, 35*(2), 232–243.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR 1998* (pp. 275–281). Melbourne: ACM.

Shin, K., & Han, S. Y. (2004). Improving information retrieval in MEDLINE by modulating MeSH term weights. In *Natural language processing and information systems* (pp. 388–394). Berlin: Springer.

Shiri, A. (2012). *Powering search: The role of Thesauri in new information environments*. Medford, NJ: Information Today Inc.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on information and knowledge management* (pp. 623–632). New York: ACM.

Srinivasan, P. (1996). Query expansion and MEDLINE. *Information Processing and Management, 32*(4), 431–443.

Stokes, N., Li, Y., Cavedon, L., & Zobel, J. (2009). Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval, 12*(1), 17–50.

Trieschnigg, D. (2010). *Proof of concept: Concept-based biomedical information retrieval*. Doctoral dissertation, University of Twente.

Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH up: Effective MeSH text classification for improved document retrieval. *Bioinformatics, 25*, 1412–1418.

van Rijsbergen, (1979). *Information retrieval* (2nd ed.). London: Butterworths.

Vechtomova, O., Robertson, S., & Jones, S. (2003). Query expansion with long-span collocates. *Information Retrieval, 6*(2), 251–273.

Voorhees, E. M. (1994). Query expansion using lexical–semantic relations. In *Proceedings of SIGIR 1994* (pp. 61–69). London: Springer.

Wang, L., Bennett, P. N., & Collins-Thompson, K. (2012). Robust ranking models via risk-sensitive optimization. In *Proceedings of SIGIR 2012* (pp. 761–770). Portland: ACM.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of SIGIR 2006* (pp. 178–185). Seattle: ACM.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of SIGIR 1996* (pp. 4–11). Zurich: ACM.

Zeng, Q. T., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., & Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association, 13*(1), 80–90.

Zeng, Q., Kogan, S., Ash, N., Greenes, R. A., & Boxwala, A. A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine, 41*(4), 289–298.

Zeng, Q. T., Kogan, S., Plovnick, R. M., Crowell, J., Lacroix, E. M., & Greenes, R. A. (2004). Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics, 73*(1), 45–55.

Zhai, C. (2002). *Risk minimization and language modeling in text retrieval*. Doctoral dissertation, University of Massachusetts, Amherst.

Zhai, C., & Lafferty, J. (2001a). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the SIGIR 2001* (pp. 334–342). New Orleans: ACM.

Zhai, C., & Lafferty, J. (2001b). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the CIKM 2001* (pp. 403–410). Atlanta: ACM.

Zhai, C., & Lafferty, J. (2002). Two-stage language models for information retrieval. In *Proceedings of the SIGIR 2002* (pp. 49–56). Tampere: ACM.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179–214.

Zhang, J., Wolfram, D., Wang, P., Hong, Y., & Gillis, R. (2008). Visualization of health-subject analysis based on query term co-occurrences. *Journal of the American Society for Information Science and Technology, 59*, 1933–1947.

Zielstorff, R. D. (2003). Controlled vocabularies for consumer health. *Journal of Biomedical Informatics, 36*, 326–333.