Pan *et al. BMC Medical Informatics and Decision Making* 2019, **19**(Suppl 9):251
https://doi.org/10.1186/s12911-019-0986-6

BMC Medical Informatics and Decision Making

RESEARCH Open Access

An adaptive term proximity based rocchio's model for clinical decision support retrieval

Min Pan[1,2], Yue Zhang[3], Qiang Zhu[3], Bo Sun[1], Tingting He[3*] and Xingpeng Jiang[3]

From 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference
Wuhan and Shanghai, China. 15-18 August 2018, 3-4 November 2018

Abstract

Background: In order to better help doctors make decision in the clinical setting, research is necessary to connect electronic health record (EHR) with the biomedical literature. Pseudo Relevance Feedback (PRF) is a kind of classical query modification technique that has shown to be effective in many retrieval models and thus suitable for handling terse language and clinical jargons in EHR. Previous work has introduced a set of constraints (axioms) of traditional PRF model. However, in the feedback document, the importance degree of candidate term and the co-occurrence relationship between a candidate term and a query term. Most methods do not consider both of these factors. Intuitively, terms that have higher co-occurrence degree with a query term are more likely to be related to the query topic.

Methods: In this paper, we incorporate original HAL model into the Rocchio's model, and propose a new concept of term proximity feedback weight. A HAL-based Rocchio's model in the query expansion, called HRoc, is proposed. Meanwhile, we design three normalization methods to better incorporate proximity information to query expansion. Finally, we introduce an adaptive parameter to replace the length of sliding window of HAL model, and it can select window size according to document length.

Results: Based on 2016 TREC Clinical Support medicine dataset, experimental results demonstrate that the proposed HRoc and HRoc_AP models superior to other advanced models, such as PRoc2 and TF-PRF methods on various evaluation metrics. Among them, compared with the Proc2 and TF-PRF models, the MAP of our model is increased by 8.5% and 12.24% respectively, while the F1 score of our model is increased by 7.86% and 9.88% respectively.

Conclusions: The proposed HRoc model can effectively enhance the precision and the recall rate of Information Retrieval and gets a more precise result than other models. Furthermore, after introducing self-adaptive parameter, the advanced HRoc_AP model uses less hyper-parameters than other models while enjoys an equivalent performance, which greatly improves the efficiency and applicability of the model and thus helps clinicians to retrieve clinical support document effectively.

Keywords: Clinical retrieval, Term proximity, Query expansion, Pseudo relevance feedback

*Correspondence: tthe@mail.ccnu.edu.cn
[3]School of Computer, Central China Normal University, 430079 Wuhan, China
Full list of author information is available at the end of the article

Background

Introduction and motivation

Retrieving more relevant articles for the clinician is helpful to improve their decision-making on the diagnosis, treatment and test of patients [1].The TREC Clinical Decision Support (CDS) tracks provides corpora for such systems and encourages the information retrieval tools and resources needed to implement these systems. The real hospital notes contain a lot of abbreviations and other language styles[1], and extracted by clinicians as queries. A note is usually longer than a summary or a description. All of those will bring new challenges for traditional retrieval systems. Thus, a clinician rewrites his or her notes in a standard report unreasonably. Instead, to reduce the time burden of clinician, the task of information retrieval system should be to operate notes directly. We retrieve full-text biomedical articles for answering questions related to one of three generic clinical information needs: Diagnosis, Test and Treatment.

To address the challenge, we use Pseudo Relevance Feedback (PRF) technique. PRF is a very famous query extension technology [2–6]. It assumes that the first retrieval top-ranked documents are relevant to the query. Then PRF refines the potentially related terms weight and adds to the original queries. Although the traditional PRF has been proved to be very effective [7, 8], it still fail in some classic IR tasks. The expansion terms are selected according to the candidate term frequency or the term distributions of the feedback documents are irrelevant.

Integrating term proximity information into PRF can improve retrieval efficiency and become a hot spot of research [9]. The Hyperspace Analogue to Language (HAL) [10] is a mental theory of term meaning calculation model, only considering the context of words, immediately surrounded by a given word. Most of these researches focus on the term proximity in the original query and apply it to the sorting documents [11–18]. However, most traditional term proximity methods ignore the significance of term frequency.

The main contributions of this paper are as follows:

Adapting a new concept of proximity information weight, we propose a proximity based PRF model, called HRoc.

Introducing three normalization methods to make a fair comparison.

Adapting the adaptive function to make the length of sliding window (denote by D value) of HAL model dynamically adjust according to the length of the document and increasing the universality of the model.

Our proposed model has been proved to be effective by TREC clinical medicine collections.

[1] http://www.trec-cds.org/

Related work

The CDS track complements the previous TREC tasks inspired by biomedicine [1], specifically, the genomics and medical records tracks. The CDS track has been heavily inspired by the TREC genomics [19], medical records [20] tracks and the medical case-based retrieval track of Image-CLEF [21]. They all shown great interest in medical ad-hoc retrieval. There are no reusable, uncertified set of medical records, in [22, 23], short case reports, proposed that the real medical records should be represented by idealized method. For a given case report, follow-up participants retrieve full-text biomedical articles and answer questions related to several types of clinical information needs. The 2016 CDS focuses on topic query expansion modeling by actual patients note [24–26].

In Information Retrieval (IR) process, original queries may lead to the absence of some important terms information. PRF method is a common but effective technique for achieving better retrieval performance in [2, 7, 8, 27], which the semantic relationship between the added terms and the original query terms is considered, including these defined relations in Rocchio's model. It brings better result. Then, many other relevance feedback techniques and algorithms were proposed and most of them were derived under the Rocchio's framework. For example, A famous and successful automatic PRF algorithm is proposed in okapi system by Robertson et al [28]. A feedback framework based on proximity (called proc) is proposed by Miao et al, which includes different proximity measures to estimate the correlation and importance of candidate options [29]. Ye and Huang propose a unified model (TF-PRF) to capture local saliency of related candidate in feedback documents [30]. These two models are strong baselines, and used for comparison in our experiments. In addition, many other competitive approaches have obtained significant performance in improving retrieval effectiveness [4, 31]. Since they are not that related to our research methods, we do not introduce them in detail.

Recently, plenty of work has been studied to integrate term proximity and other relationships into existing IR models. In [32], the authors introduced a pseudo term to the model in the Dirichlet language model, the approximate centrality of query terms is used as a parameter. LV and Zhao integrated location and proximity information into the language model from a second perspective [33]. These relations play an important role in IR field. Mbarek et al have obtained significant performance in improving retrieval effectiveness [34]. Rasolofo et al use proximity measurement in combination with the Okapi probabilistic model [17]. Peng et al incorporate term dependency in the DFR framework [35]. Metzler et al developed a general and formal framework for modeling term dependency through Markov random fields, and developed a

new method to train the model, which directly maximizes the average accuracy rather than the availability of training data. [36]. Zhao et al use Triangle Kernel functions in information retrieval applications [37]. In this paper, we propose three HAL-based co-occurrence PRF models, in which we integrate the approximate weight information of a term into the traditional PRF model: Rocchio model. In our method, we estimate their weights by considering the distance between the candidate expansion and the query item. In addition, we introduce three normalization methods for a new concept of proximity-based term weighting and an adaptive function to make the D value dynamically adjust according to the length of the document.

Methods

Traditional pRF models

Our study based on a classic traditional PRF model. In this section, we will first briefly revisit the Rocchio's model. The Rocchio's model is a classical framework to realize pseudo relevance feedback representation, which incorporates the information of pseudo relevance feedback in the first-pass retrieval [27]. In PRF, the feedback documents often contain relevant and irrelevant documents, but the irrelevant in the Rocchio equation is ignored. Finally, the query is realized by the linear combination of the initial query vector and the feedback document vector.

$$Q_1 = a * Q_0 + \beta * \Sigma_{r \in R} \frac{r}{|R|} \tag{1}$$

where Q_0 and Q_1 represent the original and new query vectors respectively, $|R|$ represents the number of the feedback documents, r is the expansion terms weight vector for feedback documents, α and β represent the original query weight and the related documents weight respectively. In Eq. (1), we can notice that α and β are actually constant values, which control how much we rely on the original query and the feedback information. In practice, we can always set α at 1.0, and only study β until we get better performance.

However, traditional Rocchio's model does not capture the relationship between an original query term and a candidate term. Generally, a candidate term is supposed to be relevant to the query topic if it occurs near to a query term.

HAL method for term co-occurrence weight

The basic motivation is that when a person encounters a new concept, he or she is inclined to infer its meaning from other concepts occurring within the same context. For example, a document that contains both "heart disease" and "China" is irrelevant to the topic "heart disease in China" when these two terms are not close to each other in the context. Therefore, term proximity is effective to discriminate against these types of documents. Vechtomova et al. using multiple distance factors and mutual information to select query extension from Windows Environment [38]. HAL model constructs a high-dimensional vector for each word by simply treating the context of a given query word as a close word [10, 39, 40]. HAL method begins by producing a co-occurrence matrix $|V| * |V|$ for each term in a specified vocabulary $|V|$. This process of counting local co-occurrences is illustrated in Fig. 1.

It is assumed that the length of a document is 18, and the D value is set to 5. For each word a, a proximity relation can be generated between a and every word b which occurs close to a. Then their distance strength can be calculated as follows. If B occurs adjacent to A, the strength is 5. Then if B and A are separated by a word, it would get the strength of 4, and it also drop to the intensity of 1. The element $W_{a,b}$ (row A, column B) of the symbiotic matrix contains the weighted sum of all occurrences of B close to A. The co-occurrence matrix contains after HAL-style weighting of the counts from the sliding window in Fig. 2.

The weighting of each co-occurred term is accumulated over the whole corpus. We adapt the original HAL model similar as in [39, 40]. Then, the weight calculation of each term can be represented by a semantic vector within a specified distance.

$$HAL(t,q) = \sum_{l=1}^{|D|} w(l) * p(t,l,q) \tag{2}$$

where l is the distance from query term q to term l, $p(t,l,q)$ is the co-occurrence frequency within the sliding windows when the distance equals l, then $w(l) = D - l + 1$ denotes the strength. In this paper, we need to construct vector for term in document, which denotes a proximity relationship with the entire query. Intergrading the average proximity information into the query term weight could make a better result. Then we also take into account

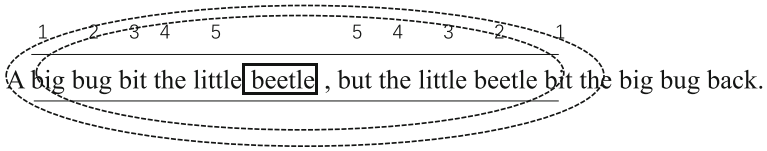

Fig. 1 An example for counting local co-occurrences

	a	big	bug	the	little	but	beetle	bit	back
a	0	5	4	2	1	0	0	3	0
big	5	0	10	8	4	0	4	8	4
bug	4	10	0	8	4	0	4	8	5
the	2	8	8	0	8	3	8	10	3
little	1	4	4	13	1	3	10	8	0
but	0	0	0	7	7	0	7	3	0
beetle	0	4	4	11	11	4	1	8	1
bit	3	8	7	12	9	3	8	0	1
back	0	4	5	3	0	0	1	2	0

Fig. 2 The co-occurrence matrix contain after HAL-style weighting of the counts from the sliding window

the distinction factor of different query terms. Therefore, each dimension is the specific representation vector weight of the query term, which can be calculated by the inverse document frequency formula below, and then the proximity-based term weight $W_{HAL}(t,q)$ in the method is computed as follows:

$$W_{HAL}(t,q) = \sum_{i=1}^{|Q_0|} HAL(t,Q_0) * IDF(q_i) \tag{3}$$

The weighted HAL model combines the two elements of term distance and co-occurrence frequency at the same time. It is the first time that the HAL model was adopted to measure the proximity between the query terms and the candidate terms in such way, and then used in the field of medical retrieval.

A hAL-based pRF model

We take into account the significance of TF-IDF and proximity information, hence propose a HAL-based co-occurrence model for PRF via integrating a term's co-occurrence information into traditional PRF models for expansion terms selection. Formally, let $Q_0 = \{q_1, q_2, \ldots, q_m\}$ represents the original query given by the user, a new query Q' will be generated by a method HRoc as follows:

$$Q' = (1-\alpha)*Q_0+\alpha*\left((1-\beta)*\Sigma_{r\in R}\frac{r}{|R|}+\beta*\Sigma_{r'\in R}\frac{r'}{|R|}\right) \tag{4}$$

where α and β are tuning coefficients of 0 to 1.0. α is constant value for tuning the contribution weight between the original query and the feedback information, and β is constant value for balancing the contribution weight between the feedback information measured through common term frequency or term distributions and the feedback information measured through the corresponding term co-occurrence.

In Eq. (4), Q', Q_0 and $|R|$ have the same meanings with those in Eq. (1). Parameter r is the vector of expansion term weight computed with BM25, and r' is the vector of expansion term proximity co-occurrence weight for feedback documents, which reveals the relationship between a candidate term and a query topic.

$$Q' = (1-\alpha)*Q_0+\alpha*\left((1-\beta)*\Sigma_{r\in R}\frac{r}{|R|}+\beta*\Sigma_{r'\in R}\frac{W_{HAL}(t,q_i)}{|R|}\right) \tag{5}$$

In order to better compare with advanced model, we design three normalization methods of term proximity weight in next part.

Normalization methods

Normalization is convenient for data processing. We get weight score ranking in Eq. (5). Due to the big difference between the multiplicative values, the effect is certainly not good if it is directly integrated into the Rocchio's model. To solve this problem, we adopt three normalization methods to optimize the weight score. The formula representation after the introduction of normalization is shown in Eq. (6):

$$Q' = (1-\alpha)*Q_0+\alpha*\left((1-\beta)*Norm\left(\Sigma_{r\in R}\frac{r}{|R|}\right)+\beta*Norm\left(\Sigma_{r'\in R}\frac{W_{HAL}(t,q_i)}{|R|}\right)\right) \tag{6}$$

Three normalization methods are presented in Table 1.

We take methods of $norm_1(t)$, $norm_2(t)$ and $norm_3(t)$ to process data respectively and t represents the different weight values in Eq. (6). We call them HRoc1, HRoc2 and HRoc3 by using three normalization methods. In our experiment, we use different normalization methods to make a comparison in next part.

An adaptive term proximity normalization model

In the HRoc model, the Mean Average Precision (MAP) of the retrieval results is closely related to the D value. The traditional HAL model uses a fixed window size, so we need to make a large number of experiments to find the optimal D value and to improve the value of the MAP. While, this process wastes a lot of time and resources. In order to solve this problem, we need to make the model automatically find the most appropriate D value. In fact, a

Table 1 The three kinds of normalization method

$Norm(t)$		
$norm_1(t)$	$norm_2(t)$	$norm_3(t)$
$t - min(t)/max(t) - min(t)$	$t/\sqrt{sum(t^2)}$	$t/max(t)$

large number of experimental results show that the optimal D value is affected by the length of the document. For this reason, we try to use three different functions to fit the relationship between the D value and the length of the document, as is depicted in Table 2. The specific experimental results and analysis are given in next chapter.

Table 2 The three adaptive function

$\|D\|$		
$f_1(dl)$	$f_2(dl)$	$f_3(dl)$
dl	$dl * (1 + dl/avg(dl))$	$dl + avg(dl)/1 + \log_2(dl/avg(dl))$

Test collections and evaluation metrics

In order to validate the effectiveness of our proposed model, we conduct a series of experiments on the standard TREC Clinical Decision Support Track collections. The document collection was updated to a more recent snapshot of PubMed Central (PMC)[2] from 730k to 1.25 million full-text clinical medicine articles. The PMC collection contains articles published by PubMed Central in the year of 2016, including pmc-00, pmc-01, pmc-02 and pmc-03. We use pmc-00 and pmc-01 as test collections, pmc-02 and pmc-03 as experimental collections respectively. The topic numbers associated with each collection are presented in Table 3, and "No. of Doc" denotes the number of documents, "No. of Queries" means topic numbers.

For the collections, each topic contains three fields (title, description and narrative), and the example is introduced as follows in Fig. 3.

In the process of indexing and querying, queries without judgments are removed. For all test sets used, each term is stemmed by using Porters English stemmer. We leverage the Mean Average Precision (MAP) for the top 1000 documents to measure model performance in our experiments[36]. We take this metric as the primary evaluation metric in our experiments, which is typically used as the main official metric in the corresponding TREC evaluations. In addition, P@k can evaluation the relevance degree of the top-ranking documents, and we set $k \in \{5, 10, 20\}$. F1-Value can be represented as $2 * PR/(P + R)$, which is the harmonic average of the Recall (R) rate and the Precision (P) rate. Statistically significant evaluation metrics values based on the two-tailed paired p_value are computed at a 95% confidence level.

Baseline models

Baseline models are very critical to verify the performance of the proposed methods. Firstly, we compare the proposed methods with the basic retrieval model BM25 and the KL-divergence language modeling (LM) retrieval [7, 28]. Okapi BM25 is one of the most classical text retrieval algorithms. It is a probability weighted model. In BM25, the number of query term appear in documents (also called frequency) and the number of documents containing the query term are defined to assign weights. The corresponding weighting functions are as follows:

$$w(q_i, D) = \frac{(k_1 + 1) * tf(q_i, D)}{K + tf(q_i, D)} * \frac{(k_3 + 1) * qtf}{k_3 * qtf} * log\left(\frac{N - df + 0.5}{df + 0.5}\right) \quad (7)$$

where $w(q_i, D)$ is the weight of query term qi in a document D, and N is the number of indexed documents in the collection.$k1$ and $k3$ are tuning constants that depend on the dataset used and possibly on the nature of the queries. K is equal to $k1 * ((1 - b) + b * dl/avdl)$, where dl is the length of the document, and $avdl$ is the average document length. df is the number of documents containing a specific term. $tf(q_i, D)$ is the within-document term frequency, and qtf is the within-query term frequency.

Language model method is another classical algorithm in traditional information retrieval. Its basic idea is to estimate a language model for each document, and then sort the documents according to the possibility of query in the estimated language model. In particular, for the basic language model, we use a Dirichlet prior (with a hyper-parameter of μ') to smooth the document language model as shown in Eq. (8) which generally results in better performance.

$$p(w|D) = \frac{c(w_D) + \mu' * p(w|c)}{|D| + \mu'} \quad (8)$$

where $c(w_D)$ is the frequency of query term w in document D , $p(w|c)$ is the occurrence probability of term w in collection C, and $|D|$ is the length of document D.

The primary estimation of relevance model is often called RM1 [7, 41]. Essentially, RM1 uses the query likelihood $P(Q|D)$ as the weight for document D and takes an average of the probability of each word given by each document language model. The relevance model $P(w|Q)$ is commonly used to estimate the feedback language model θ_F. And in order to improve performance [7, 41], θ_F is generally interpolated into the original query model θ_Q. However, it only captures the candidate terms' distribution, but neglects the co-occurrence distribution between

Table 3 Collections statistics

Collection		No. of Doc	Size	No. of Queries
PMC	pmc-00	263175	16.9GB	30
	pmc-01	240347	15.8GB	30
	pmc-02	389431	21.2GB	30
	pmc-03	357047	19.6GB	30

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

<title> A 75F found to be hypoglycemic with hypotension and bradycardia.

<desc> Description:

A 75F with a PMHx significant for severe PVD, CAD, DM, and CKD presented after being found down unresponsive at home. She was found to be hypoglycemic to 29 with hypotension and bradycardia. Her hypotension and confusion improved with hydration. She had a positive UA which eventually grew klebsiella. She had temp 96.3, respiratory rate 22, BP 102/26, a leukocytosis to 18 and a creatinine of 6 (baseline 2).

<narr> Narrative

Pt is a 75F with a PMHx significant for severe PVD, CAD, DM, and CKD who presented to [**Hospita11 * *H* *Location (un) 1375**] on [**6-25**] after being found down unresponsive at home. She was found to be hypoglycemic to 29 with hypotension and bradycardia. Her hypotension and confusion improved with hydration. She had a positive UA which eventually grew klebsiella, treated initially with levofloxacin. She had a leukocytosis to 18 and a creatinine of 6 up from presumed prior baseline of N2. On morning of transfer, pt had blood cultures result 3/3 bottles positive for GAS, her antibiotics were switched to vancomycin which was then changed to ceftriaxone. Her blood pressure dropped to the 60s. She was given a bolus of bicarb and transfered to their ICU. After an additional bolus of 500cc she was started on levophed. She was anuric throughout the day. She had a midline placed on right side. She received 80mg IV solumedrol this morning in the setting of low BPS and rare eos in urine. On arrival to the MICU pt was awake but drowsy. She was receiving levophed throughout her transfer.

Fig. 3 The example of topic style

query terms and expansion terms. We make a comprehensive comparison with BM25+Rocchio and DLM+RM3, and a detailed analysis in our preliminary experiments.

For state-of-the-art PRF models, we compare the methods of PRoc2, PRoc3 and TF-PRF with the proposed method. The method of PRoc2 uses Gaussian kernel that shown to be effective in most cases, while PRoc3 model is similar to our method. Concerning the fact that PRoc2 and PRoc3 are more effective than PRoc1 [29], we would employ the former two methods to compare with the proposed methods in the experiments. Additionally, TF-PRF [30] is proposed by incorporating three different term frequency transformation methods. These two methods are representatives of the state-of-the-art models, which are capable of achieving the best IR performance on most of the standard TREC datasets.

Parameter settings

In our model, several controlling parameters should be tuned for optimal results. Fairly, to find the optimal parameter settings, the following parameter settings for both baselines and the proposed model are used. And the related settings is well-known in IR for establishing strong baselines. First, in BM25, the value of b is swept from 0 to 1.0 with an interval of 0.1, and k_1 and k_3 are set to 1.2 and 8 respectively. Second, the Dirichlet prior smoothing μ ($\mu \in \{500, 600, \ldots, 2000\}$) in language model are then used to retrieve the documents. In other traditional medicine IR research, medicine data sets generally do not allow many candidate expansion terms based on practical application experience. So we sweep the number of feedback documents N and the feedback term $|T_f|$ from $\{10, 20, \ldots, 50\}$. Finally, the HAL parameter D is set as $D \in \{0, 100, \ldots, 2500\}$, and the interpolation parameter α, β are set as $\alpha, \beta \in \{0.0, 0.1, \ldots, 1.0\}$. In addition, we use 2-fold cross-validation to evaluate the proposed approaches, in which the TREC queries on each collection are partitioned into two sets by the parity of their numbers, and then the parameters learned from the training data set are applied to the test data set for evaluation purpose.

Results and discussion

Comparison with pRF basic models

After comparing the proposed methods with essential retrieval model BM25 and KL-divergence language modeling (LM) retrieval, we show the experimental results in Table 4.

As is shown in Table 4, BM25 performs slightly better with Dirichlet prior in terms of MAP and P@10 metrics, and LM is superior to it in the rest of metrics. These two basic models perform comparatively, without observing significant different.

Next, "BM25+Rocchio", the combination of BM25 and Rocchio's model, and relevance language model (RM3) are used as two strong baselines in this paper. These two methods can achieve better retrieval performance in most cases. Using them as the basic models of the PRF baselines is reasonable. Thus, we use them to compare the proposed model.

We demonstrate the results of baseline PRF models and the three proposed models (HRoc1, HRoc2, and HRoc3) with different evaluation metrics in Table 5. In particular, "*" and "+" indicate a statistically significant improvement over BM25+Rocchio and RM3 respectively (Wilcoxon signed-rank test with $p < 0.05$). The bold style in row corresponding to the best result. And Rocchio is used with BM25 and RM3 is used with LM for fair comparison.

Table 5 shows that the average performance of proposed models is better than that of baseline models. First, it has been proved that both Rocchio and RM3 are effective. They are considered to be strong baselines in previous studies. Rocchio model is superior to the RM3 model in terms of P@5 and P@10 metrics, but RM3 model outperforms Rocchio model in terms of MAP, P@20 and F1-value. Second, among the three proposed models, HRoc1 performs better retrieval results than HRoc2 or HRoc3

Table 4 Performance of basic retrieval models in certain metrics

Basic models	MAP	P@5	P@10	P@20	F1
BM25	0.0448	0.2533	0.2467	0.2100	0.0739
DLM	0.0424	0.2800	0.2367	0.2100	0.0786

Table 5 Comparison with baseline PRF retrieval models in certain metrics

	BM25+Rocchio	RM3	HRoc1	HRoc2	HRoc3
MAP	0.0490	0.0540	**0.0651**$^{*+}$ (32.8%,20.5%)	0.0647^{*+} (32.0%,19.8%)	0.0642^{*+} (31.0%,18.9%)
P@5	0.2733	0.2600	**0.2933**$^{*+}$ (7.32%,12.8%)	0.2733^{+} (0.00%,5.11%)	0.2733^{+} (0.00%,5.11%)
P@10	0.2533	0.2467	**0.2733**$^{*+}$ (7.89%,10.8%)	0.2667^{*+} (5.29%,8.11%)	0.2567^{*+} (1.34%,4.25%)
P@20	0.2167	0.2233	**0.2350**$^{*+}$ (8.44%,5.24%)	0.2317^{*+} (6.92%, 3.76%)	0.2317^{*+} (6.92%, 3.76%)
F1	0.0853	0.0932	**0.1112**$^{*+}$ (30.3%,19.3%)	0.1108^{*+} (29.9%,18.9%)	0.1096^{*+} (28.5%, 17.6%)

The values in parentheses represent the improvements over BM25+Rocchio and RM3 respectively. The best result obtained is shown in bold, and the superscripts "*" and "+" denote statistically significant improvements over BM25+Rocchio and RM3, respectively (Wilcoxon signed-rank test with $p < 0.05$)

in terms of all metrics. This outcome proves the general effectiveness of our model. Third, HRoc1 gets the best performance of all the other four models. HRoc1 achieves average improvements of 32.8%, 7.32%, 7.89%, 8.44% and 30.3% over BM25+Rocchio in terms of the five metrics in TREC medicine collection respectively. The proposed HRoc1 model achieves average improvements of 20.5%, 12.8%, 10.8%, 5.24% and 19.3% over RM3 in terms of the five metrics in TREC medicine collection respectively.

Comparison with the recent Progress

Because HRoc1 achieves better retrieval performance than HRoc2 and HRoc3, we just compare HRoc1 model with the state-of-the-art PRF models and proximity based PRF model (PRoc2, PRoc3, TF-PRF) using different evaluation metrics. Table 6 records the corresponding experimental results. In particular, "*", "+", and "#" represents statistically significant improvement over PRoc2, PRoc3, and TF-PRF, respectively. The bold style in row corresponding to the best result.

We can clearly see from Table 6 that the average performance of HRoc1 model is significantly better than other models on whole collections in most cases. PRoc2, Proc3 and TF-PRF have proven effective in previous studies [29, 30]. First, the PRoc2 model outperforms the PRoc3 and TF-PRF model on the MAP, P@5, P@10, P@20 and F1 metrics. Second, the HRoc1 model outperforms PRoc3 and TF-PRF in terms of all metrics. Then PRoc2 gets best result in P@20 metrics among all eight methods. Third, the proposed HRoc1 model achieves average improvements of 8.5%, 2.30%, 7.90% and 30.3% over PRoc2 on the MAP, P@5, P@10 and F1 metrics in TREC medicine collection respectively.

Table 6 Comparison with the state-of-the-art PRF retrieval models in certain metrics

	PRoc2	PRoc3	TF-PRF	HRoc1
MAP	0.0600	0.0593	0.0580	**0.0651**$^{*+\#}$ (8.50%,9.78%,12.24%)
P@5	0.2867	0.2667	0.2600	**0.2933**$^{*+\#}$ (2.30%,9.97%,12.81%)
P@10	0.2533	0.2300	0.2467	**0.2733**$^{*+\#}$ (7.90%,18.83%,10.78%)
P@20	0.2417	0.2167	0.2317	0.2350$^{+\#}$ (-2.77%,8.44%,1.42%)
F1	0.1031	0.1020	0.1012	**0.1112**$^{*+\#}$ (7.86%,9.02%,9.88%)

The best result obtained is shown in bold, and the superscripts "*", "+" and "#" denote statistically significant improvements over PRoc2, PRoc3 and TF-PRF, respectively (Wilcoxon signed-rank test with $p < 0.05$). The values in parentheses are the improvements over RM3, BM25+Rocchio, PRoc2 and TF-PRF, respectively

HRoc1 is obviously superior to PRoc2, PRoc3, and TF-PRF. In terms of P@10, it performs up to 7.9% improvement, which is significantly better than that on MAP. The results demonstrate that our model performs well, especially in applications that emphasize on top results. To conclude, our model is at least comparable to the latest progress in probabilistic model and language model framework in MAP, and can perform significantly better in P@5, P@10, P@20 and F1-value.

Comparison with different d value

In our preliminary experiments when smoothing methods, we introduced D value as a fixed value to measure proximity co-occurrence relationship of original query term and candidate query term. However, D value is also an important parameter in the HRoc model. We segment a document into a list of sliding windows, where each window has a fixed size. In our experiments, set D value varies from 10 to 2500. We use HRoc1, HRoc2, and HRoc3 to obtain the best result. Based on the results in MAP metrics, we get a Precision trend value, which is record in Fig. 4.

As is shown in Fig. 4, both HRoc1 and HRoc2 get their optimal Precision value where $D = 1500$. We can observe a steady and slow decline trend through the two sides of the optimal value. Therefore, the methods of HRoc1 and HRoc2 are relatively stable models. HRoc3 do not obey this trend, so we consider that the $norm_3(t)$ normalization method is not suitable for processing the proximity weight result data of the medical sets.

Comparison with adaptive d value

The value of parameter D is fixed in the HAL model proposed previously, and the HAL weights of each document in the dataset are calculated according to the fixed value. In fact, the lengths of documents in the dataset are

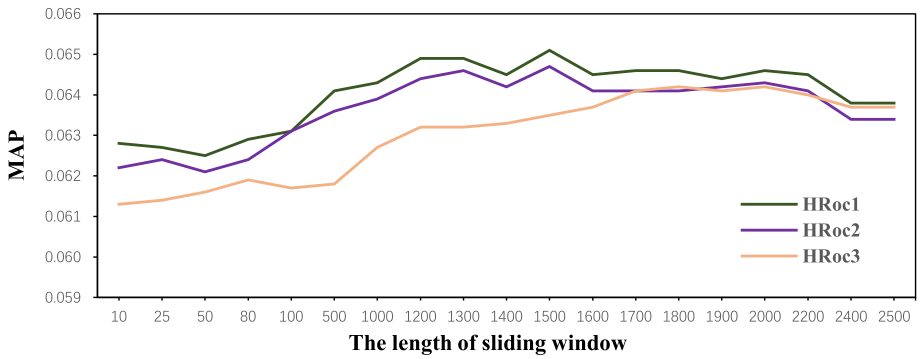

Fig. 4 The performance of HRoc with different D values in MAP

different from each other. In order to better research the relationship between *D* value and the length of the document, we calculate the lengths of the 125 million documents in the dataset, finding that the average length of all the documents is 816.3 and 87.7% of the document lengths are unevenly distributed ranging from 20 to 2500. Details can be found in Fig. 5.

Under this kind of situation, using the same value of the parameter D to calculate the HAL weights for documents with different lengths is unreasonable. Thus, we assume that it might be more appropriate to set different D values for each document. To validate this assumption, we use three different automatic parameters to replace fixed D value. Due to HRoc1 achieves better retrieval performance, the following are three different adaptive functions of HRoc1 named HRoc1_AP1, HRoc1_AP2, and HRoc1_AP3 to make a comparison with the BM25 + Rocchio, TF- PRF and HRoc1 model, and D value of HRoc1 in the range of 100 to 2500. We get the MAP, P@10, P@20 and F1 trend values in Fig. 6 as following.

It is very intuitive to find that the first adaptive function model (HRoc1_AP1) is better than the original HRoc1 model under the evaluation metrics of MAP, P@20 and F1. HRoc1_AP1 is superior to other three models in terms of P@10 at most cases. The parameter D of the adaptive function in HRoc1_AP1 is the length of the document itself, which directly confirms our previous idea that setting different D value for each document. From the comparison results of the experiment, the most suitable setting for D value is document length (dl).

Since the required parameter (dl) is an existing constant when the adaptive function (HRoc1_AP1) is calculated, the algorithm complexity of adaptive function model is not increased comparing to the original model.

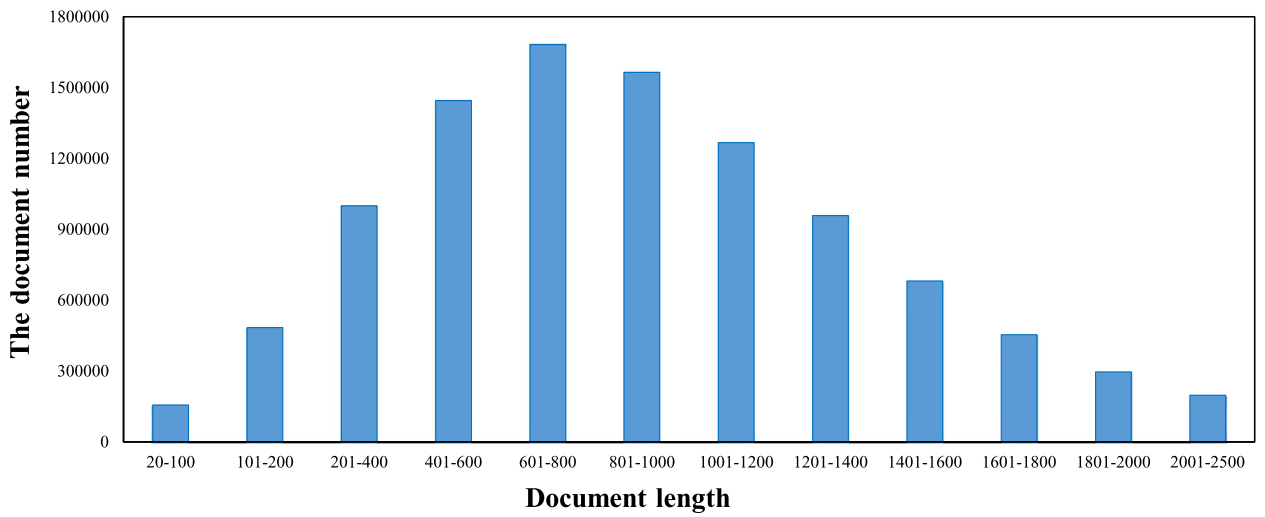

Fig. 5 The statistical distribution of document lengths

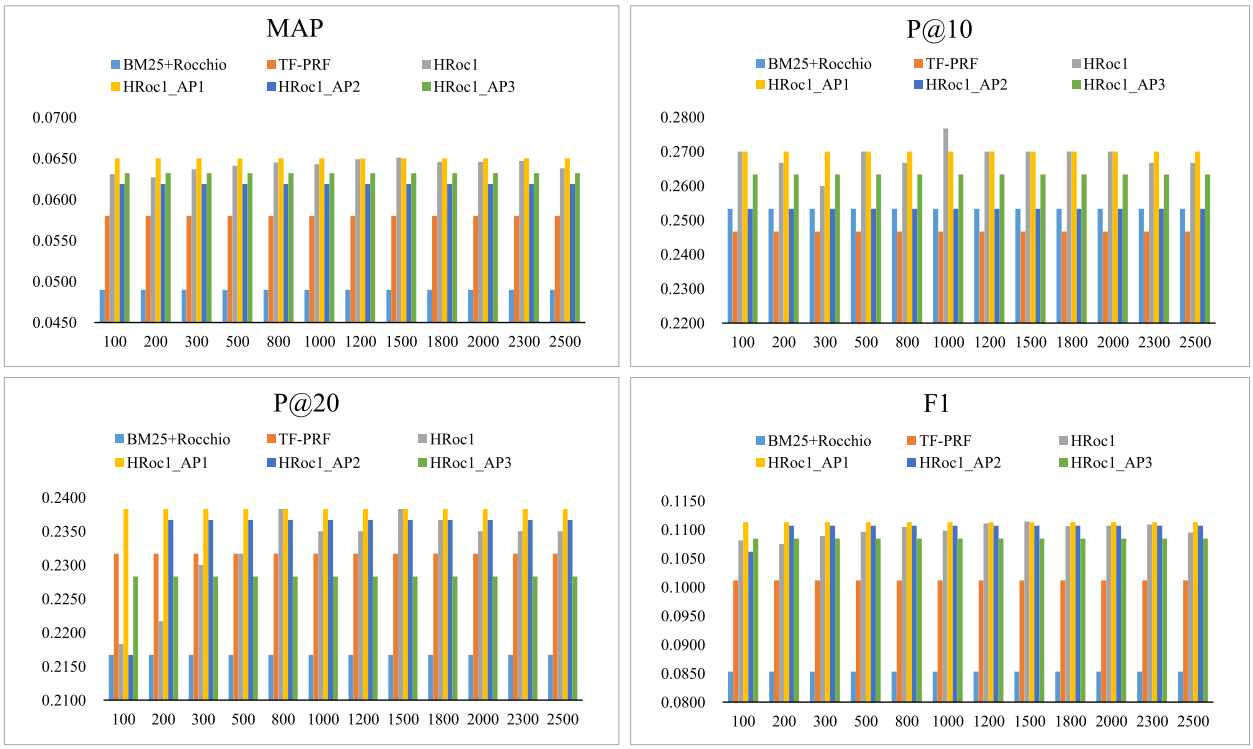

Fig. 6 Performance comparison among 6 models in MAP, P@10, P@20 and F1 metrics

Discussion

In this paper, we use five different comparing methods. In two basic retrieval models BM25+Rocchio and the KL-divergence language modeling (LM)+RM3, we get optimal values when we set $b = 0.5$ and $\mu = 2000$.

We make experiments of PRoc2 in $\sigma \in \{10, 30, 50, 100, 500, 1000, 1500\}$. As a state-of-the-art model in proximity PRF method, PRoc2 gets optimal values when $\sigma = 50$. In addition, HRoc1 achieves best result in all five different comparing methods. We get optimal value when $D = 1500$. The method of $norm_1(t)$ is the most effective in proposed three normalization methods. Both sides of the optimal value demonstrate a steady and slow decline trend.

The effectiveness of the proposed methods can be proved by the experiments on TREC medicine collections. The modification to the PRF model leads to significant and substantial (up to 32.8%) improvements. Furthermore, the proposed proximity-based function outperforms the three well-known constraints (PRoc2, PRoc3, TF-PRF). It may due to the positional semantic information in clinical records and articles. For example, the name of the drugs can represent a disease, while an article with no appearance of the disease name can be judged to be relevant when the name of a drug is close to the location of the original query. However, this phenomenon also appears in other datasets, but with higher frequency in clinical medical collection. It may also due to the significant degree of candidate term in feedback documents, which are not only decided by term frequency because some clinical jargon only appear once.

Limitations and future work

In this paper, a large number of research and experimental analysis on medical datasets, which makes the scope of application is not broad enough, the model method also has more in-depth optimization space.

Next, we will explore some other directions in the future work. First, we will do more experiments on other category collections, and therefore study more suitable relationships between D value and the length of document. Second, it is interesting to study on integrating the term's co-occurrence into other models, especially deep learning models [42, 43]. We also plan to evaluate our proposed methods on some real-world collections and applications.

Conclusions

In this paper, we proposed an enhanced HAL-based Rocchio's method. We integrate term's proximity co-occurrence weight information into classic Rocchio's model to improve retrieval performance. We then integrate three normalization methods into proposed HRoc1, HRoc2, and HRoc3 model. Proposed new method can measure the proximity co-occurrence relationships

between a candidate term and a query term. Experimental results show that our model significantly outperforms the strong baseline PRF models in terms of MAP, P@10, P@20, and F1-value. Meanwhile, our proposed methods are comparable to the state-of-art model PRoc2, PRoc3, and TF-PRF.

Additionally, we carefully analyze the D value of our proposed three HRoc methods and get a tendency figure in MAP. When D value is equal to two times of average document length, it would get the highest point of the curve. The average length of the CDS track dataset is 816.3, and we naturally associate that the D value should be related to the length of the document. Then, we make statistics and analysis on the document length of the whole dataset, and propose three adaptive functions related to dl to replace the D value and make three non-parametric adjacent normalization models. The experimental results show that the first non-parametric adjacent normalization model is not only comparable to the previous model, but also does not increase the complexity of the algorithm. At the same time, the adaptive function model has less hyper parameters than the original model, which improves the universality of the model.

Abbreviations

BM25: Best match25; CDS: Clinical decision support; DFR: Divergence from randomness models; EHR: Electronic health record; HAL: Hyperspace analogue to language; IR: Information retrieval; MAP: Mean average precision; PMC: PubMed central; PRF: Pseudo relevance feedback; PRoc: Proximity-based feedback framework; TREC: Text retrieval conference

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of *BMC Medical Informatics and Decision Making Volume 19 Supplement 9, 2019: Proceedings of the 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference: medical informatics and decision making*. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-19-supplement-9.

Authors' contributions

MP conceived and designed the study. YZ, QZ and BS analyzed data and run the experiments and contributed to writing the manuscript. TH and XJ supervised and helped conceive the study. All authors read and approved the final manuscript.

Funding

This research is supported by the Fundamental Research Funds for Central Universities (CCNU18JCK05), the National Natural Science Foundation of China (61532008), the National Science Foundation of China (61572223), and National Key Research and Development Program of China (2017YFC0909502). Publication costs have been funded by the National Key Research and Development Program of China (2017YFC0909502). We are very grateful to the anonymous reviewers and deputy editors for their valuable and excellent comments, which have greatly improved the quality of this paper.

Availability of data and materials

The presented results are available.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]National Engineering Research Center for E-Learning, Central China Normal University, 430079 Wuhan, China. [2]School of Computer and Information Engineering, Hubei Normal University, Huangshi, Hubei Normal University, 435002 Huangshi, China. [3]School of Computer, Central China Normal University, 430079 Wuhan, China.

Published: 12 December 2019

References

1. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, Pant S. Overview of the trec 2017 precision medicine track. In: Proceedings of the Twenty-Sixth Text REtrieval Conference, TREC 2017. Gaithersburg: National Institute of Standards and Technology (NIST); 2017.
2. Ksentini N, Tmar M, Gargouri F. The impact of term statistical relationships on rocchio's model parameters for pseudo relevance feedback. Int J Comput Inf Syst Ind Manag Appl. 2016;8:135–44.
3. Vaidyanathan R, Das S, Srivastava N. Query expansion strategy based on pseudo relevance feedback and term weight scheme for monolingual retrieval. arXiv preprint arXiv:1502.05168. 2015.
4. Zamani H, Dadashkarimi J, Shakery A, Croft WB. Pseudo-relevance feedback based on matrix factorization. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 2016), Indianapolis, IN, USA, October 24-28. ACM; 2016. p. 1483–92.
5. Ye Z, Huang JX. A learning to rank approach for quality-aware pseudo-relevance feedback. J Assocr Inf Sci Technol. 2016;67(4):942–59.
6. Lang H, Metzler D, Wang B, Li J-T. Improved latent concept expansion using hierarchical markov random fields. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010. Toronto: ACM; 2010. p. 249–58.
7. Lv Y, Zhai C. A comparative study of methods for estimating query language models with pseudo feedback. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009. Hong Kong: ACM; 2009. p. 1895–8.
8. Hall P. The SMART Retrieval System - Experiments in Automatic Document Processing. Information Storage & Retrieval. Elsevier Inc. 1971;9(3):199.
9. Pan M, Zhang Y, He T, Jiang X. An enhanced hal-based pseudo relevance feedback model in clinical decision support retrieval. In: Intelligent Computing Theories and Application - 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018, Proceedings, Part II; 2018. p. 93–9. https://doi.org/10.1007/978-3-319-95933-7_12.
10. Rohde DL, Gonnerman LM, Plaut DC. An improved model of semantic similarity based on lexical co-occurrence. Commun ACM. 2006;8(627-633):116.
11. Büttcher S, Clarke CL, Lushman B. Term proximity scoring for ad-hoc retrieval on very large text collections. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006. Seattle: ACM; 2006. p. 621–2.
12. Clarke CL, Cormack GV, Tudhope EA. Relevance ranking for one to three term queries. Inf Process Manag. 2000;36(2):291–311.
13. Qiao Y-n, Du Q, Wan D-f. A study on query terms proximity embedding for information retrieval. Int J Distrib Sensor Networks. 2017;13(2): 1550147717694891.
14. He B, Huang JX, Zhou X. Modeling term proximity for probabilistic information retrieval models. Inf Sci. 2011;181(14):3017–31.
15. Lv Y, Zhai C. Positional language models for information retrieval. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009. Boston: ACM; 2009. p. 299–306.
16. Chun LI. Science and technology information retrieval techniques. Education Teaching Forum. Hebei: Hebei education press; 2017, pp. 278–280.
17. Rasolofo Y, Savoy J. Term proximity scoring for keyword-based retrieval systems. In: European Conference on Information Retrieval, ECIR 2003. Pisa: Springer; 2003. p. 207–18.

18. Song R, Taylor MJ, Wen J-R, Hon H-W, Yu Y. Viewing term proximity from a different perspective. In: European Conference on Information Retrieval, ECIR 2008. Glasgow: Springer; 2008. p. 346–57.
19. Hersh W, Voorhees E. TREC genomics special issue overview. Inf. Retr. Netherlands: Springer. 2009;12(1):1-15. https://doi.org/10.1007/s10791-008-9076-6.
20. Voorhees EM, Hersh WR. Overview of the trec 2012 medical records track. In: Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012. National Institute of Standards and Technology (NIST); 2012.
21. de Herrera AGS, Kalpathy-Cramer J, Demner-Fushman D, Antani SK, Müller H. Overview of the imageclef 2013 medical tasks. Working Notes for CLEF 2013 Conference. Valencia; 2013.
22. Roberts K, Simpson MS, Voorhees EM, Hersh WR. Overview of the trec 2015 clinical decision support track. In: Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015. Gaithersburg; 2015. National Institute of Standards and Technology (NIST).
23. Simpson MS, Voorhees EM, Hersh W. Overview of the trec 2014 clinical decision support track. Technical report, LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD. 2014.
24. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. Mimic-iii, a freely accessible critical care database. Sci Data. 2016;3:160035.
25. Liu H, Song Y, He Y, Wang Y, Hu Q, He L. Ecnu at trec 2016: Web-based query expansion and experts diagnosis in medical information retrieval. In: Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016. National Institute of Standards and Technology (NIST); 2016.
26. Wang Y, Rastegar-Mojarad M, Elayavilli RK, Liu S, Liu H. An ensemble model of clinical information extraction and information retrieval for clinical decision support. In: proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016. National Institute of Standards and Technology (NIST); 2016.
27. Lavrenko V, Croft WB. Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001. New Orleans: ACM; 2001. p. 120–7.
28. Robertson SE, Walker S, Beaulieu M, Gatford M, Payne A. Okapi at trec-4. Nist Special Publication Sp. 199673–96.
29. Miao J, Huang JX, Ye Z. Proximity-based rocchio's model for pseudo relevance. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012. Portland: ACM; 2012. p. 535–44.
30. Ye Z, Huang JX. A simple term frequency transformation model for effective pseudo relevance feedback. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2014. Gold Coast: ACM; 2014. p. 323–32.
31. Colace F, De Santo M, Greco L, Napoletano P. Improving relevance feedback-based query expansion by the use of a weighted word pairs approach. J Assoc Inf Sci Technol. 2015;66(11):2223–34.
32. Zhao J, Huang JX, He B. Crter: using cross terms to enhance probabilistic information retrieval. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2011. p. 155–64.
33. Lv Y, Zhai C. Positional relevance model for pseudo-relevance feedback. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2010. p. 579–86.
34. Mbarek R, Tmar M, Hattab H, Boughanem M. Pseudo-relevance feedback method based on the cross product of irrelevant documents. IJWA. 2017;9(1):8–15.
35. Peng J, Macdonald C, He B, Plachouras V, Ounis I. Incorporating term dependency in the dfr framework. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2007. p. 843–4.
36. Metzler D, Croft WB. A markov random field model for term dependencies. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2005. p. 472–9.
37. Zhao J, Huang JX, Ye Z. Modeling term associations for probabilistic information retrieval. ACM Trans Inf Syst (TOIS). 2014;32(2):7.
38. Vechtomova O, Wang Y. A study of the effect of term proximity on query expansion. J Inf Sci. 2006;32(4):324–33.
39. Lund K, Burgess C, Atchley R. Semantic and associative priming in high-dimensional semantic space. In: proceedings of the 17th Annual Conference of the Cognitive Science Society, LEA; 1995. p. 660–665.
40. Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav Res Methods Instrum Comput. 1996;28(2): 203–8.
41. Hazimeh H, Zhai C. Axiomatic analysis of smoothing methods in language models for pseudo-relevance feedback. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval. ACM; 2015. p. 141–50.
42. Sun Q, Yang Y, Sun J, Yang Z, Zhang J. Using deep learning for content-based medical image retrieval. In: Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications, vol. 10138. International Society for Optics and Photonics; 2017. p. 1013812.
43. Mohan S, Fiorini N, Sun K, Lu Z. Deep learning for biomedical information retrieval: Learning textual relevance from click logs. In: Proceedings of the BioNLP 2017, Vancouver, Canada. Association for Computational Linguistics; 2017. p. 222–31.

Publisher's Note