Leverage the Associations between Documents, Subject Headings and Terms to Enhance Retrieval

Jin Mao
Center for the Studies of Information Resources
Wuhan University
No.16 Luojiashan Road, Wuhan, Hubei, China, 430072
danveno@163.com

Kun Lu
School of Library and Information Studies
University of Oklahoma
401 West Brooks, Norman, Oklahoma, USA, 73019
kunlu@ou.edu

ABSTRACT
Literatures in medical domain are often annotated with subject headings by professionals to help information seeking via manifesting the subjects of documents, where subject headings serve as the pivot language between documents and users. Current information retrieval methods using subject headings have not fully exploited the potential of subject headings yet. Both positive and negative results have been reported. In this paper, we explored the three-layer structure of documents annotated with subject headings, including document layer, concept layer (i.e. subject headings) and term layer, and then we proposed a concept-enhanced relevance model. The document-concept associations are mined to generate conceptual representations for documents and the concept-term associations are quantified and used to represent concepts as language models. By embedding these associations, subject headings are applied to enrich the document models in the estimation process of relevance model[1]. The experiments carried out on two medical collections showed the improvements of our model by comparing with three state-of-the-art baselines. Therefore, if exploited appropriately, such manually curated annotations as subject headings can become an effective tool to enhance information retrieval.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Relevance feedback; Retrieval models

General Terms
Algorithms, Performance, Experimentation

Keywords
Relevance model; Concept; Retrieval models; Health information retrieval

1. INTRODUCTION
In the situation of health domain, literatures are often annotated with subject headings by professionals. For example, documents in the PubMed database are labeled with terms in Medical Subject Headings (MeSH). The subject headings are often taken from controlled vocabularies, which are constructed to bridge the terminology gap between information resources (i.e. documents) and information needs (i.e. queries). Essentially, subject headings play a role of the pivot language between documents and users. In practice, the subject headings that are assigned to the documents represent the subjects/topics of the documents, aiding users to access to information resources in the situation of information seeking. Considering this characteristic, how to fully exploit these human contributed subject headings in health information retrieval is still under discussion to date. The subject headings have been used to improve retrieval in a number of ways, such as query expansion[2], or terminology assistance for users[3]. In these methods, subject terms that are related to users' information needs are first identified, either automatically or manually by users. These related subject terms are then added to the original queries to relieve the vocabulary mismatch between user queries and documents. However, both positive and negative results have been reported from existing methods of using subject headings[4]. Similar manually curated annotations, such as author keywords and user tags, are often viewed as topic metadata that discloses the topics of texts. Some doubts also exist about the usefulness of topic metadata for retrieval[5].

Most of the former studies regarded subject headings as ordinary terms rather than conceptual representations. However, manually curated knowledge can be considered as conceptual representations of documents[6]. In this paper, we explore a different approach to formally integrating the subject headings into retrieval models in a principled way. The subject headings are considered as the explicit conceptual representations of the documents that they are assigned to and then are used to enhance the relevance model[1]. The documents annotated with subject headings have three layers, including document layer, concept layer (i.e. subject headings) and term layer. Correspondingly, the assumption of our proposed model is that a document is about several concepts, and these concepts are further elaborated by the terms. In our model, the associations between concepts and terms are mined to represent the concept layer in a language modeling manner, and then the concept layer is added to the process of estimating the document language models of the pseudo feedback documents to enrich the document language models with the concept level characteristics.

2. CONCEPT-ENHANCED RELEVANCE MODEL
The underlying assumption of the relevance model[1] is that there exists a relevance model for a given query, represented in a language model. The relevance model is often estimated from pseudo relevant documents. The base relevance model that our model extends from, RM3[7], is essentially a combination of the weighted document language models in the pseudo relevant feedback set. It is not difficult to infer that the document language models play a crucial role in the final relevance model. Our concept-enhanced relevance model(CERM) is to add a concept layer in the document language model estimation process to

ESAIR'14, November 07 2014, Shanghai, China

http://dx.doi.org/10.1145/2663712.2666195

capture the concept level characteristics. Document 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 estimation process as shown in Figure 1.

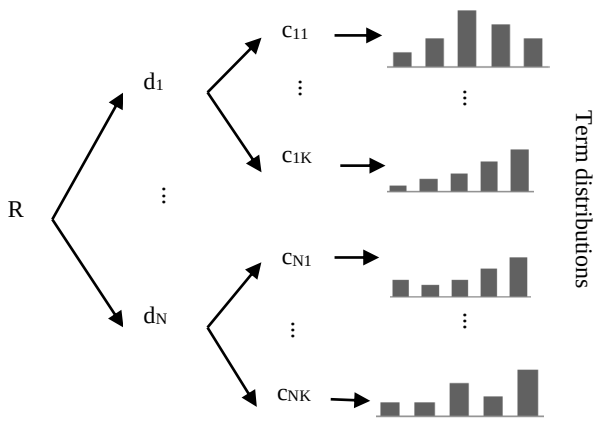

Figure 1. Generative process of concept-enhanced relevance model.

The generative process of our model 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.

In CERM, we employ a different approach for estimating $P(w|d)$. Instead of using MLE directly, we add a concept layer between the document model and the terms. As each of the documents assigned with a number of MeSH terms, 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 $P(w|d')$, shown in Equation (1).

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

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

In CERM, each concept is represented as a distribution of terms using a multinomial unigram language model. 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, as that each document can be represented with a multinomial distribution over concepts that are assigned to the document. The final estimation of the concept-enhanced relevance model can be obtained as:

$$P(w|R) \approx \sum_{d \in \Theta} (\lambda_{m1} \sum_{c \in \Gamma_d} P(w|c)P(c|d) + (1 - \lambda_{m1})P(w|d'))\, P(d|R) \quad (2)$$

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. MINING THE DOCUMENTS, CONCEPTS AND TERMS ASSOCIATIONS

In the final relevance model estimation (Equation 3), the generative concept model $P(w|c)$ and the conceptual document model $P(c|d)$ need to be estimated.

In our model, each concept is considered to be a probability distribution over terms in the unigram model, i.e., the generative concept model *P(w|c)*. The set of documents that are assigned with the concept is treated as a sub-collection for the concept. Terms from this sub-collection are used to represent 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:

$$\text{tfidf}_{w,c} = (0.5 + \sum_{d \in \Gamma_c} tf_{w,d}) * \log(\frac{N+0.5}{df_w+0.5}) \quad (3)$$

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 document frequency(the number of documents) containing the term *w*, and *N* is the number of documents in the entire corpus.

We obtain the generative concept model by normalizing all the term weightings of the concept:

$$P(w|c) = \frac{\text{tfidf}_{w,c}}{\sum_{w \in V} \text{tfidf}_{w,c}} \quad (4)$$

And also, in our model, we quantify the associations between documents and their assigned concepts, and regard a document as 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. 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[8].

4. EXPERIMENTS

Two standard IR test collections are used in the experiments: Ohsumed and TREC Genomics Track 2006. The language model toolkit, Lemur[1], was used to index the two collections. The Ohsumed collection was indexed by fields and the documents in the Genomics collection were indexed as a whole not by field. The Krovetz stemmer and the InQuery's standard stoplist with 418 stop words were used.

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.

Three baseline models are included in the study: query likelihood model[9], the base model of our model, *RM3*[7], and one similar model, MLGC[6]. Mean average precision and precision at top cutoffs (P@5 and P@10) are used to evaluate retrieval performance.

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

Table 1 lists the performance of the new concept-enhanced relevance model and three baseline models. According to Table 1, the performance of the RM3 is better than that of the QLH almost in terms of all evaluation measures in both collections. The performance of the MLGC is better than the QLH, but worse than the RM3 in both collections. In terms of the new models, it is observed that in the Ohsumed collection the CERM 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, the performance of the CERM is significantly better than that of the RM3 in terms of MAP and P@5. And the CERM improved the results significantly over the MLGC in all terms. In the Genomics collection, the CERM shows significantly better performance over the QLH model in terms of MAP and P@5. As for comparing with the RM3, the CERM improved the results over the RM3 significantly in MAP and P@5. Also, the performance of the CERM is significantly better than that of the MLGC. In sum, our concept-enhanced relevance model is more effective than the state-of-the-art models.

Table 1. The results of different models.

Collection	Metrics	QLH	RM3	MLGC	CERM
Ohsumed	MAP	0.2487	0.3058	0.2895	0.3269$^{*+\dagger}$
	P@5	0.4095	0.4590	0.4705	0.5086$^{*+\dagger}$
	P@10	0.3657	0.4295	0.4229	0.4629*†
Genomic	MAP	0.3527	0.3857	0.3568	0.4277$^{*+\dagger}$
	P@5	0.5385	0.5693	0.5231	0.6154$^{*+\dagger}$
	P@10	0.4808	0.4885	0.4654	0.5115†

*means statistically significant differences from the QLH with a 95% confidence according to Wilcoxon test. + means statistically significant differences from the RM3 and † denotes statistically significant differences from the MLGC.

5. CONCLUSIONS

Subject headings exhibit the subjects/topics of the document that they are assigned to and thus subject headings can accordingly be helpful resources for information retrieval. A novel retrieval model that integrates subject headings into the relevance model is proposed by considering the structure of conceptual representations. The document-concept associations and the concept-term associations are mined properly and represented in language models. The effectiveness of our model is evidenced by the experiments comparing with three strong state-of-the-art models. As a matter of fact, mining the document-concept associations and the concept-term associations is crucial to our model and guarantees its effectiveness. Such manually annotations as subject headings provide auxiliary information apart from the contents of documents and can enrich the representations of documents. As our model does, manually annotations indeed become effective resources for retrieval if an appropriate approach is applied.

6. ACKNOWLEDGMENTS

Jin Mao thanks the support from the National Natural Science Foundation of China funded projects under grant No. 71273196 and No. 71373286.

7. REFERENCES

[1] Lavrenko, V. and Croft, W. B. 2001. Relevance based language models. In *Proceedings of SIGIR 2001*(New Orleans, Louisiana, USA, September 13-14, 2001). ACM, New York, NY, 120-127. DOI= http://doi.acm.org/10.1145/383952.383972.

[2] Stokes, N., Li, Y., Cavedon, L., and Zobel, J. 2009. Exploring criteria for successful query expansion in the genomic domain. *Inform. Retrieval.* 12, 1(Jan. 2009), 17-50.

[3] Zeng, Q. T., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., and Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *J. Am. Med. Inform. Assn.* 13, 1(Jan/Feb.2006), 80-90.

[4] Lu, Z., Kim, W., and Wilbur, W. J. 2009. Evaluation of query expansion using MeSH in PubMed. *Inform. Retrieval.* 12, 1(Jan. 2009), 69-80.

[5] Hawking, D. and Zobel, J. (2007). Does topic metadata help with Web search?. *J. Am. Soc. Inf. Sci. Tec*.58, 5(Feb. 2007), 613-628.

[6] Meij, E., Trieschnigg, D., De Rijke, M., and Kraaij, W. 2010. Conceptual language models for domain-specific retrieval. *Inform. Process. Manag.*46, 4(Jul.2010), 448-469.

[7] Lv, Y., and Zhai, C. 2009. A comparative study of methods for estimating query language models with pseudo-feedback. In *Proceedings of CIKM 2009* (Hong Kong, China, November 2-6, 2009). ACM New York, NY, 1895-1898. DOI= http://doi.acm.org/10.1145/1645953.1646259.

[8] Lu, K. and Mao, J. 2013. Automatically infer subject terms and documents associations through text mining. In *Proceedings of the 76th ASIS&T Annual Meeting* (Montréal, Canada, November 1-5, 2013). ASIST, Maryland, 1-3.

[9] Ponte, J. M. and Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of SIGIR 1998*(Melbourne, Australia, August 24-28 1998). ACM, New York, NY, 275-281. DOI= http://doi.acm.org/10.1145/290941.291008.