JAMIA Open, 2(3), 2019, 360–368
doi: 10.1093/jamiaopen/ooz016
Advance Access Publication Date: 4 June 2019
Research and Applications

Research and Applications

Test collections for electronic health record-based clinical information retrieval

Yanshan Wang,[1] Andrew Wen,[1] Sijia Liu,[1] William Hersh,[2] Steven Bedrick[3], and Hongfang Liu[1,*]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA, [2]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA and [3]Department of Computer Science and Electrical Engineering, Oregon Health & Science University, Portland, Oregon, USA

*Corresponding Author: Hongfang Liu, PhD, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905 (liu.hongfang@mayo.edu)

Received 13 March 2019; Revised 26 April 2019; Editorial Decision 30 April 2019; Accepted 3 April 2019

ABSTRACT

Objectives: To create test collections for evaluating clinical information retrieval (IR) systems and advancing clinical IR research.

Materials and Methods: Electronic health record (EHR) data, including structured and free-text data, from 45 000 patients who are a part of the Mayo Clinic Biobank cohort was retrieved from the clinical data warehouse. The clinical IR system indexed a total of 42 million free-text EHR documents. The search queries consisted of 56 topics developed through a collaboration between Mayo Clinic and Oregon Health & Science University. We described the creation of test collections, including a to-be-evaluated document pool using five retrieval models, and human assessment guidelines. We analyzed the relevance judgment results in terms of human agreement and time spent, and results of three levels of relevance, and reported performance of five retrieval models.

Results: The two judges had a moderate overall agreement with a Kappa value of 0.49, spent a consistent amount of time judging the relevance, and were able to identify easy and difficult topics. The conventional retrieval model performed best on most topics while a concept-based retrieval model had better performance on the topics requiring conceptual level retrieval.

Discussion: IR can provide an alternate approach to leveraging clinical narratives for patient information discovery as it is less dependent on semantics. Our study showed the feasibility of test collections along with a few challenges.

Conclusion: The conventional test collections for evaluating the IR system show potential for successfully evaluating clinical IR systems with a few challenges to be investigated.

Key words: electronic health records, information retrieval, test collections, relevance judgment, evaluation

INTRODUCTION

The rapid adoption of electronic health record (EHR) systems has led to an unprecedented expansion in the volume of available free-text EHR information. Information retrieval (IR), which returns relevant documents from a large collection of documents for a user's textual queries, can be adopted to find useful free-text EHR information for clinical practice and research efficiently. Clinical IR can facilitate a variety of applications including diagnosis and treatment recommendations by finding similar patients,[1] patient recruitment for a clinical trial,[2,3] and characterization of population-scale epidemiological realities.[4] However, the evaluation of a clinical IR

system depends on a test collection environment, consisting of a set of topics or information need descriptions, a corpus consisting of a set of documents to be searched, and relevance judgments indicating which document is relevant for which topic.

The primary objective of this study is to create test collections for advancing clinical IR research. In the test collections, we will use a set of search queries (also known as topics in IR)[5] and a corpus consisting of free-text EHR documents associated with the Mayo Clinic BioBank (MCB) cohort. We first present materials, including curation of corpus and topics. Since the EHR data is different from textual data in the general domain because it contains not only free-text documents but also structured data, we describe how we leveraged the additional structured data, adopted the strategies used in test collections from the general domain and adjusted them to clinical IR. We demonstrate how to create a to-be-evaluated document pool using five retrieval models. We then report the guidelines for manual relevance judgment, and the analysis results of relevance judgment in terms of human agreement and time spent and results of three levels of relevance. We also compare the performance of five testing clinical IR models. We conclude this study with a few challenges experienced during creation of the test collections and insights for the future work.

BACKGROUND

The automatic retrieval of relevant EHR documents to meet physicians' or clinical researchers' information needs is a prerequisite to many downstream clinical applications.[6] A majority of EHR retrieval tools are based on Boolean retrieval model, also known as exact-match retrieval, where documents are retrieved if they match the query terms exactly, without ranking of documents according to the level of relevance.[7] To address the limitation of Boolean retrieval model, researchers developed IR models that incorporate ranking algorithms when retrieving documents.

Natural language processing (NLP) techniques have shown promise in their ability to be leveraged for secondary use of EHRs for finding relevant EHR documents.[8–10] These clinical NLP systems have been developed to encode information from free-text EHR documents into standard terminologies, which is usually called concept encoding. However, concept encoding in clinical NLP is a semantic processing task and a majority of the existing NLP systems have shown unsatisfactory performance[11] and portability issues.[12] IR is a technique used by search engines for storing, retrieving, and ranking documents from a large collection of text documents based on users' queries. It can provide an effective, versatile, and scalable solution to leverage clinical narratives for cohort discovery as it is less dependent on semantics.[13,14] Figure 1 illustrates the framework of an IR system used for the retrieval of relevant clinical documents. A few IR systems have been successfully developed for EHR document retrieval, such as EMERSE,[13] Léon Bérard Cancer Center System,[15] and StarTracker.[16] However, these EHR retrieval systems were rarely tested by the evaluation methods being utilized in the general IR domain due to the public unavailability for researchers.

As shown in Figure 1, in theory, the relevance judgment should be made on each retrieved document given information needs for evaluating an IR system. However, this evaluation method is not practical and requires a tremendous amount of human effort. Thus, evaluating IR systems using test collections has been widely used in general IR.[17,18] To create test collections, disparate IR models are used to create multiple runs, each of which contains a ranking list of documents from the document set for each topic with descending relevance

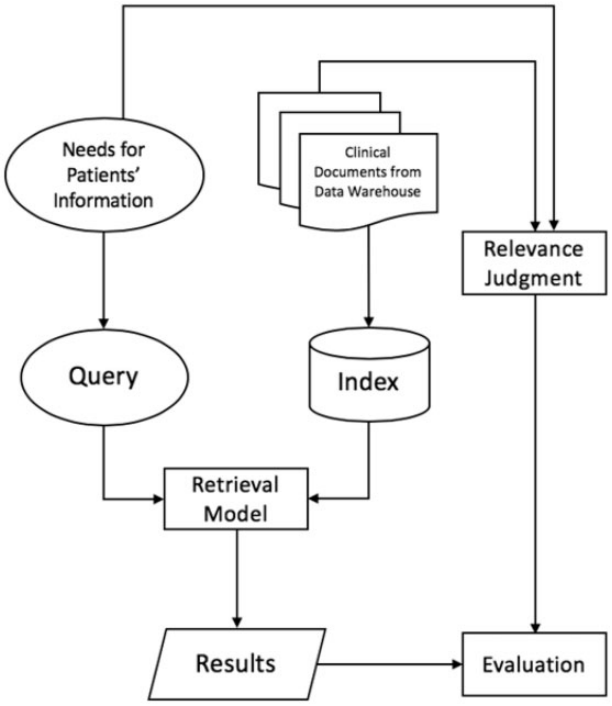

Figure 1. A general IR framework used for retrieval of clinical documents. IR, information retrieval.

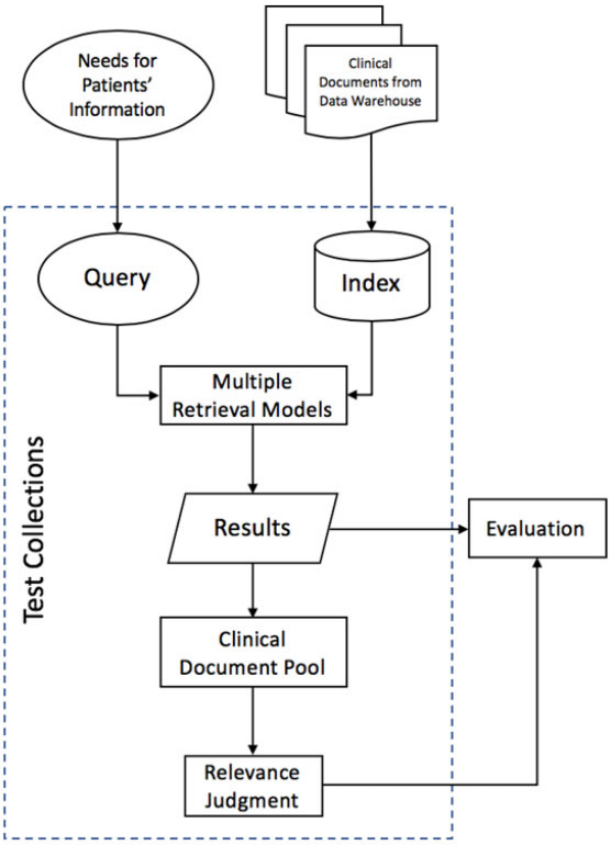

Figure 2. Test collections in the evaluation of a practical IR framework. IR, information retrieval

scores. Then a pool of sample documents retrieved by those runs for each topic is created for human annotators to make relevance judgments on each document. The judged relevant documents can be utilized as gold standards to evaluate any IR system. Figure 2 shows the use of test collections in the evaluation of a practical IR framework.

The evaluation approach using test collections has been prevalently adopted in IR shared tasks, such as the Text Retrieval Conference (TREC; http://trec.nist.gov), the Cross-Language Evaluation Forum (CLEF; http://www.clef-initiative.eu), the Forum for Information Retrieval Evaluation (FIRE; http://fire.irsi.res.in/fire/2018/home), and the NII Testbeds and Community for Information Access Research project (NTCIR; http://research.nii.ac.jp/ntcir/index-en.html). However, most of these previous tasks and studies focus on retrieval of text data in the general domain, as opposed to clinical domain. A few tracks, such as CLEF-eHealth in CLEF and the Precision Medicine track in TREC, which address medical text data, utilized Web health contents or a biomedical literature test. Only the Medical Records tracks in TREC 2011 and 2012 used the EHR data in the IR task. However, some limitations of these tracks are that topics were simple one sentence queries that could not represent real-world use cases and that documents were just unstructured de-identified clinical notes, which were only a portion of EHR data that include both structured (eg, demographics) and unstructured data. Clinical IR has been understudied due to a lack of confidence that healthcare institutions have with sharing clinical datasets due to the Health Insurance Portability and Accountability Act (HIPAA) privacy rule and security issues.[19,20] Therefore, test collections for evaluating clinical IR systems are rarely examined and investigated in the literature.

MATERIALS AND METHODS

Dataset

In this work, we used the most recent MCB cohort.[21] The cohort consists of more than 45 000 patients, mainly from the Upper Midwest states, who have been recruited since 2009 when the Biobank was established by the Mayo Clinic Center for Individualized Medicine. The goal of MCB is to support a wide array of health-related research studies, especially those with the potential to improve patient care. Many of the participants have more than 15 years of EHR history at Mayo Clinic, and many have additional years of data available for manual abstraction from paper medical records, if needed. Since the MCB cohort has been successfully utilized for the translation of individualized medicine, using this cohort for clinical IR will unlock the information in unstructured EHR data for use in future translational research. The cohort's EHR data was retrieved from the clinical data warehouse (CDW), including structured data (demographics and diagnosis codes) and free-text data (clinical notes). Since the patients in the MCB cohort might have multiple clinical visits for which clinical notes are generated, and clinical notes have standard event type labels (eg, consult, therapy, and limited evaluation) and section labels (eg, family history, diagnosis, immunizations, and impression), each patient's clinical notes were stratified into documents named by "PatientID_DocumentID_EventType_Date_SectionName," which results in more than 42 million documents. This study was approved by the Mayo Clinic institutional review board (IRB #12–009059) for human subject research.

Topics

Previous analysis suggests that a minimum of 50 topics should be used to obtain stable effectiveness estimates for an IR system.[22,23] Thus, a total of 56 topics, each illustrating a patient cohort, were developed through a collaboration between Mayo Clinic and Oregon Health & Science University (OHSU).[5] Initially there were 29 topics generated from OHSU and 30 from Mayo Clinic. Three topics with similar characteristics from both sides were merged during topic development to avoid redundancy. Clinical study data requests, as submitted by researchers to the Oregon Clinical and Translational Research Institute (OCTRI), OHSU's Research Data Warehouse (RDW), provided the basis for 26 topics at OHSU. The 30 topics from Mayo Clinic were modeled after cohorts from the Phenotype KnowledgeBase (PheKB) (7 topics), Rochester Epidemiology Project (REP) (9 topics), National Quality Forum (NQF) (12 topics), and Mayo RDW (2 topics). Since 10 topics specific to the OHSU RDW had no documents returned by the clinical IR system at Mayo Clinic, we reported the results of the remaining 46 topics in our experiment.

Each topic is written in XML format, composed of topic id, source of the topic, topic summary statement, patient statement (eg, signs, symptoms, and treatment), and inclusion and exclusion criteria (eg, patient's demographics, lab test, and diagnosis). The entire set of topics can be found in the Supplementary Material.

Indexing

In our study, we implemented the clinical IR system using Elasticsearch (https://www.elastic.co/), which is an open-source IR platform. Elasticsearch is a search engine implemented using Lucene library with additional enhanced search functionality. Compared with other search engines based on Lucence library, such as Solr, Elasticsearch is easier and more intuitive to set up a fully distributed search engine. Therefore, Elasticsearch is utilized as the search engine in this study. The structured data were indexed as value elements in Elasticsearch. Figure 3 illustrates the hierarchical index structure. We used parent/child relationship to group the documents (indexed in document fields) by patient encounters (indexed in encounter fields), and then to group encounters by patient (indexed in person fields). The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standardized medical concepts were extracted from the clinical documents using NLP algorithms and indexed as child fields to clinical documents. The relation of indexing medical concepts and more technical details related to the NLP algorithm can be found in our previous study.[24]

Document pool creation

Following the conventional test collections, we created a pool of clinical documents for relevance judgment. As aforementioned, the main distinction of clinical IR from general IR is that EHR data contains both structured and unstructured data. Since it is redundant to manually judge whether a patient's demographics satisfy the cohort criteria, we simply filter patients using the demographic data in the first step. We did not filter patients using structured diagnostic codes since patients with more than one condition may be coded with only one code for billing purposes. Using structured diagnostic codes will filter out many eligible patients, which is one of the main drawbacks in current clinical IR systems using only structured EHR data. After filtering patients, we applied five retrieval models, namely term frequency-inversed document frequency (tf-idf)-based Vector Space Model (VSM), BM25, Dirichlet Language Model (LM), Markov Random Field (MRF) model, and Cohort Retrieval Enhanced by Analysis of Text from EHRs (CREATE), to alleviate the bias of constructing the document pool toward a particular IR system. The reason we chose these five IR models is that they cover three main

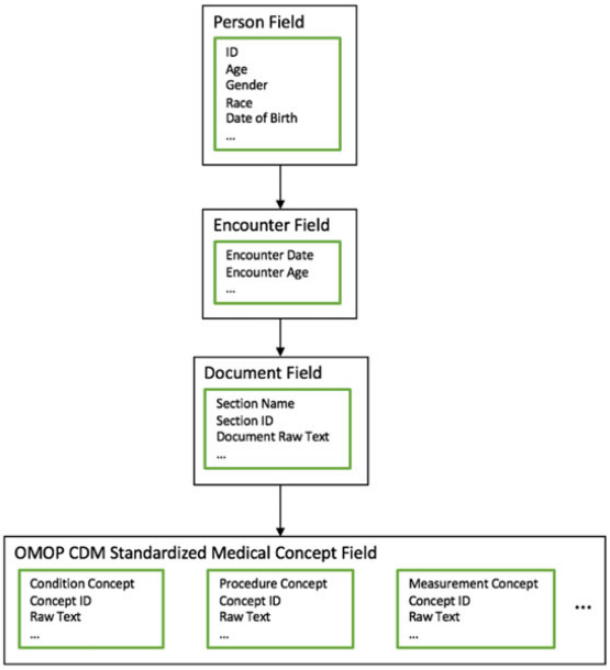

Figure 3. Hierarchical index structure in Elasticsearch.

categories of IR models: geometric models (Tf-idf based VSM), probabilistic models (BM25, Dirichlet LM, and MRF), and semantic models (CREATE). The retrieval models used to generate the pool of documents are described in Table 1.

After running the five IR systems, we followed the steps below to construct the document pool: (1) For each topic and the corresponding results, all documents retrieved in ranks 1–15 by each IR system in union with a 20% sample of documents not retrieved in the first set that were retrieved in ranks 16–100 by the IR systems were selected for input to the pool; (2) These results were merged across the IR systems and sorted randomly; and (3) Duplicate documents were removed for each topic.

Since there exist duplicates or highly similar sections in the EHR (eg, family history section is most likely duplicated at different encounters if no new contents are added), many retrieved clinical documents are duplicated in contents. In order to remove duplicate or similar documents, we calculate the similarity based on Euclidean distance and condensed all highly similar matches into a single document entry, with document ID as a concatenation of all similar document IDs. By doing so, a list of documents to be judged by human annotators was generated for each topic.

Human assessment guidelines

Each document was judged by two human annotators with clinical expertise: one with a registered nursing degree and the other with a

Table 1. Information Retrieval models for the generation of document pool

Retrieval models	Description
tf-idf based VSM	Vector Space Model (VSM)[25] represents documents and queries using t-dimensional vectors $q = [q_1, q_2, \ldots, q_t]$ and $d = [d_1, d_2, \ldots, d_t]$, where t is the number of index terms. Documents are retrieved based on the similarity between vectors of documents and queries (eg, cosine similarity). The element of each vector represents the term weight. In the tf-idf based VSM, the vector element is the tf-idf weight for each index term.
BM25	Okapi BM25[26] uses probabilistic arguments and defines an empirical scoring function to rank documents: $$s = \sum_{i \in Q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1) f_i}{K + f_i} \cdot \frac{(k_2 + 1) qf_i}{k_2 + qf_i}$$ where i is a query term in query Q, r_i is the number of relevant documents containing term i, n_i is the number of documents containing term i, R is the number of relevant documents for query Q, N is the total number of documents in the corpus, f_i is the frequency of term i in the document; qf_i is the frequency of term I in query Q, and k_1, k_2, and K are empirical parameters. If no relevant documents are available, r_i and R are set to 0.
Dirichlet LM	Language models[27] estimate a probabilistic language model for each document and rank documents by the likelihood of the query: $$\log p(Q|D) = \sum_{i \in Q} \log \left((1 - \lambda) \frac{f_{i,D}}{|D|} + \lambda \frac{c_i}{|C|} \right)$$ where i is a query term in query Q, D is a document model, $f_{i,D}$ is the number of times term i occurs in D, c_i is the number of times term i occurs in the corpus of documents, $\|D\|$ is the number of words in D, $\|C\|$ is the total number of word occurrences in the corpus, and λ is a coefficient defined as $\lambda = \mu/(\|D\| + \mu)$ with an empirical parameter μ in the Dirichlet language model.
MRF	Markov Random field (MRF) model[28] incorporates three relationships between query terms in the ranking function by leveraging the Markov property, namely full independence (T), sequential dependence (S), and full dependence (F). The scoring function is defined as: $$p(QD) = \sum_{i \in T} \lambda_T f_T(i) + \sum_{i \in S} \lambda_S f_S(i) + \sum_{i \in F} \lambda_F f_F(i)$$ where λ_T, λ_S, and λ_F are weights, and $f_T(\cdot)$, $f_S(\cdot)$, and $f_F(\cdot)$ are ranking functions for three independence, respectively.
CREATE	CREATE[24] incorporates medical concepts matching into the BM25 scoring function: $$s = s_{BM25} + \frac{1}{\|M\|} \sum_{i \in Q_M} s_i$$ where s_i is the ranking score for concept i in query concept set Q_M and $\|M\|$ is the number of concepts.

Abbreviation: IR, information retrieval; LM, Language Model; MRF, Markov Random field; VSM, Vector Space Model.

medical degree. Since the structured data in the topics (demographics) can be simply matched without errors, annotators just needed to judge whether all aspects of medical conditions and medications in clinical documents met the cohort criteria. We take Topic 4 "Postherpetic neuralgia treated with topical and systemic medication" as an example. If a patient takes gabapentin and uses a lidocaine patch, but those are actually prescribed to treat their low back pain and not postherpetic neuralgia (which they may have in a separate location), then the medications do not meet that criteria. Three levels of relevance were defined: (1) a definitely relevant judgment meant that the patient mentioned in the clinical note was unequivocally a candidate for the study; (2) a possibly relevant judgment meant that the patient mentioned in the clinical note might be a candidate for the study but insufficient information was available for a definitive decision; and (3) a not relevant judgment meant that the patient was not a candidate for the clinical study mentioned in the clinical note. Table 2 lists three examples of different levels of relevance. Each topic was judged independently by two judges with medical background. In order to measure the difficulty of judging a topic, we also asked the judges to record the time spent reviewing each topic. After the first round of judgment, any disagreements were discussed by the judges and were ultimately resolved with a final, consistent answer between the judges.

Table 2. Examples of three levels of relevance

Document	Judgment	Reason
. . . The patient with autism and cerebral palsy was treated today . . .	Nonrelevant	The patient has cerebral palsy which is the exclusion criteria.
. . . He appears to have autism . . .	Partially relevant	The result could be relevant because it mentions autism. But it does not mention any of exclusion conditions.
. . . Patient has autism and doesn't have any of neurodevelopmental disorders	Relevant	Content meets all criteria.

RESULTS

Annotation results

The total number of documents to be judged for the 46 topics is 5815, with an average of 126 documents per topic. The overall agreement amongst two judges was 0.49 in terms of unweighted Kappa, which is consistent with the moderate agreement in previous IR evaluations.[29] The moderate agreement has, in general, been found to have little impact on the relative effectiveness ranking of different IR systems. Figure 4 shows the agreement for each topic. We can observe that the agreement between two judges varies across topics. The top 5 topics with the highest agreement are Topics 47, 37, 43, 24, and 7 (in descending order), which are close or greater than 0.8. The top 5 topics with the lowest agreement are Topics 42, 31, 11, 34, and 1 (in ascending order).

Figure 5 shows the amount of time the two judges spent per document for each topic. Since one judge had no experience of IR relevance judgment, she spent more time on Topics 1 and 2 to get familiar with the process. Overall, the time spent was consistent amongst two judges. We can find that both judges spent more time on some topics than others. For example, both judges spent more than 0.5 min/document on Topics 35, 39, 42, 43, 45, and 49. These topics are treated as "difficult" topics for human judges, since they contain more inclusion and exclusion criteria than the other topics. For example, Topic 42 is "Elderly patients with dementia taking an antipsychotic and don't have schizophrenia, bipolar disorder, Huntington's disease, or Tourette's Syndrome."

Relevance judgment results

Figure 6 depicts the results of three levels of relevance for each topic judged by human experts. The averaged proportions for the query set in terms of "definitely relevant," "partially relevant," and "not relevant" are 20.2%, 37.8%, and 42.0%, respectively. The top 5 topics for which the most retrieved documents are "definitely relevant" (in percentage) are Topics 31 (96.6%), 9 (82.9%), 2 (81.1%), 7 (74.2%), and 17 (62.5%). The top 5 topics for which the most retrieved documents are "not relevant" (in percentage) are Topics 37 (100%), 53 (96.3%), 48 (95.8%), 36 (95.0%), and 11 (93.3%). The top 5 topics for which the most retrieved documents are "partially relevant" (in percentage) are Topics 18 (88.2%), 4 (84.0%), 15 (78.9%), 33 (78.4%), and 39 (76.2%).

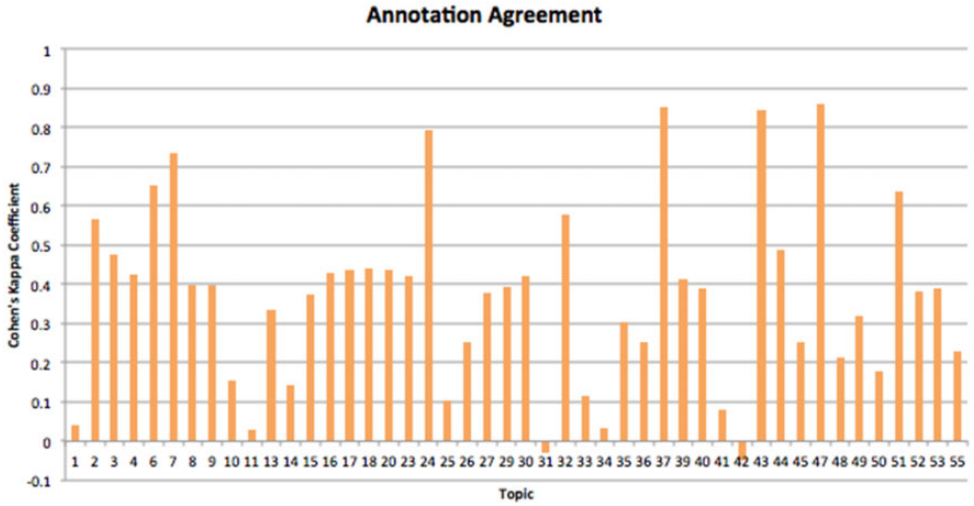

Figure 4. Agreement amongst two expert judges for each topic.

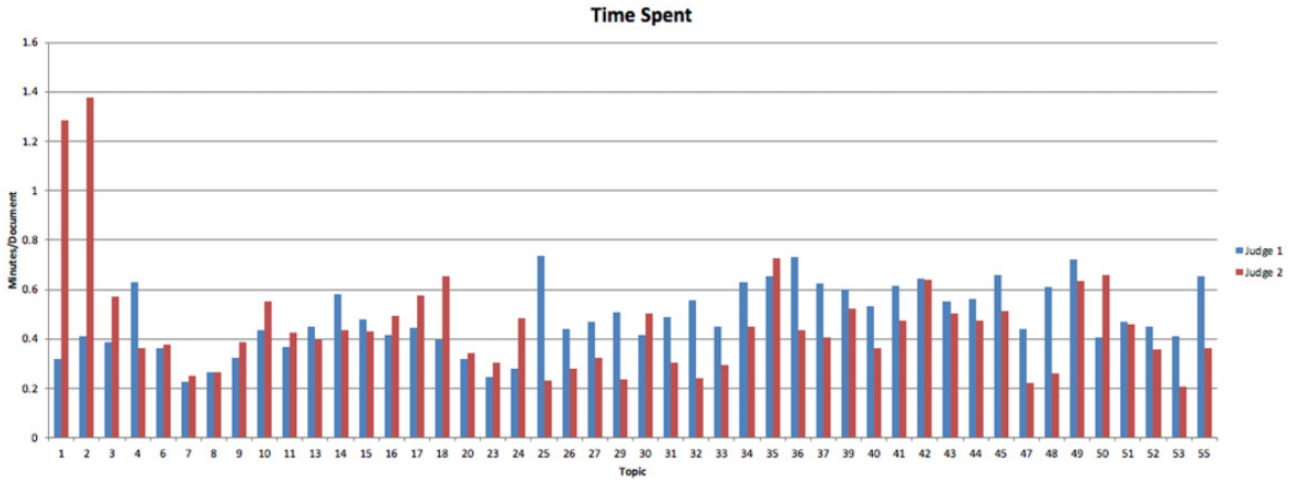

Figure 5. Time two judges spent per document for each topic.

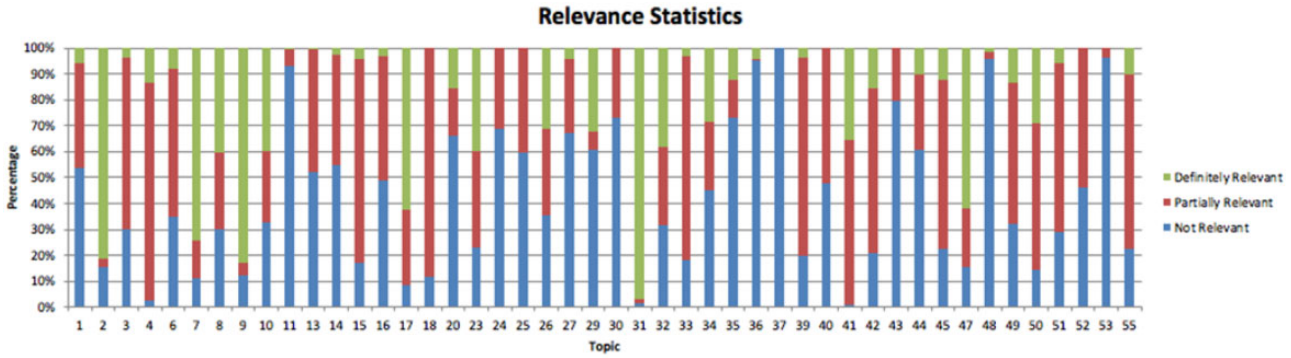

Figure 6. Results of three levels of relevance for each topic.

IR system performance

Using the relevance judgment result as the gold standard, we can evaluate the IR systems that generated the document pool. In this experiment, we used five prevalent IR metrics, including mean average precision (MAP), R-precision (Rprec), precision at the 10th document (P@10), the normalized discounted cumulative gain (NDCG), and inference average precision (infAP). MAP is defined as the mean of averaged precision over all the topics, where average prevision is the prevision at each relevant document, averaged over all relevant documents for a topic. It is mostly used in IR research to represent the overall effectiveness of an IR system.[30] Rprec is the precision after all the relevant documents have been retrieved for a topic. It measures precision at a comparable point of the retrieval process for every topic.[30] P@10 is the precision at the 10th ranked documents. It usually measures the performance on the first search results page. Discounted cumulative gain (DCG) uses a graded relevance scale of documents, for example, rating the relevance of a document from 1 (nonrelated) to 5 (very relevant), to evaluate the usefulness of a document based on its position in the retrieval list. Suppose r_i is the graded relevance of the document at ranked position i, DCG accumulated at a particular rank position p is defined as $DCG_p = \sum_{i=1}^{p} \frac{r_i}{\log_2(i+1)}$. NDCG is the normalized DCG that is computed as $NDCG_p = \frac{DCG_p}{IDCG_p}$ where $IDCG_p = \sum_{i=1}^{|R|} \frac{2^{r_i}-1}{\log_2(i+1)}$ and $|R|$ represents the list of relevant documents (ordered by their relevance) in the corpus up to position p. InfAP is an estimated metric that measures the full collection average precision from the pool subsample directly.[31]

Table 3. Performance of IR systems in terms of MAP, Rprec, P@10, NDCG, and infAP

IR model	MAP	Rprec	P@10	NDCG	infAP
tf-idf-based VSM	**0.3529**	**0.3900**	**0.6761**	**0.6035**	**0.3529**
BM25	0.3091	0.3524	0.6239	0.5622	0.3091
Dirichlet LM	0.2027	0.2577	0.6370	0.4556	0.2027
MRF	0.2060	0.2576	0.4783	0.4088	0.2060
CREATE	0.2343	0.2852	0.6065	0.4316	0.2343

Abbreviation: IR, information retrieval; LM, Language Model; MRF, Markov Random field; VSM, Vector Space Model.

A bold value indicates the best performance for that metric.

Table 3 shows the performance of five IR systems where tf-idf-based VSM outperforms other IR systems in terms of all five metrics.

Table 4 lists the complete performance of five IR systems per topic in terms of MAP. Since the CREATE is the only IR model among five systems that considers medical concepts, we compare it with the tf-idf-based VSM in the following analysis. The tf-idf-based VSM performs better than the CREATE on most topics, which is consistent with its overall performance in Table 3. However, the CREATE outperforms the tf-idf-based VSM on a few topics. For example, the performance gain of CREATE over tf-idf-based VSM is almost 100% on Topics 36 (97.2%), 43 (92.1%), 48 (91.5%), and 53 (91.5%). From Figure 6, we can observe that the percentage of retrieved documents are mostly "not

Table 4. Performance of IR systems per topic in terms of MAP

Topic ID	tf-idf based VSM	BM25	Dirichlet LM	MRF	CREATE
1	0.2996	0.0688	**0.4212**	0.2359	0.0003
2	**0.3983**	0.2936	0.2996	0.0908	0.2612
3	**0.4235**	0.3126	0.1719	0.0987	0.0892
4	0.5801	0.5670	0.2059	**0.5910**	0.5902
6	**0.4316**	0.3093	0.2616	0.0066	0.0946
7	**0.4380**	0.3286	0.3492	0.2544	0.3499
8	0.2770	0.2169	**0.3114**	0.0040	0.2326
9	**0.4259**	0.3134	0.3561	0.1827	0.1681
10	**0.2748**	0.0830	0.1089	0.1036	0.0791
11	0.0230	0.0642	**0.0708**	0.0202	0.0030
13*	0.1402	0.2031	0.0226	0.1570	**0.3275**
14	0.3229	0.2543	0.1353	0.2659	**0.3388**
15	**0.3378**	0.3199	0.2323	0.1938	0.3200
16	0.4026	**0.4574**	0.4300	0.2818	0.2969
17	**0.4821**	0.4365	0.2042	0.2407	0.1776
18	0.3832	**0.4496**	0.1164	0.4491	0.2366
20	0.2403	0.1316	**0.2990**	0.2021	0.0274
23	**0.5657**	0.4116	0.2499	0.2752	0.3718
24*	0.0752	0.1731	0.0792	**0.2511**	0.1895
25	**0.7314**	0.7149	0.6093	0.0295	0
26	**0.4976**	0.4222	0.4102	0.3195	0.0484
27	**0.2923**	0.2377	0.0799	0.0039	0.0827
29	**0.1778**	0.1725	0.1343	0.1224	0.0534
30	0.3351	**0.3579**	0.1647	0.0648	0.0387
31	**0.7397**	0.7392	0.1099	0.6502	0.6503
32	**0.6305**	0.5397	0.1635	0	0.4634
33	**0.7013**	0.6960	0.5149	0.4315	0.6180
34	0.2579	0.2254	0.1957	**0.2720**	0.0675
35	**0.1224**	0.0690	0.0625	0.0771	0.0095
36*	0.0144	0.0226	0.0584	0.0226	**0.5137**
37	0	0	0	0	0
39	0.4873	0.5137	0.4354	**0.5139**	0.2009
40	**0.4316**	0.2128	0.1686	0	0.1181
41	**0.6020**	0.5269	0.1582	0.4686	0.3117
42	**0.5740**	0.4290	0.2942	0.4482	0.1368
43*	0.0628	0.1250	0.0026	0.0483	**0.7980**
44*	0.2251	0.1085	0.1394	0.0105	**0.3948**
45	**0.4935**	0.3931	0.2616	0.2719	0.1486
47	0.4217	**0.4314**	0.1989	0.1265	0.2530
48*	0.0144	0.0163	0.0051	0.0833	**0.1695**
49	**0.3882**	0.3130	0.2970	0.2467	0.1911
50	0.3704	**0.3819**	0.2421	0.3767	**0.3819**
51*	0.2048	0.1182	0.0916	0.0017	**0.2972**
52	0.4720	**0.5531**	0.0359	0.5447	0.2462
53*	0.0216	0.0042	0.0197	0.0062	**0.2542**
55	0.4435	**0.5019**	0.1463	0.4316	0.1766

Note: The topics for which the CREATE significantly outperforms the tf-idf-based VSM using *t*-test ($P < .01$) are marked by the asterisk (*).

Abbreviation: IR, information retrieval; LM, Language Model; MRF, Markov Random field; VSM, Vector Space Model.

A bold value indicates the best performance for that topic.

relevant" for these topics. This result indicates that the CREATE, a concept-based system, might be effective for topics requiring conceptual level retrieval.

DISCUSSION

The widespread adoption of EHRs has enabled secondary use of EHR data for clinical research and healthcare delivery. Many institutions have established CDWs in conjunction with patient information discovery tools (eg, i2b2) to enable investigators to use EHR data for identifying patient information. A majority of those patient information discovery tools are, however, solely based on structured EHR data (eg, billing codes, lab tests, and demographic information). This limitation leads to reduced retrieval performance for patient information discovery tasks since a significant portion of relevant patient information is embedded in clinical narratives. To compensate, NLP techniques have shown promise in their ability to be leveraged for secondary use of EHRs for clinical research. Many clinical NLP systems have been developed to encode information from unstructured data into standard terminologies for various downstream applications.[8] However, concept encoding in clinical NLP is a semantic processing task and a majority of the existing NLP systems have shown unsatisfactory performance[11] and portability issues.[12]

IR, a technique used in search engines for storing, retrieving, and ranking documents from a large collection of text documents based on users' queries, can provide an alternative approach to leverage clinical narratives for patient information discovery as it is less dependent on semantics.[13] Since test collections are the most widely used evaluation tool in the development of an IR system, our work attempts to investigate the feasibility of test collections in clinical IR using the real-world EHR data and topics. There are no similar studies in the literature, to the best of our knowledge. The experimental results showed the feasibility of test collections on most topics in our clinical IR task. However, the pooling method may require more efforts in choosing a good range of different kinds of IR systems for the topics that were conceptually relevant. We showed that most documents to be judged in the pool for a few topics (eg, Topics 36, 48, 53) were not relevant. Robertson suggested that some manual systems involving human-designed search strategies should be used for generating the document pool.[32] In our future work, we will design rule-based IR systems and incorporate more medical concept-based systems for creating the document pool.

We faced a few challenges while conducting test collections in this study. First, it was challenging to choose a range of IR models that generated a reasonable variety of relevant documents. Unlike TREC that pooled documents from a large number of participant systems, we could only leverage a limited number of IR models. The tf-idf-based VSM, BM25, and Dirichlet LM are models implemented in Elasticsearch. The MRF model is a comprehensive term dependency IR model. The CREATE is designed to incorporate extracted medical concepts. We hoped that the diverse set of IR approaches could alleviate the bias of constructing the document pool toward a particular IR approach. Second, duplicates were found in the document pool due to the nature of how clinical notes were generated (eg, copy-and-paste). Thus, we utilized Euclidean distance to condense all highly similar matches into a single document entry. The third challenge was that there existed no published evaluation guidelines for clinical IR relevance judgment. We designed our guideline following a conventional TREC guideline and tailored it to the clinical IR task. However, this guideline is not optimal in that the annotators still had difficulty in judging many documents due to the complexity of topics and EHRs. The annotators found it challenging on how to judge if a topic had exclusions while a retrieved document did not have the exclusion information.

We take Topic 29 "Adults 20–73 years old who have had radioiodine thyroid ablation, thyroid lobectomy, or thyroidectomy, and who have never had ischemic heart disease, including myocardial infarction or coronary atherosclerosis, and have never had cerebrovas-

cular disease, including stroke or transient ischemic attack" as an example, the top 2 retrieved documents were "Hypothyroidism on replacement Graves' disease status post radioactive iodine with subsequent hypothyroidism," and "Graves' disease status postradioiodine ablation, now on thyroid replacement therapy. Hypertension. De Quervain's tenosynovitis". The exclusion information in Topic 29 was "ischemic heart disease, including myocardial infarction or coronary atherosclerosis" and "cerebrovascular disease, including stroke or transient ischemic attack". However, this information was not mentioned in the retrieved documents. The reasons might be (1) the patient never had these conditions; (2) the patient had these conditions but not present; or (3) the patient had never examined these conditions. The document relevance should be (1) definite for 1, (2) partial, and (3) unknown. Since patient-level judgment requires tremendous efforts in reviewing the large longitudinal data and currently there are no IR systems for retrieving relevant patients for the pooling method, we focused on document-level judgment in our relevance judgment. We chose to judge the document as definitely relevant if there was no evidence of such exclusion information in the document.

In our future work, we would like to improve test collections for clinical IR, specifically for patient-level retrieval. We will investigate creating patient-level test collections and judging patient-level relevance. Since current IR models focusing on retrieving documents and patient retrieval is not simply combining relevance scores of the patient's clinical records, we will need to propose novel IR models for patient cohort retrieval. We will introduce granularity and priority of information components (eg, inclusions and exclusions) in the topics, and define multilevels of relevance for manual judgment. We will provide more detailed annotation guidelines for annotators based on the definition of relevance. Accordingly, we may need to introduce new IR metrics for evaluating clinical IR systems to account for the different levels of relevance. Since our results show that the medical concept-based CREATE is more effective than the traditional IR models on some topics, we will develop a categorization method for topics to select the optimal IR model for each topic category. Finally, we would like to investigate user variables and perspectives in clinical IR since a majority of clinical IR applications, such as cohort identification in retrospective clinical studies and patient recruitment in clinical trials, are for clinical purposes where expert's knowledge and experience play a crucial role.

There are limitations in this study. First, the experiment was only conducted at one institution. Since different institutions have disparate implementations of EHR systems and infrastructures, information available in EHRs and topics, and document pooling strategy may be different in different EHR systems. Analysis of such differences may be of interest across institutions. We will conduct similar experiment at OHSU, which may of be interest to analyze the different results. Second, we only used one IR platform (ie, Elasticsearch) despite we utilized disparate retrieval models to create test collections. Third, the IR systems could only partially address the lexical variation issue in clinical notes. There are multiple data quality issues and substantial variations in how medical concepts are represented in clinical notes.[33] One of the IR systems, CREATE, leveraged an automatic NLP algorithm to extract the OMOP CDM standardized medical concepts from clinical notes. For example, "type 2 diabetes" or "type II diabetes" are extracted and mapped to the OMOP CDM standardized medical concept "Diabetes Type 2." However, this system is subject to the performance of NLP algorithm for identification of medical concepts. Therefore, the IR systems could not fully address the lexical variation issue.

CONCLUSION

In this study, we investigated the feasibility of test collections, the most widely used evaluation tool for IR system development in the general domain, in evaluating clinical IR systems using the previously defined corpus and topics at Mayo Clinic. We described in detail the IR models for generating a pool of EHR documents and human relevance judgment guidelines. Finally, we analyzed the annotation results in terms of human agreement and time spent, and results of three levels of relevance, and performance of IR systems. The experimental results showed that test collections were feasible on most topics in our clinical IR task and that the pooling method might require more efforts in choosing a good range of different kinds of IR systems or incorporating rule-based IR systems for the topics that were conceptually relevant. We reported a few challenges during conducting test collections in this study and insights for our future work.

FUNDING

This work was supported by National Institutes of Health grant numbers R01LM11934, U01TR002062, and UL1TR02377 Supplement.

CONTRIBUTORS

Y.W.: conceptualized the study; wrote the manuscript; and designed and implemented the methods. A.W.: implemented the system; retrieved the data; and edited the manuscript. S.L.: implemented the system; and edited the manuscript. W.H.: edited the manuscript. S.B.: edited the manuscript. H.L.: conceptualized the study, and edited the manuscript. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

The authors would like to thank Donna M. Ihrke, R.N., and Xin Zhou, M.D. for the relevance judgment.

Conflict of interest statement. None declared.

REFERENCES

1. Frost JH, Massagli MP. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. *J Med Internet Res* 2008; 10 (3): e15.
2. Pathak J, Kiefer RC, Chute CG. Using semantic web technologies for cohort identification from electronic health records for clinical research. *AMIA Jt Summits Transl Sci Proc* 2012; 2012: 10–9.
3. Sarmiento RF, Dernoncourt F. Improving Patient Cohort Identification Using Natural Language Processing. *Secondary Anal Electron Health Rec* 2016; 405–17.
4. Wu ST, Sohn S, Ravikumar K, *et al*. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. *Ann Allergy Asthma Immunol* 2013; 111 (5): 364–9.
5. Wu S, Liu S, Wang Y, Timmons T, Uppili H, Bedrick S, Hersh W, Liu H. Intrainstitutional EHR collections for patient-level information retrieval.

Journal of the Association for Information Science and Technology 2017; 68(11): 2636–48.
6. D'Avolio LW, Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010; 17 (4): 375–82.
7. Goodwin TR, Harabagiu SM. Learning relevance models for patient cohort retrieval. *JAMIA Open* 2018; 1 (2): 265–75.
8. Kang T, Zhang S, Tang Y, *et al*. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc* 2017; 24 (6): 1062–71.
9. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
10. Liu H, Bielinski SJ, Sohn S, *et al*. An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc* 2013; 2013: 149–53.
11. Pradhan SE, Chapman W, Manandhar S, Savova G. SemEval-2014 task 7: analysis of clinical text. *SemEval* 2014; 199 (99): 54.
12. Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012; 19 (e1): e162–9.
13. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J Biomed Inf* 2015; 55: 290–300.
14. Goodwin TR, Harabagiu SM. Multi-modal patient cohort identification from EEG report and signal data. *AMIA Annu Symp Proc* 2016; 2016: 1794–803.
15. Biron P, Metzger MH, Pezet C, Sebban C, Barthuet E, Durand T. An information retrieval system for computerized patient records in the context of a daily hospital practice: the example of the Leon Berard Cancer Center (France). *Appl Clin Inform* 2014; 5 (1): 191–205.
16. Gregg W, Jirjis J, Lorenzi NM, Giuse D. StarTracker: an integrated, web-based clinical search engine. In *AMIA Annu Symp Proc* 2003 (Vol. 2003, p. 855). American Medical Informatics Association.
17. Voorhees EM. The philosophy of information retrieval evaluation. *Workshop Proceedings of the Cross-Language Evaluation Forum for European Languages* 2001; 2001; 355–370.
18. Sanderson M. Test collection based evaluation of information retrieval systems. *FnT Inf Retrieval*. 2010; 4(4): 247–375.
19. Lee J, Sun J, Wang F, Wang S, Jun CH, Jiang X. Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR Med Inform* 2018; 6 (2): e20.
20. Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inf* 2018; 77: 34–49.
21. Olson JE, Ryu E, Johnson KJ, *et al*. The Mayo Clinic Biobank: a building block for individualized medicine. *Mayo Clin Proc* 2013; 88 (9): 952–62.
22. Voorhees EM, Buckley C. The effect of topic set size on retrieval experiment error. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM; 2002: 316–323.
23. Sakai T. Topic set size design. *Inf Retrieval J* 2016; 19 (3): 256–83.
24. Liu S,Y, Hong N, Shen F, Wu ST, Hersh WR, Liu H. On mapping textual queries to a common data model. 2017 IEEE International Conference on Health Informatics (ICHI) 2017: 21–5.
25. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM* 1975; 18 (11): 613–20.
26. Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. Nist Special Publication Sp. 1995;109: 109.
27. Zhai C, Lafferty J. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM; 2001: 334–342.
28. Metzler D, Wb C. 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*. Salvador, Brazil: ACM, 2005: 472–79.
29. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press; 2008.
30. Voorhees EM, Harman DK. *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge: MIT Press; 2005.
31. Estimating average precision with incomplete and imperfect judgments. Proceedings of the 15th ACM international conference on Information and knowledge management; 2006. ACM.
32. Robertson S. On the history of evaluation in IR. *Journal of Information Science* 2008; 34 (4): 439–56.
33. Yilmaz E, Aslam JA. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management* 2006 Nov 6 (pp. 102-111). ACM.