Journal of Korean Library and Information Science Society
Vol. 52, No. 1

The MeSH-Term Query Expansion Models using LDA Topic Models in Health Information Retrieval[*]

MeSH 기반의 LDA 토픽 모델을 이용한 검색어 확장

Sukjin You (유석진)[**]

Contents

ABSTRACT : Information retrieval in the health field has several challenges. Health information terminology is difficult for consumers (laypeople) to understand. Formulating a query with professional terms is not easy for consumers because health-related terms are more familiar to health professionals. If health terms related to a query are automatically added, it would help consumers to find relevant information. The proposed query expansion (QE) models show how to expand a query using MeSH terms. The documents were represented by MeSH terms (i.e. Bag-of-MeSH), found in the full-text articles. And then the MeSH terms were used to generate LDA (Latent Dirichlet Analysis) topic models. A query and the top *k* retrieved documents were used to find MeSH terms as topic words related to the query. LDA topic words were filtered by threshold values of topic probability (TP) and word probability (WP). Threshold values were effective in an LDA model with a specific number of topics to increase IR performance in terms of infAP (inferred Average Precision) and infNDCG (inferred Normalized Discounted Cumulative Gain), which are common IR metrics for large data collections with incomplete judgments. The top *k* words were chosen by the word score based on (TP *WP) and retrieved document ranking in an LDA model with specific thresholds. The QE model with specific thresholds for TP and WP showed improved mean infAP and infNDCG scores in an LDA model, comparing with the baseline result.

KEYWORDS : MeSH, LDA, Topic Model, Information Retrieval, Query Expansion

요 약 : 헬스 분야에서 정보 검색의 어려움 중의 하나는 일반 사용자들이 전문적인 용어들을 이해하기가 어렵다는 점이다. 헬스와 관련된 전문 용어들은 일반 사용자들이 검색어로 사용하기 어렵기 때문에 이러한 전문 용어들이 자동적으로 검색어에 더해질 수 있다면 좀 더 검색의 효과를 높일 수 있을 것이다. 제안된 검색어 확장 모델은 전문 용어를 포함하는 MeSH (Medical Subject Headings)를 검색어 확장을 위한 단어 후보 군으로 이용하였다. 문서들은 MeSH용어들로 표현이 되고 이렇게 표현된 문서들의 집합에 대해서 LDA(Latent Dirichlet Analysis) 토픽들이 생성된 후, (검색어 + 초기 검색어에 의해 검색된 상위 *k*개 문서들)에 연관된 토픽 단어들이 원래의 검색어를 확장하는 데 쓰여졌다. MeSH로 구성된 토픽 단어들은 임의로 정해진 토픽 확률 임계값과 토픽을 구성하는 단어의 확률 임계값보다 높았을 때 초기의 검색어에 포함되었다. 특정수의 토픽을 갖는 LDA 모델에서 이러한 적절한 임계값의 설정을 통해 선택된 토픽 단어들은 검색어 확장에 이용되어 검색시에 infAP(inferred Average Precision)와 infNDCG(inferred Normalized Discounted Cumulative Gain)를 높이는데 효과적으로 작용하였다. 또한 토픽 확률값과 토픽 단어의 확률값을 곱하여 계산된 토픽 단어의 스코어가 높은 상위 k개의 단어를 검색어를 확장하는 데 이용하였을 때에도 검색의 성능이 향상될 수 있음을 확인하였다.

주제어 : MeSH, LDA, 토픽 모델, 정보검색, 검색어 확장

[*] This study is based on the part of the author's dissertation at the School of Information Studies at the University of Wisconsin-Milwaukee.

[**] Ph.D., Information studies, University of Wisconsin-Milwaukee (sukjinyou@gmail.com / ISNI 0000 0004 9414 2965)

• 논문접수: 2021년 2월 18일 • 최초심사: 2021년 2월 25일 • 게재확정: 2021년 3월 20일
• 한국도서관·정보학회지, 52(1), 79-108, 2021. http://dx.doi.org/10.16981/kliss.52.1.202103.79

Ⅰ. Introduction

Information retrieval (IR) is the process and activity to find information matching a user's information need. Differently from the past, a huge amount of information is being created and shared in electronic formats on the Web in the world every day. It is getting challenging to find relevant information today.

In query-based search, one popular way to improve IR performance is to add meaningful words following the original query. LDA (Latent Dirichlet Allocation, Blei, Ng, & Jordan, 2003) is one of the most common algorithms for topic modeling these days. An LDA model generates meaningful topic words representing the collection. Topic words generated by LDA can be candidate terms for query expansion (QE).

One weakness of the QE model using LDA topic words is that topic words might be too general to represent topics. Those words might not be helpful to retrieve a relevant document. Therefore, some dictionaries would be useful to filter out general words and identify key terms appropriate for a specific field. In the health domain, a health-related special terminology, such as MeSH (Medical Subject Headings) terms can provide more effective terms for QE.

Another challenge in QE is how to select relevant terms from candidate terms. An LDA model identifies topics related to a given text, such as a query or a retrieved document and generates topic words related to the topics. Although the LDA model is a good tool to collect candidate words for QE, the selection of appropriate words is the following concern. Some topic words might be relevant for QE, but others are not. If there is a way to filter out irrelevant words for QE, it would help to improve information retrieval.

Ⅱ. Purpose and objectives

The purpose of this research work is to identify how effective the application of LDA topic words based on MeSH terms is for QE in health IR.

Referring to the word probability and the topic probability for a topic word is one way to measure the extent to which the word is related to the query or the top *k* retrieved document by the query. If the topic for the word is highly related to the query and retrieved documents and the word is highly related to the topic generated by LDA models, the word would be likely to be related to the query. For effective word filtering, thresholds for topic probability (TP), word probability (WP), and the values of (TP * WP) were set up. It was assumed that the LDA topic words contribute to achieving better performance in terms of infAP and infNDCG, compared with the original query, especially when topic words are selected with thresholds. Mean infAP and infNDCG scores of 40 LDA models were compared with the result of the baseline run by two-sample t-test.

Ⅲ. Background and related work

IR models and topic models share a similar process and concepts. Text processing based on the Bag-of-Words Model and vector representation of a document is also required in topic modeling. The Language Model estimates the probability distribution of words to find documents related to a query. The probability distribution of words for a topic is an output in topic modeling.

1. The Bag-of-Words model

The Bag-of-Words model (Harris, 1954) is a basic assumption underlying in Boolean and Vector Space models, and even simple Probabilistic and Language models. In the Bag-of-Words model, a document is represented as a group of words. Syntactical structure and semantic implications are ignored, but only the lexicon is considered. The order of words in a document is meaningless. Only the term (word) is an element for representation. Information needs are represented by a set of assigned keywords (a query) and matched to index terms representing documents. A query ignoring the order or proximity between words is likely to miss an exact meaning that can be captured from multi-gram words. The specification of information needs or problems is limited in IR models based on the Bag-of-Words model. Meanwhile, the representation process can be finished easily in a short time in comparison to NLP representation. Although IR systems based on n-gram words or semantics (e.g. context and situations) cannot be easily implemented in the Bag-of-Words model, those IR systems might be effective in reflecting users' information needs. The use of proximity operators in Bag-of-Words models has been discussed (Mitchell, 1973) and implemented (Munro, Bolanos, & May, 1978; Schütze, Manning, & Raghavan, 2008) to consider the relationship between words for IR.

2. Query Expansion in IR

QE is the process to reformulate query for finding relevant documents by adding some terms related to an original query. A query can be expanded manually, automatically, and interactively (between a user and a system, Efthimiadis, 1996). Differently from manual QE, automatic QE consists of several steps before query reformulation (Azad & Deepak, 2019; Carpineto & Romano, 2012). In automatic QE, query reformulation is the final step of QE, where unnecessary terms are removed and new terms are added. For query reformulation, meaningful terms are extracted from internal or external collections, hand-built data sources (dictionaries, thesaurus, ontologies). Of the terms, terms related to the query are selected, weighted, ranked for term selection.

QE approaches have been categorized into two types of techniques: global analysis and local analysis (Azad & Deepak, 2019). For term selection, global analysis has employed various techniques according to data type: 1) linguistic techniques including syntactic, semantic, or contextual analyses on external data sources (e.g. ConceptNet, Liu, & Singh, 2004; WordNet & Miller, 1995), 2) concept extraction using term clustering, co-relation analysis between terms, term feature extraction using mutual information on an internal resource (corpus), 3) query-document relationship analysis on search logs (e.g. user, query, or search logs), and 4) query enrichment using semantic annotations and (hyper or linked) text on web-based resources (e.g. Wikipedia, anchor texts, and FAQs). Meanwhile, QE terms in the local analysis are selected from retrieved documents based on (pseudo) relevance feedback.

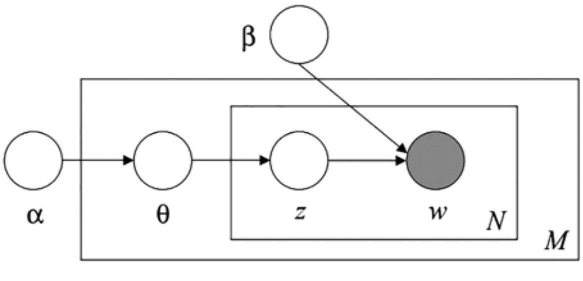
A concept is a group of related nouns. Not only an individual term but also a concept can help increase IR performance when it is used for QE. A concept can be extracted from corpora or retrieved documents based on data mining or machine learning techniques. A concept is a group of clustered terms, which may include not only synonyms but also adjacent (co-occurred) in term of context in the collection. For example, a concept might correspond to a topic in topic modeling. Generated concepts can be named by concept lexicons (e.g. LSCOM, Yanagawa et al., 2007). Natsev et al. (2007) expanded queries by mapping text, visual queries, and initially retrieved results to LSCOM-Lite 39 concepts. The presence of concepts related to a query was used to re-rank initial results in multimedia retrieval. Terms frequently occurred in concepts related to a query or the terms with high probability in a topic are likely to be appropriate terms for QE. Those concept-based or topic-based QEs has shown improvements in IR performance in terms of precision, recall, or F-measure (Chang, Ounis, & Kim, 2006; Xu &

Croft, 2017; Zeng et al., 2012).

3. Topic modeling

Topic clustering is one of the main interests in IR. A topic model is a statistical model to find abstract topics from a set of documents. Identification of the topics related to a query (or a tag) might be helpful for query reformulation or clustering of the related documents. The Topic model is focused on the identification of related topics by calculating the probability distributions over words, while a classical clustering algorithm (like K-means or hierarchical clustering) matches, in case of non-disjunctive clustering models, one cluster per document rather than multiple matches.

Blei, Ng, and Jordan (2003) developed a topic modeling method, LDA, which is one of the unsupervised learning techniques. The LDA algorithm categorizes the document into a mixed group of multiple topics. Under several topic categories, each topic word is distributed with a probability that shows how much the topic word presents the corresponding topic category. LDA transforms a document-term matrix into document-topic and topic-term matrices. Topic words are the terms associated with a topic in the second matrix where each is given a weight (i.e. probability). Blei, Ng, and Jordan (2003) introduced an inference technique based on variational methods and an EM (Expectation-Maximization, Hofmann, 1999) algorithm for Bayes parameter estimation, which is an optimization approach (figure 1).



<figure 1> LDA plate notation (Blei, Ng, & Jordan, 2003)

A word is the basic unit in a vocabulary indexed by {1..., V}. A document is a sequence of N words: $\boldsymbol{w} = (w_1, w_2, ..., w_n)$, where w w_n is the *n*th word in the sequence. A corpus is a collection of *M* documents: $D = \{ w_1, w_2, ..., w_m \}$. Dirichlet prior $\alpha$is given for the topic distributions (θ_d) for each document *d*. For the observed words (***w***) and the number of topics (*k*), topic z_n is assigned for w_n over the multinomial variable (θ). the word probabilities are para meterized by a $k \times V$ matrix, β The LDA problem is to solve the probabilities of topic-document and topic-word for a document.

$$p\,(\boldsymbol{\theta},\mathbf{z}\mid\mathbf{w},\boldsymbol{\alpha},\,\boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta},\ \mathbf{z},\ \mathbf{w}\mid\boldsymbol{\alpha},\ \boldsymbol{\beta})}{p(\mathbf{w}\mid\boldsymbol{\alpha},\ \boldsymbol{\beta})} \tag{1}$$

Two free variational parameters, φ and γ, which are used to estimate the topic-word distributions (z) and the topic distributions (θ) for each document, are updated by stochastic iteration process - minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior. Original Variational Bayes (VB) is not practical for a collection including large amounts of documents because it is based on batch processing needing all documents in a collection. Hoffman, Bach, and Blei (2010) developed the Online VB inference for LDA to overcome this weakness of the batch VB inference by just looking at parts of documents.

LDA has been applied to IR by combining with other IR algorithms. Traditional IR approaches such as BM25 (Beaulieu et al., 1997), MATF (Multi Aspect TF, Paik, 2013), and Dirichlet LM (Zhai & Lafferty, 2004) increased MAP, P@5 and P@20 when LDA was applied together (Jian et al., 2016). Zhou et al. (2017) showed that QE expansion models based on the integration of LDA and word embedding can increase IR performance in terms of MAP, NDCG, and MRR (Mean Reciprocal Rank), compared to common IR approaches using Kullback-Leibler divergence, pseudo-relevance feedback, or co-occurrence statistics. Sometimes, LDA-based IR approaches are used to compare IR approached with concept-based approach (Wang, Huang, & Feng, 2017).

Ⅳ. Data and methods

1. Dataset

Setting up a dataset is a costly process in quantitative research. The TREC CDS (Clinical Decision Support) track has provided several sets of data collections (e.g. PubMed Central) with a gold standard for a judgment set to participants for IR tasks. 108, 012 documents were assessed with three relevance levels (“Definitely Relevant”, “Possibly Relevant”, or “Not Relevant”) by physicians, most of whom were either biomedical informatics students or postdoctoral fellows (Roberts et al., 2016). Two categories (“Definitely/Possibly Relevant” and “Not Relevant”) were deployed for infAP, while three categories ("Definitely Relevant", "Possibly Relevant", and "Not Relevant") were employed for infNDCG.

Dataset for indexing. Using the TREC data and evaluation scheme is an easy way to save the evaluation cost as well as the data collection cost. The 2016 CDS track dataset (http://www.trec-cds.org/2016.html) was indexed by the search engine, Terrier (http://terrier.org/). Terrier was used to generate search results. For the 2016 dataset, it includes 6,970 folders－journals, and 1,495,289 files－full-text articles (52G). The indexing was conducted before this study. Indexing a huge number of documents takes much time (e.g. a few weeks). Even though the data for indexing are slightly different from the data for LDA modeling, assuming the difference would not be critical to this study, the indexing data previously generated based on NXML format, which is XML encoded using the U.S. National Library of Medicine (NLM) Journal Archiving and Interchange Tag Library (http://jats.nlm.nih.gov/archiving/versions.html), was used.

Dataset for LDA topic models. In this study, a PMC (PubMed Central) snapshot (12/04/2016, ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/) was used. There are 6966 OA (open access, https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist) journals included in PMC, which were categorized into three types: 1) full participation－depositing the complete contents of each volume and issues, starting with a specific volume and issue, 2) NIH Portfolio －depositing all NHI-funded articles, and 3) selective deposit－including a selected set of articles by publishers (NIH, 2015). The number of OA journals might be slightly different by the time when data are collected. This dataset was used for the LDA model generation. Of 1,451,661 text files, 1,451,651 (50.3 GB) documents were represented by MeSH terms to create LDA topic models. Any MeSH terms were not included in 10 documents, therefore, those documents were ignored.

2. The Bag of MeSH

For IR in health information, terms more related to health can contribute to IR performance. The National Library of Medicine publishes a controlled vocabulary thesaurus called MeSH (Medical Subject Headings). MeSH data consist of three types of data: 1) MeSH descriptor, 2) MeSH qualifier, and 3) MeSH Supplemental Concept Records (SCR). Generally, LDA topic models are constructed based on all words included in a collection. When including all the words in a document, an LDA model includes many general terms in topics. MeSH can be a more effective terminology than general terms in health IR (Díaz-Galiano et al., 2007; Lu, Kim, &

Wilbur, 2009; Merabti et al., 2009; Mu, Lu, & Ryu, 2014). Another benefit of using MeSH is that preprocessing of documents represented by only MeSH terms is more efficient than using all words included in documents. For these reasons, a document was represented by MeSH terms, which are included in the full-text article. MeSH terms (n-gram) used in this study were extracted from the descriptor field (MH) in the 2016 MeSH descriptor file, which comprises 27,883 descriptors (https://www.nlm.nih.gov/mesh/download_mesh.html), but 24,883 MeSH terms were observed in the collection.

3. QE models using LDA models

One of the key processes to QE is finding words related to a query. Queries in this study consist of one or two sentences in most cases. A query can be considered as a document in IR. A query and retrieved documents by a search engine can be used to collect candidate QE words related to the query. LDA is a popular technique to predict a topic as a concept given a document. LDA topics are generated for a collection. In IR on the same collection in the same domain, words selected by LDA would be more appropriate for QE than words chosen from other terminologies that are created based on external sources in different domains. Of candidate words, more relevant words for QE can be selected by thresholds of Topic Probability (TP), word probability (WP), the multiplication of TP and WP. An LDA topic consists of words defined in a dictionary. TP of a document is the topic distribution showing how closely the document is related to each topic, while WP of a topic shows how closely the word is related to each topic. Therefore, TP and WP can be used to identify more relevant words to a document or query. TP and WP were calculated using online Variational Bayes (Hoffman, Bach, & Blei, 2010) by *genism*, a python library. The QE model uses an LDA topic words with thresholds for topic and word probability.

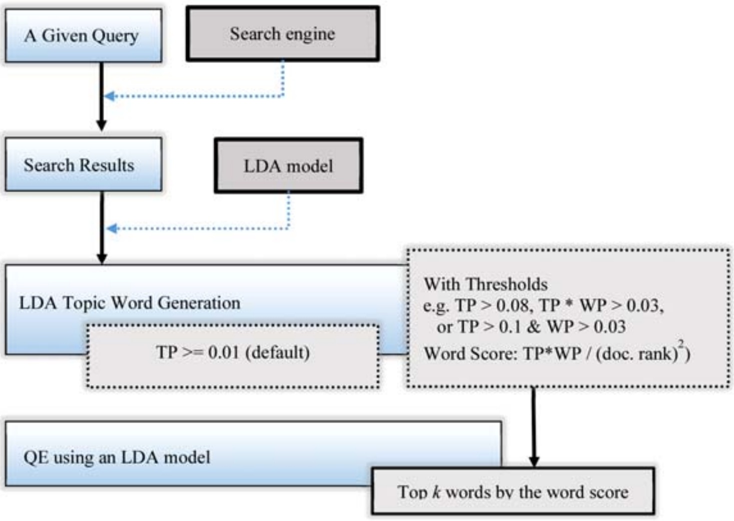
Parameter Setting. Parameters related to word selection for QE affect the baseline results. Two parameters, 1) the number of top-ranked retrieved documents to generate topic words and 2) a power value (k) to weight word scores regarding the rank, $1/\text{rank}^k$, were adjusted to generate better performance by evaluating IR performance for k in terms of infAP and infNDCG.

Topic word (MeSH term) scoring. To rank topic words for QE, a basic word score (S_w) was calculated using Topic Probability (TP), Word Probability (WP), and Document Rank (DR):

$$S_w = TP * WP \ / \ (DR)^2 \qquad (2)$$

LDA topics were generated based on only MeSH terms included in a query and retrieved documents. When the LDA topic probability of a query (the summary field, http://www.trec-cds.org/topics2016.xml) or a retrieved document is higher than or equal to 0.01, the LDA topic is considered as a related topic to the query or the retrieved document. LDA topic words were identified by the topic of a query and the retrieved top-ranked documents. The query text was included in the text of the first ranked document because the query is as an important source as the retrieved document in extracting LDA topics related to the query. Using the rank of the retrieved document can be helpful to score the word for QE. For example, the topic words generated by the first-ranked documents have more weight than the topic words generated by the second-ranked document. A maximum of the top 10 words, empirically, was selected as terms for QE by the descending order of the word score in 40 (the more, the better, but limited by the time and cost for generating LDA models in this study) LDA models with different numbers of topics. If a word has more than two scores, the highest score was given to the word.

The overall steps for QE models using LDA models were illustrated in Figure 2.



<Figure 2> QE model using an LDA models with threshold for TP and WP

1. Search result generation by the search engine, Terrier.
2. LDA topic words were generated by LDA models that were created on the PMC snapshot before search process: topic words were generated with different thresholds for topic probability

(TP), word probability (WP), and TP*WP.

- The default topic probability (TP) threshold was set up as 0.01. If the topic probability of the retrieved documents is higher than 0.01 or equal to 0.01, the topic was considered as a related topic to the document. Retrieved documents have a rank. The top 1 ranked document or the top 2 ranked documents were used to generated LDA topic words. Topic words are scored by (TP*WP/ (document rank)^2).
- Topic words were filtered by specific thresholds for 1) TP (e.g. 0.08 or 0.1), 2) WP (e.g. 0.03), or 3) TP*WP (e.g. 0.08). The threshold values were determined by the result in an LDA model with 1700 topics in terms of infAP, infNDCG, and the ratio of the number of positive words and negative words in improving IR performance (Section 5). Threshold values generating high infAP and infNDCG scores and high ratio values were preferred. Topic words were sorted by word score. The top 10 words were added to the original query and search results by this new query were evaluated in terms of infAP and infNDCG.

To sum up, the documents related to a query were retrieved by a search engine, and then the query and top k retrieved documents were used to generate LDA topic words assuming the word would be relevant for QE. The LDA topics words were filtered for QE by the LDA-related thresholds to choose more relevant words for the given query.

Preprocessing. Preprocessing was required before generating LDA models. LDA models with different numbers of topics were trained based on the 2016 PMC snapshot (Dec. 4). Before training, each document was represented by only MeSH terms that are included in the document. MeSH terms were extracted from the "MH" field (MeSH) in the 2016 MeSH descriptor file (https://www.nlm.nih.gov/mesh/download_mesh.html) so that LDA topics consist of MeSH terms. Only a complete MeSH term described in the MeSH was considered as a unit of analysis. MeSH terms are multi-gram based, which might be one word or more than one word. If a MeSH term consists of a word, the word is fine as a unit. Special characters (e.g. “,”, “(“, and “)”) in the descriptors were ignored. Document representation is based on the Bag of MeSH (n-gram) model. Each document was represented in the form of a pair of words and frequency. A dictionary (including MeSH terms) and a corpus (including document representations) were created for generating LDA topic models. The dictionary for LDA models includes all MeSH terms used in the collection. The average numbers of all and unique MeSH terms included in a document were

286.2 and 75.9, respectively.

Because the Python library (Rehurek & Sojka, 2010), *genism*, has been updated steadily and relatively stable and efficient to handle large size of datasets (documents), *genism* was used to create LDA topic models, which was implemented using the Variational Bayesian inference algorithm. 40 LDA models were generated using *gensim* with the maximum number of iterations, 50, which is called *pass* in *genism*. In most cases, topics were converged after one pass.

IR Evaluation. TREC datasets and evaluation scheme of the TREC 2016 Clinical Decision Support (CDS) track was used. The TREC 2016 CDS track provides a snapshot of an open-access subset on March 28, 2016, for ad hoc retrieval tasks. Full-text articles were distributed in the NXML format (XML encoded using the NLM Journal Archiving and Interchange Tag Library). There are 30 queries (called topic) given in the CDS track. 30 queries were used for LDA topic generation along with retrieved documents. The text for original queries was integrated with the text for the top 1 retrieved (ranked) document.

LDA models were used to expand the queries. LDA top *n* topic words for the top *k* retrieved documents were added to the original query for QE. In addition to the baseline run for the original queries, several runs based on QE models using LDA models, were generated:

- Queries for the baseline run. 30 texts in the *summary* fields of original (query) topics (http://www.trec-cds.org/topics2016.xml, e.g., *An elderly female with past medical history of right hip arthroplasty presents after feeling a snap of her right leg and falling to the ground*) have been used as queries. The baseline run was created by the search engine using the original query without QE. 1000 search results per query were included in the baseline run. The search algorithm is based on the Language Model using Bayesian smoothing with Dirichlet Prior (Zhai & Lafferty, 2004). Porter stemmer was set up as the default for the retrieval in Terrier.
- Query Expansion (QE) using the LDA top 10 topic words. LDA topics words, which are related to the query and the top *k* documents that were retrieved by the query, were generated by an LDA model. The top 10 words were selected by the descending order of the word score based on the topic probability, word probability, and the rank of the retrieved document: TP * WP * (1 / (document rank)^2) for the top *k* retrieved documents. Two types of LDA models with thresholds for TP, WP, and TP*WP were created:

1) The basic QE model using the LDA model with a topic probability threshold, 0.01 (by default), because it is not effective to consider many topics with low topic probability values as

related topics, topics with TP lower than 0.01 were ignored as unrelated topics.

2) The QE model using the LDA model with specific LDA threshold values - e.g. the threshold, 0.08 or 0.1 for TP and 0 .03 for WP or 0.03 for TP*WP.

Evaluation measures commonly used, such as average precision, R-precision, and precision-at cutoff k, are not robust to incomplete relevance judgments. bpref (Buckley & Voorhees, 2004) was more effective to incomplete relevance than R-precision and precision at 10 (P@10) in terms of Kendall's correlation between the system ranking evaluated by the original judgment set and the system ranking produced using the reduced judgment set. Yilmaz and Aslam (2006) proposed three evaluation measures for an incomplete judgment set: induced AP (Average Precision), subcollection AP, and inferred AP. Kendall's τ, linear correlation coefficient ρ and root mean squared (RMS) error were calculated to see the changes as the judgment set is reduced. Compared with bpref, three evaluation measures were robust to the reduced judgment set. Similarly, inferred NDCG (Normalized Discounted Cumulative Gain) consistently outperformed infAP and nDCG on random judgments in terms of Kendall's τ and root mean squared (RMS) error (Yilmaz, Kanoulas, & Aslam, 2008).

Two inferred measures, including inferred AP and inferred NDCG, have become popular measures for large data collections with incomplete judgments (Bompada et al., 2007; Voorhees, 2014)－especially, in TREC (Lupu et al. 2011; Roberts et al., 2015; Roberts et al., 2017).

Evaluation for the IR tasks depended on the scheme of the TREC 2016 CDS track based on infAP (inferred Average Precision) and infNDCG (inferred Normalized Discounted Cumulative Gain) as IR evaluation measures, which are robust evaluation measures for an incomplete judgment set.

Ⅴ. Results

1. The number of the top-ranked retrieved documents

For 40 LDA models, infAP and infNDCG scores for 30 queries were calculated for 1000 results when terms from first-ranked document are selected for QE (Table 1). Even though there were five scores (in bold) shown more than the scores of the baseline run (infAP: 0.0209 & infNDCG: 0.1808), most LDA models showed lower infAP and infNDCG scores.

<Table 1> Mean infAP and infNDCG scores of 40 LDA models with different numbers of topics for the top 1 retrieved document (TP threshold: 0.01)

no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0163	0.0191	0.0203	0.0175	0.0152	0.0168	0.0166	0.0176	0.02	0.0175
infNDCG	0.1479	0.1596	**0.1817**	0.1645	0.1489	0.1678	0.1539	0.1729	0.1743	0.1731
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0179	0.0188	0.017	0.0156	0.0158	**0.0223**	**0.0247**	0.0188	0.0182	0.0167
infNDCG	0.1651	0.1754	0.1526	0.1565	0.1613	**0.1917**	**0.1845**	0.1837	0.1604	0.1594
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0179	0.0166	0.0205	0.0164	0.0165	0.0202	0.0188	0.0184	0.0208	0.0186
infNDCG	0.1783	0.1466	0.1747	0.1584	0.1644	0.175	0.1696	0.1627	0.1717	0.1744
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0186	0.0173	0.0188	0.0204	0.0181	0.0172	0.0169	0.0202	0.0192	0.0184
infNDCG	0.1744	0.1778	0.1674	0.1746	0.1761	0.1707	0.1689	0.1733	0.1712	0.1709

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

The top2 retrieved documents include two documents when searching terms for query expansion: the first-ranked document and the second-ranked document. To compare the results for the top2 retrieved documents with the results for the top 1 document, infAP and infNDCG scores of 40 LDA models using the top2 retrieved documents were listed in Table 2. Results were generated based on the ranking weight, the inverse value of the document rank to the power of two: 1 / (document rank)^2.

<Table 2> Mean infAP and infNDCG scores of 40 LDA models with different numbers of topics for the top2 retrieved documents (TP threshold, 0.01) with score weighting of the rank number to the power of 2

no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0161	0.0197	**0.023**	0.0202	0.0159	0.019	0.0202	0.0184	**0.0236**	0.0199
infNDCG	0.1513	0.1641	**0.1948**	0.1723	0.1567	0.1734	0.1738	0.1795	**0.1823**	**0.1841**
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0197	0.0193	0.0198	0.0201	0.0199	**0.0239**	**0.0255**	**0.0221**	**0.0211**	0.0204
infNDCG	0.1688	0.1746	0.1703	0.1768	0.1804	**0.1954**	**0.1935**	**0.1962**	0.1744	0.1786
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0203	0.0201	0.0208	**0.0213**	0.0181	**0.0223**	0.0197	**0.0213**	0.0206	**0.0212**
infNDCG	0.1868	0.1548	0.1777	0.1807	0.1682	0.1773	0.172	0.1718	0.1732	0.1778
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0198	0.0198	**0.0216**	**0.0227**	0.0189	0.0199	**0.0232**	**0.022**	**0.0216**	0.0199
infNDCG	0.1662	**0.1822**	0.171	**0.184**	**0.1814**	0.1751	**0.1934**	0.1803	**0.1832**	0.1744

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

Being compared with the average mean infAP (0.0183) and infNDCG (0.1684) scores for the top 1 document, the average mean infAP (0.0206) and infNDCG (0.1768) scores for the top 2 documents were higher, although the scores were lower than the scores for the baseline run, 0.0209 for infAP and 0.1808 for infNDCG. Several LDA models showed better infAP and infNDCG scores than the scores of the baseline run: 15 LDA models for infAP and 11 LDA models for infNDCG, which might imply that LDA application can be effective in increasing IR performance. LDA models for the top 2 retrieved documents have shown relatively better infAP and infNDCG scores than the LDA models for the top 1 and top 3 (Table 3) retrieved documents. Top 2 retrieved documents might be more appropriate to refer to for QE in applying LDA to IR rather than consulting only top1 retrieved document in that, in general, more expanded queries show higher recall scores. Meanwhile, using top 3 retrieved documents for QE might not be effective in increasing precision. Therefore, top 2 retrieved documents were used in generating topic words for proposed QE models.

〈Table 3〉 Mean infAP & infNDCG scores of 40 LDA models with different numbers of topics for the top 3 retrieved documents with score weighting of the rank number to the power of 2

no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0161	0.0195	**0.0233**	0.0203	0.0153	0.0186	0.0186	0.0188	**0.0236**	0.02
infNDCG	0.1515	0.1629	0.196	0.1738	0.1537	0.1766	0.1671	0.1801	**0.1827**	0.1771
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0197	0.0189	0.0196	0.0201	0.0195	**0.0233**	**0.0243**	**0.0215**	0.0209	0.02
infNDCG	0.1698	0.1716	0.1701	0.1772	0.176	**0.1936**	**0.1876**	**0.1943**	0.1696	0.1725
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0203	0.0199	0.0201	**0.021**	0.0172	**0.0214**	0.0194	0.0201	0.0188	**0.0228**
infNDCG	**0.1898**	0.1534	0.1745	**0.1809**	0.171	0.1728	0.1675	0.1691	0.1725	**0.1821**
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0168	0.0188	0.0192	**0.0224**	0.0189	0.0178	**0.0232**	0.0193	**0.0229**	0.0194
infNDCG	0.1744	0.1778	0.1674	0.1746	0.1761	0.1707	0.1689	0.1733	0.1712	0.1709

* baseline run - infAP: 0.0209 & infNDCG: 0.1808

2. Ranking weight

For the LDA model with 3700 topics and when top 2 retrieved documents are searched for expansion terms, which showed relatively high mean infAP (0.0232) and infNDCG (0.1934) scores, the rank of a retrieved document was used to score a word for QE. The score was weighted by the inverse value of the rank to the power of *k*: 1 / (document rank) ^{k}. Mean infAP and infNDCG scores were compared according to the power value, *k*, in Table 4.

<Table 4> Mean infAP and infNDCG scores for the power values and the number of top retrieved documents (the LDA model with 3700 topics)

no. top docs	1	2	3	4	5	6	7	8	9	10
The power of 0.5										
infAP	0.0241	0.0251	0.0191	0.0203	0.0204	0.0196	0.02	0.0203	0.0191	0.0194
infNDCG	0.1955	0.1915	0.1811	0.1877	0.1833	0.1805	0.1754	0.1729	0.1752	0.1768
The power of 1										
infAP	0.0241	**0.0254**	0.0215	0.0225	0.024	0.0228	0.0214	0.0214	0.021	0.0211
infNDCG	0.1955	0.1942	0.1894	0.1951	0.1932	0.19	0.1864	0.1878	0.1826	0.1847
The power of 2										
infAP	0.0241	0.025	0.0251	0.0245	0.0245	0.0245	0.0245	0.0245	0.0245	0.0245
infNDCG	0.1955	**0.1988**	0.1971	0.1961	0.1961	0.1961	0.1961	0.1961	0.1961	0.1961
The power of 3										
infAP	0.0241	0.0242	0.0246	0.0242	0.0242	0.0242	0.0242	0.0242	0.0242	0.0242
infNDCG	0.1955	0.1944	0.1968	0.1944	0.1944	0.1951	0.1944	0.1951	0.1944	0.1944

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

QE using the power of 2 and 3 showed the stable infAP scores for the top 2 or 3 retrieved documents. Although the best mean infAP (0.0254) score for the top 1 document was observed when the power of 1 was applied, the mean infAP score for the power of 1 showed a high variance of infAP scores regarding 10 different numbers of retrieved documents.

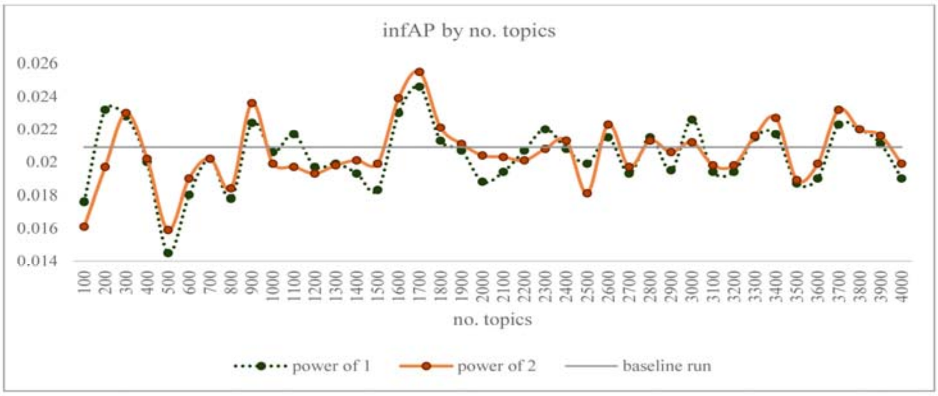
Similarly, QE using top 2 or 3 retrieved documents showed stable and high infNDCG scores when the power of 2 and 3 were applied to the word score. The best infNDCG score (0.1988) for the top 2 retrieved documents was observed when the power of 2 was applied.

Mean InfAP and infNDCG scores of 40 LDA models for the top 2 retrieved documents were compared according to two different power values: the power of 1 (Table 5) and 2 (Table 2). The QE based on word scores weighted by the power of 2 showed slightly better average mean infAP (0.0206) and infNDCG (0.1768) scores of 40 LDA models than QE using the power of 1 (0.0204 for infAP and 0.1743 for infNDCG), although the QE using the power of 1 showed better mean scores in several LDA models: 11 LDA models with 100, 200, 1000, 1100, 1200, 1300, 2200, 2300, 2500, 2800, and 3000 topics (Figure 3) for infAP and 9 LDA models with 100, 200, 1100, 1300, 2200, 2500, 2800, 3100, and 3800 topics (Figure 4) for infNDCG.

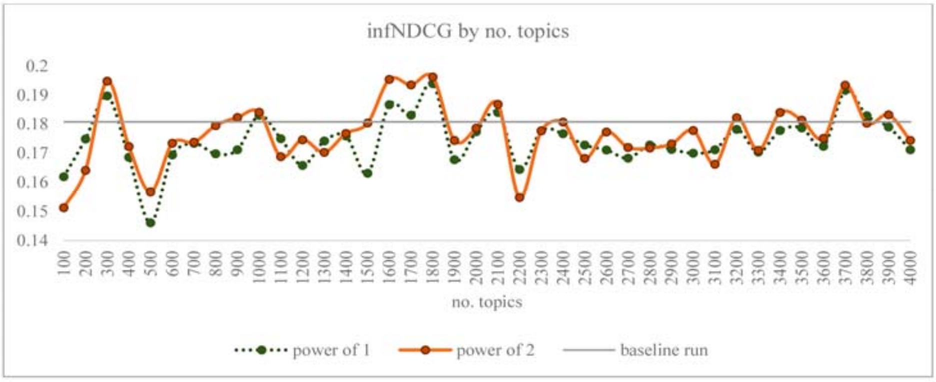
〈Table 5〉 Mean infAP and infNDCG scores for the LDA models with the weighting - the inverse value of the rank to the power of 1 for the top 2 retrieved documents (the best in bold)

no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0176	**0.0232**	0.0228	0.02	0.0145	0.018	0.0202	0.0178	0.0224	0.0206
infNDCG	0.1619	0.1749	**0.1897**	0.1686	0.1461	0.1694	0.1734	0.1698	0.1712	0.183
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0217	0.0197	0.0199	0.0193	0.0183	0.023	0.0246	0.0213	0.0207	0.0188
infNDCG	0.175	0.1658	0.1742	0.1759	0.1631	0.1867	0.1831	0.1939	0.1678	0.1774
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0194	0.0207	0.022	0.0208	0.0199	0.0215	0.0193	0.0215	0.0195	0.0226
infNDCG	0.184	0.1644	0.1777	0.1767	0.1728	0.1711	0.1683	0.1728	0.1713	0.17
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0194	0.0194	0.0215	0.0217	0.0187	0.019	0.0223	0.022	0.0211	0.019
infNDCG	0.1712	0.1782	0.1704	0.1778	0.1787	0.1724	0.1916	0.1829	0.179	0.1712

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)



〈Figure 3〉 Mean infAP scores of 40 LDA models for the top 2 retrieved documents-weighted by the power of 1 and 2



〈Figure 4〉 Mean infNDCG scores of 40 LDA models for the top 2 retrieved documents-weighted by the power of 1 and 2

3. QE using Thresholds for LDA TP, WP, and TP * WP

The thresholds for TP, WP, and TP*WP might affect IR performance. A high TP threshold would filter out minor topics from the top retrieved documents, while a high WP threshold would filter out less important words for a topic. For the model with 1700 topics, which showed a relatively high average infAP and infNDCG scores. infAP and infNDCG scores were measured by the thresholds for TP, WP, and TP * WP between 0 and 1.0 at 100 probability levels (level distance: 0.01). For example, there were 10,000 runs for TP * WP threshold determination.

3.1 QE using thresholds for TP and WP

According to different TP and WP threshold values, the mean infAP and infNDCG scores for the model with 1700 topics were calculated. The LDA model was generated based on the top 1 retrieved document. The top 9 results by mean infAP and infNDCG scores were listed along with the number and ratio of positive and negative words in Table 6 and 7, respectively. Positive words for QE mean words that increase IR performance (infAP and infNDCG) when used for QE, while negative words decrease IR performance. In the case that there is no negative word, 1 is added for the divisor, preventing from being zero.

〈Table 6〉 Mean infAP and infNDCG scores for TP and WP thresholds sorted by infAP score (1700 topics based on the top 1 retrieved document)

TP	WP	infAP	infNDCG	No. positive words	No. negative words	No. positive words / (No. negative words +1)
0.15	0.02	0.0277	0.1813	25	54	0.45
0.14	0.02	0.0276	0.1856	38	66	0.567
0.16	0.02	0.0274	0.18	21	49	0.42
0.09	0.02	0.0273	0.1876	64	110	0.577
0.08	0.02	0.0272	0.1869	64	110	0.577
0.15	0.03	0.0267	0.1892	23	51	0.442
0.11	0.02	0.0267	0.1872	49	92	0.527
0.19	0.02	0.0267	0.182	13	29	0.433
0.2	0.02	0.0267	0.182	13	29	0.433

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

TP values for the top 9 infAP scores were distributed between 0.08 and 0.2 for infAP, while there were only two WP values, 0.02 and 0.03. The highest infAP score was observed in the LDA models with the thresholds, 0.15 for TP and 0.02 for WP. The highest ratio of the number of positive words and negative words was shown in the LDA model with the thresholds, 0.09 and 0.08 for TP and 0.02 for WP.

〈Table 7〉 Mean infAP and infNCDG scores for TP and WP thresholds sorted by infNDCG score (1700 topics based on the top 1 retrieved document)

TP	WP	infAP	infNDCG	No. positive words	No. negative words	No. positive words / (No. negative words + 1)
0.07	0.03	0.0258	0.1963	75	135	0.551
0.06	0.03	0.0253	0.1948	77	142	0.538
0.07	0.24	0.0237	0.1939	30	27	1.071
0.14	0.03	0.0266	0.1936	35	62	0.556
0.08	0.03	0.0263	0.1926	60	99	0.6
0.09	0.03	0.0263	0.1926	60	99	0.6
0.06	0.24	0.0233	0.1923	32	28	**1.103**
0.07	0.6	0.0219	0.1921	13	12	1.0
0.07	0.61	0.0219	0.1921	13	12	0.433

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

TP values were distributed between 0.06 and 0.14, while WP values were between 0.03 and 0.61 (0.03 for five cases) for the top 9 infNDCG scores. The highest inNDCG score was measured in the LDA model with the thresholds, 0.07 for TP and 0.03 for WP. The highest ratio (1.103) of positive words and negative words was shown in the LDA model with the thresholds, 0.06 for TP and 0.24 for WP where positive words were more than negative words.

The top 2 ranked TP (0.07 and 0.06) and WP (0.03) values were relatively low and positive and negative words were more than others. Based on the results, the thresholds for TP (0.1) and WP (0.03) were applied to 40 LDA models. The mean infAP and infNDCG scores were shown in Table 8. The average mean infAP and infNDCG scores of 40 LDA models with the thresholds for TP, 0.1, and WP, 0.03, were 0.0188 and 0.1633, respectively. The better scores than the scores of the baseline run were in bold. In most LDA models except for the LDA model with 1700 topics, infAP and infNDCG scores were lower than the scores of the baseline run.

〈Table 8〉 Mean infAP and infNDCG scores of 40 LDA models with the thresholds for TP - 0.1 and WP - 0.03 for the top 1 retrieved document

no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0135	0.019	0.019	0.0171	0.0156	0.0151	0.0191	0.0175	0.0203	0.0173
infNDCG	0.1265	0.1545	0.1598	0.1707	0.1534	0.1502	0.1648	0.1627	0.172	0.1692
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0171	0.019	0.017	0.0097	0.0166	**0.021**	**0.0254**	0.0188	0.0178	0.0192
infNDCG	0.152	0.1541	0.1531	0.1245	0.1615	0.1681	**0.1888**	0.1614	0.1551	0.1676
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0175	0.0185	0.0219	0.0168	0.0187	0.0194	0.018	0.0194	0.02	0.0175
infNDCG	0.1708	0.1573	0.1759	0.1541	0.1629	0.1561	0.1587	0.166	0.1599	0.17
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0197	0.0202	0.019	**0.0227**	**0.0218**	0.0208	0.0205	0.0194	0.0203	**0.021**
infNDCG	0.1664	0.1667	0.1698	0.1731	**0.1842**	0.1683	0.1778	0.177	0.1745	0.1744

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

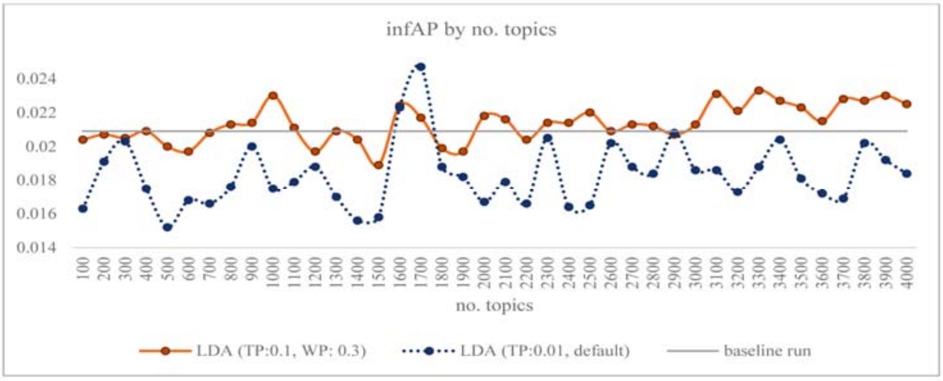
The ratio of positive and negative words (no. positive words / (no. negative words + 1)) can be another indicator to decide threshold values. The high ratios were shown in the LDA models with the threshold for WP, 0.24 (1.103 for TP, 0.06, and 1.071 for TP, 0.07) in Table 9. Because the infAP and infNDCG scores were low, another WP, 0.3 (more than 0.24, but roughly similar), was applied instead of 0.03 (Table 9). The ratio of positive (16) and negative words (16) generated by the LDA model with the thresholds (TP: 0.1 and WP: 0.3) were 0.9412 (16 / (1+16)).

〈Table 9〉 Mean infAP and infNDCG scores of 40 LDA models with the thresholds for TP (0.1) and WP (0.3) for the top 1 retrieved document

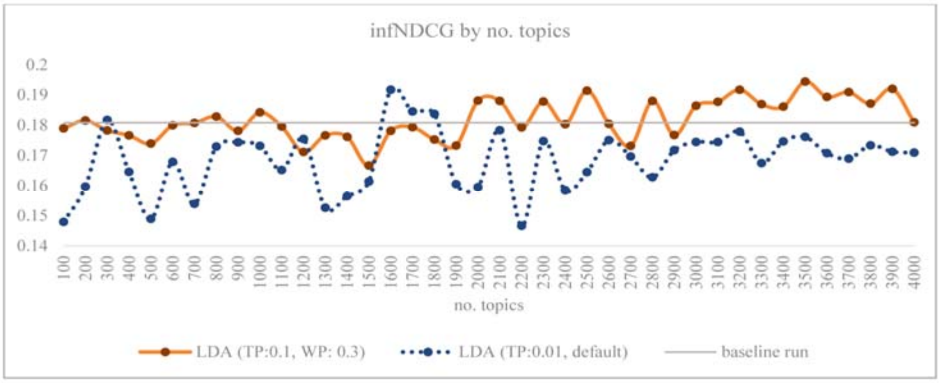
no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0204	0.0207	0.0205	0.0209	0.02	0.0197	0.0208	**0.0213**	**0.0214**	**0.023**
infNDCG	0.1789	0.1815	0.1782	0.1766	0.1739	0.1799	0.1807	**0.1828**	0.1781	**0.1842**
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	**0.0211**	0.0197	0.0209	0.0204	0.0189	**0.0224**	**0.0217**	0.0199	0.0197	**0.0218**
infNDCG	0.1795	0.1711	0.1766	0.1761	0.1666	0.1781	0.1793	0.1752	0.1732	**0.1881**
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	**0.0216**	0.0204	**0.0214**	**0.0214**	**0.022**	**0.0209**	**0.0213**	**0.0212**	0.0207	**0.0213**
infNDCG	**0.188**	0.1792	**0.1878**	0.1803	**0.1914**	**0.1804**	0.1731	**0.188**	0.1767	**0.1864**
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	**0.0231**	**0.0221**	**0.0233**	**0.0227**	**0.0223**	**0.0215**	**0.0228**	**0.0227**	**0.023**	**0.0225**
infNDCG	**0.1877**	**0.1917**	**0.1869**	**0.1861**	**0.1944**	**0.1893**	**0.1909**	**0.1871**	**0.192**	**0.1809**

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

Mean infAP and infNDCG scores were compared between two LDA models with different thresholds for TP and WP: 1) TP: 0.01 and 2) TP: 0.1 & WP: 0.3 (Figure 5 and 6). The average mean infAP and infNDCG scores of 40 LDA models with the thresholds for TP, 0.1, and WP, 0.3, were 0.0213 and 0.1819, respectively, which are higher than the scores of 40 LDA models with the thresholds for TP, 0.1, and WP, 0.03 (0.0188 for infAP and 0.1633 for infNDCG) as well as the scores for the baseline run. Better mean infAP and infNDCG scores were observed in 39 and 36 LDA models with the thresholds, 0.1 for TP and 0.3 for WP, respectively. Meanwhile, the average mean infAP and infNDCG scores of 40 LDA models with the threshold for TP, 0.01 were 0.0183 and 0.1684. There were statistically significant differences in the average mean infAP and infNDCG scores (paired t-test, alpha = 0.05, p-value = 2.3E-12 for infAP and 1.7E-09 for infNDCG).



〈Figure 5〉 Mean infAP scores of 40 LDA models with different thresholds for TP and WP for the top 1 retrieved documents



〈Figure 6〉 Mean infNDCG scores of 40 LDA models with different thresholds for TP and WP for top 1 retrieved document

27 and 19 LDA models showed better mean infAP and infNDCG scores, respectively, than the baseline run. It implies that it might be effective to have specific thresholds for TP and WP, considering that just 2 and 3 LDA models with the threshold for TP, 0.01 (Table 5), showed better infAP and infNDCG scores, respectively, than the baseline run (Figure 8 & Figure 9). Compared with the scores of the baseline run, there was a statistically significant difference in the average mean infAP score, but not in the average mean infNDCG score (two-sample t-test, alpha = 0.05, p-value = 0.0135 for infAP and 0.2813 for infNDCG.

In a similar fashion, mean infAP and infNDCG scores of 40 LDA models with the thresholds for TP, 0.1, and WP, 0.3, were measured for the top 2 retrieved documents (Table 10). The average mean infAP and infNDCG scores were 0.0201 and 0.1696, respectively, which are lower than the scores of the baseline run as well as the scores of the LDA models for the top 1 retrieved document, 0.0213 for infAP and 0.1819 for infNDCG. Compared with the scores of the baseline run, there was a significant difference in the average mean score for infAP and infNDCG (two-sample t-test, alpha = 0.05, p-value = 0.0235 for infAP and 2.72E-11 for infNDCG).

<Table 10> Mean infAP and infNDCG scores with the thresholds for TP - 0.1 and WP - 0.3 for the top 2 retrieved documents

no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0178	**0.0231**	0.0203	0.0213	0.0176	0.0175	0.0209	0.0172	**0.024**	0.02
infNDCG	0.149	0.1764	0.1663	**0.1816**	0.161	0.1722	0.1683	0.1649	0.1796	**0.1833**
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0202	**0.0215**	**0.0212**	0.0139	0.0194	**0.0219**	**0.0232**	0.0202	0.0204	0.0183
infNDCG	0.1677	0.1769	0.1662	0.1452	0.1624	0.1775	0.1781	0.1756	0.1655	0.1691
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0166	0.0191	**0.0217**	0.0188	0.0189	**0.0213**	0.0193	0.0202	0.0187	**0.0236**
infNDCG	0.1671	0.1573	0.1803	0.1662	0.1567	0.1706	0.169	0.1704	0.1602	0.1802
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0189	0.0203	**0.0234**	0.0201	0.0197	**0.0214**	**0.0238**	0.0188	0.0207	0.0197
infNDCG	0.1533	0.1727	0.1706	0.1739	0.1725	0.1688	0.1862	0.1752	0.1777	0.168

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

3.2 QE using thresholds for TP and (TP * WP)

To find general threshold values for 40 LDA models, the ratio of positive words and negative words was referred rather than infAP and infNDCG scores. For the TP threshold, the ratio of positive words and negative words generated by the thresholds was calculated. TP values were sorted by the ratio for the top 1 retrieved document in the LDA model with 1700 topics (Table 11). In the threshold, 0.08, negative words (315) were generated more than positive words (133) by more than twice.

〈Table 11〉 TP thresholds sorted by no. positive words and no. negative words (1700 topics) for the top 1 retrieved document

TP	No. positive words	No. negative words	No. positive words / (No. negative words +1)
0.08	133	315	0.421
0.09	131	311	0.420
0.07	168	403	0.416
0.06	175	426	0.410
0.14	74	192	0.383
0.13	79	207	0.380
0.05	185	495	0.373
0.11	99	265	0.372
0.12	83	223	0.371

In a similar way, the TP * WP values were sorted by the ratio of positive and negative words for the top 1 retrieved document in the LDA model with 1700 topics (Table 12).

〈Table 12〉 TP * WP thresholds sorted by no. positive words and no. negative words (1700 topics) for the top 1 retrieved document

TP * WP	No. positive words	No. negative words	No. positive words / (No. negative words +1)
0.13	5	5	0.833
0.05	22	26	0.815
0.08	11	13	0.786
0.04	28	35	0.778
0.06	16	20	0.762
0.07	12	17	0.667
0.09	8	12	0.615
0.02	50	81	0.610
0.14	3	4	0.600

TP * WP values showing a high ratio were between 0.02 and 0.14., TP * WP values less than 0.4 looked better because the LDA model with the threshold, 0.04, generated more positive words (28) than the LDA model with the threshold, 0.13 (5).

To improve infAP and infNDCG, two thresholds, 0.08 for TP and 0.03 for (TP * WP) were applied. A maximum of the top 10 words was chosen by the descending order of TP * WP / (document rank)^2. The infAP and infNDCG scores for the top 1 retrieved document were listed in Table 13. For more information, two LDA models with 50 topics and 4800 topics were generated but did not show interesting scores. The mean infAP and infNDCG for the LDA models with 50 topics and 4800 topics were 0.0196 & 0.1759 and 0.0224 & 0.1873, respectively.

〈Table 13〉 Mean infAP and infNDCG scores of 40 LDA models with different thresholds: TP (0.08) and TP * WP (0.03) for the top 1 retrieved document

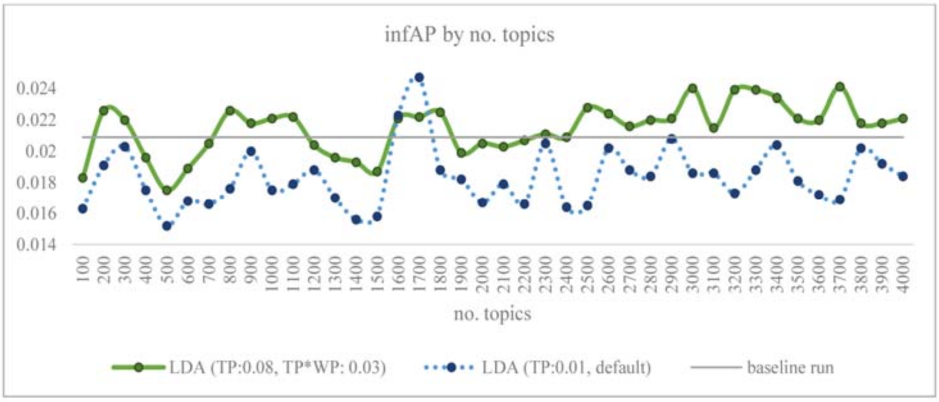
no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0183	0.0226	0.022	0.0196	0.0175	0.0189	0.0205	0.0226	0.0218	0.0221
infNDCG	0.1688	0.1865	0.1917	0.1645	0.1606	0.1806	0.1742	0.1871	0.1837	0.1855
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0222	0.0204	0.0196	0.0193	0.0187	0.0221	0.0222	0.0225	0.0199	0.0205
infNDCG	0.1792	0.1731	0.1745	0.1786	0.1659	0.1875	0.1805	0.1965	0.1793	0.1873
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0203	0.0207	0.0211	0.0209	0.0228	0.0224	0.0216	0.022	0.0221	0.024
infNDCG	0.1859	0.1803	0.1847	0.1876	0.1904	0.1856	0.1811	0.1856	0.1795	0.1945
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0215	0.0239	0.0239	0.0234	0.0221	0.022	**0.0241**	0.0218	0.0218	0.0221
infNDCG	0.1827	0.1885	0.1885	0.1917	**0.1977**	0.1894	0.1955	0.1866	0.1838	0.1848

* baseline run - infAP: 0.0209 and infNDCG: 0.1808

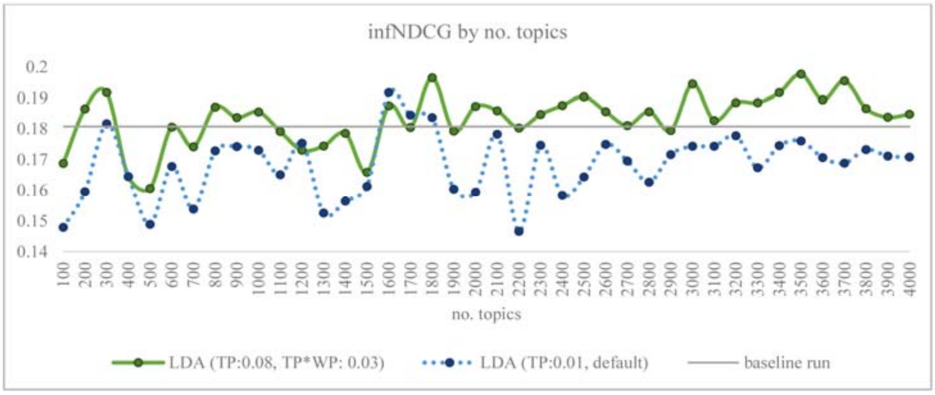
When the mean infAP and infNDCG scores for the top 1 retrieved document were compared with the mean scores of the baseline run, the average mean infAP and infNDCG scores of 40 LDA models were higher: 0.0214 and 0.183, respectively.

The LDA models with large numbers of topics showed better performance (Figure 10 & 11). Compared with the score of the baseline run, the LDA models with more topics than 2200 showed higher infAP scores. Meanwhile, the LDA model with smaller topics than 2300, 9 LDA models showed higher infAP scores than the score of the baseline run, but 13 models showed lower infAP scores. For LDA models with more topics than 2000, most LDA models showed higher infNDCG scores than the score of the baseline run, even though 2 LDA models with 2200 topics (0.1803) and 2900 topics (0.1795) showed lower infNDCG scores. Of the LDA models with 2000 or smaller numbers of topics than 2000, 8 LDA models showed higher infNDCG scores and 12 models showed lower infNDCG scores. There is statistically significant difference in the average mean infAP score in the two-sample t-test (alpha = 0.05, p-value = 0. 0335 for infAP), but not for infNDCG (p-value = 0.0712).

Also, the mean infAP and infNDCG scores of the LDA models with the thresholds (TP:0.08 & TP*WP: 0.03) were compared with the scores of 40 LDA models with only the default TP threshold value (0.01) in Figure 8 and 9. There were improvements in the LDA models the thresholds (TP:0.08 & TP*WP: 0.03): 38 LDA models for infAP and 36 LDA models for infNDCG (Figure 7 & 8). There were statistically significant differences of the average mean infAP and infNDCG scores between two groups in the paired t-test (alpha = 0.05, p-value = 3.3E-13 for infAP and 7.7E-13 for infNDCG).



<Figure 7> Mean infAP scores of 40 LDA models with different thresholds for TP and TP*WP for the top 1 retrieved document



<Figure 8> Mean infNDCG scores of 40 LDA models with thresholds for TP and TP*WP for the top 1 retrieved document

One reason why infAP and infNDCG scores were not that high in 40 LDA models, might be that the threshold values for TP, WP, and TP * WP were optimized for a specific model (with 1700 topics). The ideal TP values would be different depending on individual LDA models, therefore, TP values would be standardized or normalized to be compared between models.

Two thresholds for TP (0.08) and TP*WP (0.03) were effective in increasing infAP for the top 2 retrieved documents (Table 14), while the average mean infNDCG score was lower than that the score of the baseline run. The average mean infAP and infNDCG scores were 0.0217 and 0.1804, respectively. The optimized threshold values would be found in a similar way to top 1 retrieved document.

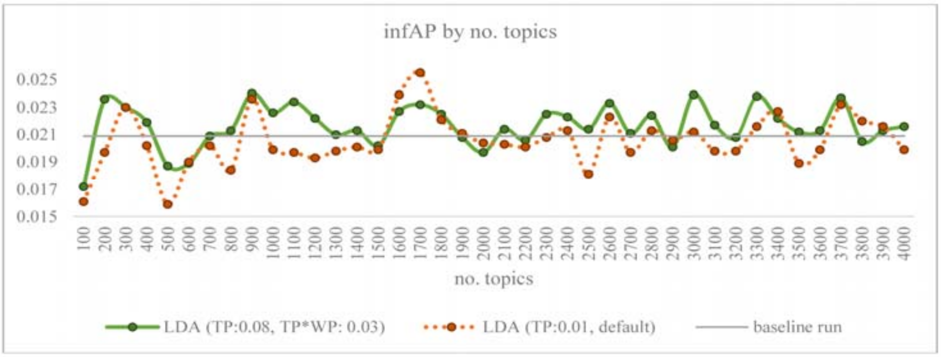
Figure 9 and 10 shows better mean infAP and infNDCG scores for the top 2 retrieved documents in 30 and 26 LDA model, respectively, compared with the scores of the LDA model with only one threshold for TP (0.01). There were statistically significant differences in mean infAP and infNDCG scores (paired t-test, alpha = 0.05, p-value = 0.00002 for infAP, 0.014 for infNDCG) between the LDA model with two thresholds (TP, 0.08 and TP * WP, 0.03) and the LDA model with the threshold (TP: 0.01, Table 2).

Compared with the baseline run, LDA models with two thresholds for TP (0.08) and TP*WP (0.03) showed a statistically significant difference in the average mean infAP score, but not in the average mean infNDCG score (two-sample t-test, alpha = 0.05, p-value = 0.0022 for infAP and 0.7341 for infNDCG).

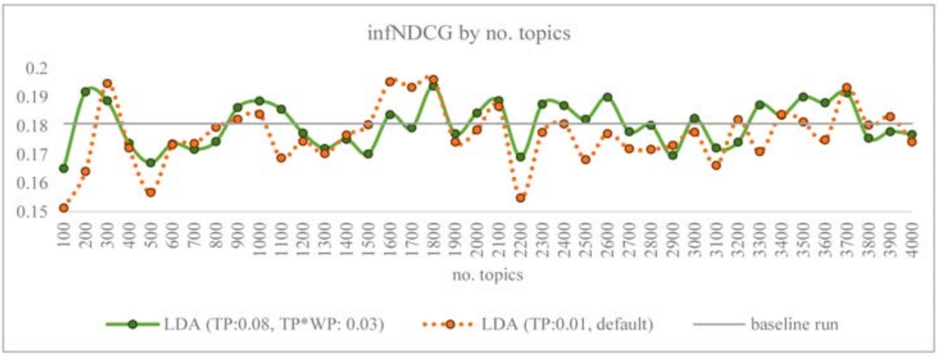
〈Table 14〉 Mean infAP and infNDCG scores of 40 LDA models with the thresholds: TP (0.08) and TP * WP (0.03) for the top 2 retrieved documents

no. topics	100	200	300	400	500	600	700	800	900	1000
infAP	0.0172	0.0236	0.023	0.0219	0.0187	0.0189	0.0209	0.0213	0.024	0.0226
infNDCG	0.1651	0.1919	0.1888	0.1739	0.1671	0.1737	0.1718	0.1745	0.1864	0.1887
no. topics	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
infAP	0.0234	0.0222	0.021	0.0213	0.0202	0.0227	0.0232	0.0225	0.0208	0.0197
infNDCG	0.1858	0.1775	0.1721	0.1753	0.1702	0.1839	0.1793	**0.1939**	0.1773	0.1845
no. topics	2100	2200	2300	2400	2500	2600	2700	2800	2900	3000
infAP	0.0214	0.0206	0.0225	0.0223	0.0214	0.0233	0.0211	0.0224	0.0201	0.0239
infNDCG	0.1888	0.1691	0.1876	0.1871	0.1823	0.19	0.178	0.1802	0.1697	0.1826
no. topics	3100	3200	3300	3400	3500	3600	3700	3800	3900	4000
infAP	0.0217	0.0208	**0.0238**	0.0222	0.0212	0.0213	0.0237	0.0205	0.0213	0.0216
infNDCG	0.1723	0.1743	0.1873	0.1837	0.1901	0.1881	0.1916	0.1757	0.178	0.177

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)



〈Figure 9〉 Mean infAP scores of 40 LDA models with different thresholds for TP and TP*WP for the top 2 retrieved documents



〈Figure 10〉 Mean infNDCG scores of 40 LDA models with different thresholds for TP and TP*WP for the top 2 retrieved documents

4. Summary of Results

The average mean infAP and infNDCG scores of the QE models using the LDA models with different threshold values were listed with p-values calculated in two-sample t-tests, comparing with the baseline run (Table 15). The improved results showing a significant difference (alpha = 0.05) are in bold.

The thresholds for TP (0.1 and 0.08), WP (0.03 and 0.3), and TP * WP (0.03) were applied for the top 1 retrieved document based on an LDA model with 1700 topics. High infAP and infNDCG scores were observed, such as 0.0277 (TP: 0.15 & WP: 0.02) for infAP and 0.1963 (TP: 0.07 & WP: 0.03) for infNDCG. However, because the threshold values were chosen on a specific condition including an LDA model with a specific number of topics (1700) and top 1 retrieved document, they were not effective when applied to other LDA models with different numbers of topics and different numbers of top retrieved documents (e.g. top 2).

<Table 15> Average mean infAP and infNDCG scores of the LDA models with different thresholds for TP, WP, or TP * WP for the top 1 and top 2 retrieved documents

Docs ranked	TP	WP	TP*WP	Ave (mean infAP)	Ave (mean infNDCG)	p-value (infAP)	p-value (infNDCG)
top 1	0.01	-	-	0.0183	0.1684	6.98E-13	5.88E-11
top 2	0.01	-	-	0.0206	0.1768	0.2766	0.0167
top 1	0.1	0.03	-	0.0188	0.1633	2.20E-06	3.33E-13
top 1	0.1	0.3	-	**0.0213**	0.1819	**0.0135**	0.2813
top 2	0.1	0.3	-	0.0201	0.1696	0.0235	2.72E-11
top 1	0.08	-	0.03	**0.0213**	0.1819	**0.0335**	0.0712
top 2	0.08	-	0.03	**0.0217**	0.1804	**0.0022**	0.7341

* baseline run (infAP: 0.0209 and infNDCG: 0.1808)

Although LDA models with specific thresholds for TP, WP, and TP*WP showed overall better mean infAP and infNDCG scores than the scores of 40 LDA model with the default threshold for TP (0.01), the IR performance of each LDA model was not always better in comparison with the baseline run. There were two pairs of thresholds increasing infAP: 1) TP: 0.1 & WP: 0.3 for the top 1 retrieved document, 2) TP: 0.08 & TP * WP: 0.03 for the top 1 and top 2 retrieved documents. Three average mean infAP scores of 40 LDA models were statistically significantly better than the infAP score of the baseline run (in bold).

To find more general thresholds, the optimized thresholds from several LDA models based on different conditions (e.g. different numbers of topics and different numbers of retrieved documents) would be compared.

Ⅵ. Conclusion & Limitations

LDA topic models generated topic words (MeSH terms) using a query or documents retrieved by the query. Because generated topic words include many irrelevant words for QE, selecting relevant words is the key point to increase the IR performance. The results showed that setting up thresholds for topic probability (TP), word probability (WP), or (TP * WP) can filter out negative words for QE. The applications of several LDA models resulted in high infAP and infNDCG scores, such as 0.0277 (TP: 0.15 & WP: 0.02) for infAP and 0.1963 (TP: 0.07 & WP: 0.03) for infNDCG.

However, one limitation is that threshold values can vary in individual LDA models according to the number of LDA topics. Although thresholds values for filtering words were effective to increase infAP and infNDCG scores on several individual LDA models, optimized thresholds for an individual LDA model did not function well in other LDA models with different numbers of topics. The development of the solution to select more relevant words of the candidate LDA topic words would be helpful in improving IR performance regardless of the number of LDA topics. In addition to adjusting threshold values, the applications of classifiers or external terminologies, thesauri, or ontologies would be effective to identify relevant words for QE.

References

Azad, H. K. & Deepak, A. (2019). Query expansion techniques for information retrieval: a survey. Information Processing & Management, 56(5), 1698-1735.

Beaulieu, M., Gatford, M., Huang, X., Robertson, S., Walker, S., & Williams, P. (1997). Okapi at TREC-5. Nist Special Publication SP, 143-166.

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

Bompada, T., Chang, C. C., Chen, J., Kumar, R., & Shenoy, R. (2007, July). On the robustness of relevance measures with incomplete judgments. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 359-366.

Buckley, C., & Voorhees, E. M. (2004, July). Retrieval evaluation with incomplete information. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 25-32.

Carpineto, C. & Romano, G. (2012). A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR), 44(1), 1-50.

Chang, Y., Ounis, I., & Kim, M. (2006). Query reformulation using automatically generated query concepts from a document space. Information Processing & Management, 42(2), 453-468.

Díaz-Galiano, M. C., García-Cumbreras, M. Á., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2007, September). Integrating mesh ontology to improve medical information retrieval. In Workshop of the Cross-Language Evaluation Forum for European Languages. Springer, Berlin, Heidelberg, 601-606.

Efthimiadis, E. N. (1996). Query expansion. Annual Review of Information Science and Technology (ARIST), 31, 121-87.

Harris, Z. S. (1954). Distributional structure. Word, 10(2/3), 146-62.

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 50-57. ACM.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In Advances in Neural Information Processing Systems, 856-864.

Jian, F., Huang, J. X., Zhao, J., He, T., & Hu, P. (2016, July). A simple enhancement for ad-hoc information retrieval via topic modelling. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 733-736.

Liu, H. & Singh, P. (2004). ConceptNet－a practical commonsense reasoning tool-kit. BT Technology Journal, 22(4), 211-226.

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

Lupu, M., Zhao, J., Huang, J., Gurulingappa, H., Fluck, J., Zimmermann, M., ··· & Tait, J. (2011, November). Overview of the TREC 2011 Chemical IR Track. In TREC.

Merabti, T., Letord, C., Abdoune, H., Lecroq, T., Joubert, M., & Darmoni, S. J. (2009). Projection and inheritance of SNOMED CT relations between MeSH terms. In MIE, 233-237.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

Mitchell, P. C. (1973). A note about the proximity operators in information retrieval. ACM SIGPLAN Notices, 10(1), 177-180.

Mu, X., Lu, K., & Ryu, H. (2014). Explicitly integrating MeSH thesaurus help into health information retrieval systems: an empirical user study. Information Processing & Management, 50(1), 24-40.

Munro, R. J., Bolanos, J. A., & May, J. (1978). LEXIS vs. WESTLAW: an analysis of automated education. Law Libr. J., 71.

Natsev, A., Haubold, A., Tešić, J., Xie, L., & Yan, R. (2007, September). Semantic concept-based query expansion and re-ranking for multimedia retrieval. In Proceedings of the 15th ACM International Conference on Multimedia, 991-1000.

Paik, J. H. (2013, July). A novel TF-IDF weighting scheme for effective ranking. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 343-352.

Roberts, K., Simpson, M. S., Voorhees, E. M., & Hersh, W. R. (2015, November). Overview of the TREC 2015 Clinical Decision Support Track. In TREC.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., & Hersh, W. R. (2016, November). Overview of the TREC 2016 Clinical Decision Support Track. In TREC.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A. J., & Pant, S. (2017, November). Overview of the TREC 2017 Precision Medicine Track. In TREC.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Voorhees, E. M. (2014, July). The effect of sampling strategy on inferred measures. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, 1119-1122.

Wang, Y., Huang, H., & Feng, C. (2017, April). Query expansion based on a feedback concept model for microblog retrieval. In Proceedings of the 26th International Conference on World Wide Web, 559-568

Xu, J. & Croft, W. B. (2017, August). Quary expansion using local and global document analysis. In Acm Sigir Forum. New York, NY, USA: ACM, 51(2), 168-175.

Yanagawa, A., Chang, S. F., Kennedy, L., & Hsu, W. (2007). Columbia university's baseline detectors for 374 lscom semantic visual concepts. Columbia University ADVENT Technical Report, 222-2006.

Yilmaz, E. & Aslam, J. A. (2006, November). Estimating average precision with incomplete and imperfect judgments. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, 102-111. ACM.

Yilmaz, E., Kanoulas, E., & Aslam, J. A. (2008, July). A simple and efficient sampling method for estimating AP and NDCG. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 603-610. ACM.

Zeng, Q. T., Redd, D., Rindflesch, T., & Nebeker, J. (2012). Synonym, topic model and predicate-based query expansion for retrieving clinical documents. In AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2012, 1050.

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

Zhou, D., Wu, X., Zhao, W., Lawless, S., & Liu, J. (2017). Query expansion with enriched user profiles for personalized search utilizing folksonomy data. IEEE Transactions on Knowledge and Data Engineering, 29(7), 1536-1548.