Chapter 11
Information Retrieval and Query Expansion for Biomedical Data

Sadika Sood and Hazra Imran

Abstract The field of biomedical research generates vast amount of data from various sources, including electronic health records, scientific publications, clinical trials, and experimental studies. This data provides valuable insights into the underlying mechanisms of diseases and their treatments. However, biomedical data's sheer volume and complexity pose significant challenges in retrieving and analyzing relevant information. This chapter provides an overview of information retrieval with focus on retrieving biomedical data. It begins by introducing information retrieval (IR) and highlights the significance of IR. Further, it discusses the role of query expansion in improving the performance of an information retrieval system (IRS). The insights gained from this chapter can help researchers in designing and developing more effective IR systems that can unlock the full potential of biomedical data and accelerate progress in biomedical research.

Keywords Information retrieval · Vector space model (VSM) · Word embedding · Query expansion (QE) · Pseudo-relevance feedback-based query expansion

Abbreviations

IRS	Information Retrieval System
IR	Information Retrieval
TREC	Text REtrieval Conference
VSM	Vector Space Model
EHR	Electronic Health Records
QE	Query Expansion
AQE	Automatic Query Expansion

S. Sood
TeOra, Banglore, India

H. Imran (✉)
Northeastern University, Vancouver, BC, Canada
e-mail: h.imran@northeastern.edu

A. Sharan et al. (eds.), *Text Mining Approaches for Biomedical Data*, Transactions on Computer Systems and Networks, https://doi.org/10.1007/978-981-97-3962-2_11

CUI	Concept Unique Identifier
RF	Relevance Feedback
PRF	Pseudo-Relevance Feedback
CDS	Clinical Decision Support
IC	Information Content
DD	Descendant Distance
TF	Term Frequency

11.1 Introduction to Information Retrieval

Information retrieval (IR) plays a crucial role in satisfying users' information needs. Traditionally information retrieval means searching the relevant documents from a large text corpus based on the user's query. As the digital information grows, the size of the corpus may go to millions of documents, where a very few will be relevant to a query. Even if there are many relevant documents, the user, especially the web user, may be interested in the top few documents at times. Thus, the main challenge for the information retrieval system (IRS) is to accept a user's query and return a ranked list of relevant documents to the users. Due to heterogeneous and semi-structured documents available on the web (instead of plain text), information retrieval involves searching for relevant information from the documents, information within documents, and metadata about documents, as well as searching relational databases and the World Wide Web. Information retrieval is an interdisciplinary field that draws on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, and statistics. While not a new concept, the history of IR dates back almost 4000 years, as humans developed methods to store and organize information for later retrieval. The science of IR focuses on representing, storing, organizing, and accessing information items. An old definition of information retrieval by Mooers (Mooers 1951) recited from Savino (Savino and Sebastiani 1998) is as follows:

> Information retrieval is the process or method whereby a prospective user of information can convert his need for information into an actual list of citations to documents in storage containing information useful to him.

Biomedical information is proliferating electronically as many journals, clinical records, and books are using open access models for publishing the scientific information and distributing it on the web. A lot of medical domain data is also available on social media from past decades in the form of images or unstructured/free text (Piwowar et al. 2018). These heterogeneous resources contain a lot of hidden information, the information that cannot be extracted implicitly. Discovering this hidden knowledge is tedious, time-consuming, labor-intensive, and a challenging task. According to Baeza-Yates (Baeza-Yates and Ribeiro-Neto 1999), information retrieval is driven by an information need, which may be explicitly or implicitly

stated. With the advancements in technology, there are now multiple sources from which information should be retrieved, including digital libraries, digital repositories, and the World Wide Web. This has led to the development of information retrieval systems, automatically retrieving relevant information from digital sources. An information retrieval model is required to develop an information retrieval system.

In general, information retrieval models operate on large, fixed collections of documents (corpora), from which the model attempts to find out the useful information that best matches (is most relevant to) a user's need (query). Baeza-Yates (Baeza-Yates and Ribeiro-Neto 1999) gives a general definition of an IR model:

An information retrieval (IR) model is a quadruple [D, Q, F, R(qi, dj)] consisting of:

- D is a set of logical views for the documents in the collection.
- Q is a set of logical views for the user's information needs expressed as queries.
- F is a framework for modeling document representations, queries, and relationships.
- R(qi, dj), a ranking function that associates a real number with a pair of query qi from Q and a document dj from D. This ranking function defines an ordering among the documents relevant for the query qi.

Figure 11.1 presents an abstract view of the working of an information retrieval system.

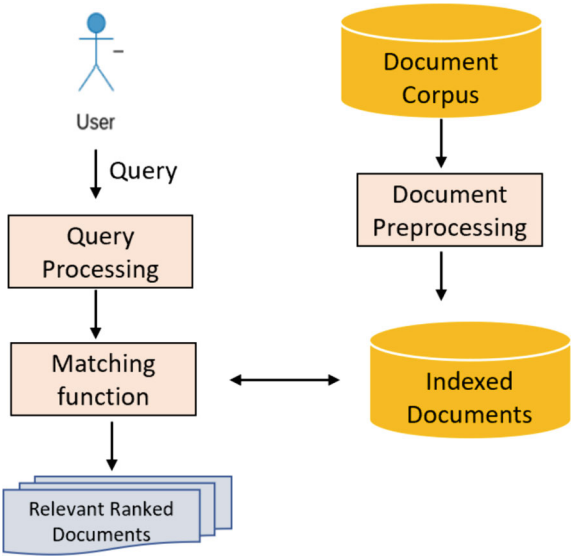


Fig. 11.1 Typical information retrieval system

11.2 Vector Space Model (VSM) for Information Retrieval

Over the years, different information retrieval models have been proposed. These models can be categorized based on their mathematical approach or their properties. The models based on mathematical approaches include: set-theoretic models, algebraic models, probabilistic models, vector space models, and machine-learned ranking models. However, vector space models have succeeded as most popular models for developing and deploying information retrieval systems. The most important reason being the simplicity and computational efficiency of vector based representation. Also this representation allows to provide a numeric score to pair of query and document based on their match, thus allowing the ranking of the documents based on relevance to the query.

Classical VSM represents each document in the space of the corpus vocabulary, by giving a weight to each word in the document, whereas advance VSM represents each document as sequence of word embeddings.

11.2.1 Classical Vector Space Model

The classical vector space model represents documents and queries as vectors in a |V|-dimensional space, where V is the vocabulary of distinct terms in the collection (corpus). The details of the text representation in the classical vector space model have been presented in Chap. 9. In this model, the collection of documents can be represented as a document-term matrix D_{m*n} where m is number of documents and n is the vocabulary size. Each term d_{ij} of D represents the weight of the jth term in ith document. The query is also represented as a document.

Let us consider following sample query along with collection of documents as given in the code block:

```
# User query and document corpus used in Classical VSM model
query="are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19"
documents=[
"Role of age, comorbidity and renin-angiotensin-aldosterone system in COVID-19. Effects of ACE inhibitors and angiotensin receptor blockers", "Bioinformatic characterization of angiotensin-converting enzyme 2, the entry receptor for SARS-CoV-2",
"COVID-19 in diabetic patients: related risks and specifics of management",
"Genetic diversity and evolution of SARS-CoV-2",
"Many patients receiving 5-alpha-reductase inhibitors be in higher risk of  COVID-19 complications",
"Characterization of Angiotensin Converting Enzyme-2 (ACE2) in Human Urine"
      ]
```

This is a sample query from standard collection of Cord-19 dataset.

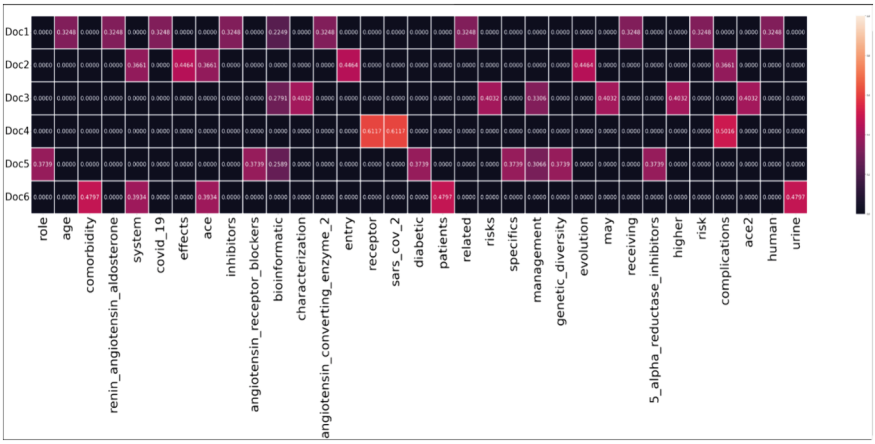


Fig. 11.2 Document term matrix for a document corpus

We will try to understand, how to represent this information in VSM. Firstly, the query and document are preprocessed by standard preprocessing task (like stopword removal, numword removal, remove punctuation marks, etc.). We further preprocess it in context of biomedical domain. In our case, we consider biomedical entities only to represent the document. For each preprocessed document, the biomedical entities are extracted using a tool called Metamap (Aronson and Lang 2010). The multiword entities (if any) were concatenated with underscore ('_') so that these entities can be recognized as a single entity. Finally, the document-term matrix is created using TF-IDF approach and is presented in Fig. 11.2. (You may refer to Chap. 9 for details of TF_IDF approach for document-term matrix construction).

It can be observed that each document is represented as a term vector (dimensionality of vector is the vocabulary size of the corpus). Query will also considered as a document and is thus represented as a vector.

The main drawback of this classical VSM approach is that it is a keyword-based model and is that it is not able to capture the semantic and syntactic information hidden in the text. Another important limitation is the high dimension and sparsity of text vectors. In spite of these limitations, classical VSM survived for a long time because of its simplicity, lesser computational complexity and its reasonable efficiency for retrieving the relevant documents. Another important reason for the success of this model was that none of the model by that time was able to capture the syntactic and semantic information hidden in the text, while maintaining reasonable computational complexity (Rawal 2020). To overcome this problem, advanced vector space models were developed. These models are based on pre-trained embeddings for representing the text. The next section focuses on these advanced models for text representation.

11.2.2 Advanced VSM Based on Word Embedding

Through embedding-based models, the linguistic entities (i.e., words, sentences, or documents) can be represented by pre-trained vectors/embeddings. Most primitive of the linguistic embedding are the word embedding. Word2vec is one of the most popular techniques for representing word as embedding. These embedding are significant in the sense that semantically similar words have similar embedding. Thus embedding-based VSM is able to capture the semantic relation between words in the text corpus to a certain extent. The details of these embedding models are presented in Chap. 9 and 10.

In this approach, each document and the query are represented as a sequence of word embedding. Multiple words are generally represented by summing the embedding of each word or any other aggrgeation method. Due to multiple length of the documents, a fixed vector length needs to be selected and necessary padding needs to be done as per requirement. These models are able to overcome many limitation of classical VSM, including the problem of high dimensionality and sparsity. Also the models facilitate capturing of the contextual information of the words in the text representation itself.

11.2.3 Indexing and Matching in VSM

Once the documents are represented in VSM, documents are then indexed using an inverted index on terms, in order to optimize speed and performance of searching the documents relevant to the query. The user query can also be represented as a document in term document matrix. The query is then matched with the documents using a vector-based similarity measure (Pedersen et al. 2007, 2009; Hao and Fan 2017). Mostly cosine similarity is used for matching query and document.

Let us extend the example of the query and document presented in code block of Sect. 11.2.1 for ranking the documents based on similarity match between query and the documents. We present here the comparison of results of ranked document obtained via classical VSM and advanced VSM. For evaluation of the result, we judge it form the benchmark results on the cord-19 dataset that provides the ground truth. According to ground truth, doc1, doc2, doc5, and doc6 are relevant.

Figure 11.3 presents the ranked list of relevant documents obtained by finding cosine similarity between the query and the document for the selected query using the classical VSM.

It can be observed that only one document Doc1 has a nonzero similarity with the query. It can also be observed (from the benchmark data) that Doc2 and Doc3 are relevant but still have a zero similarity. As this model is a keyword-based model, the query and document (Doc 2 and Doc 3) may be semantically similar to the query, however documents might not be containing any query word (leading to zero cosine similarity) still might be relevant to the query.

	Doc_id	Documents	Cosine_Similarity	Relevancy
0	Doc1	Role of age, comorbidity and renin_angiotensin...	0.459332	relevant
1	Doc2	Bioinformatic characterization of angiotensin_...	0.000000	relevant
2	Doc3	COVID_19 in diabetic patients: related risks a...	0.000000	relevant
3	Doc4	Genetic_diversity and evolution of SARS_CoV_2	0.000000	irrelevant
4	Doc5	May patients receiving 5_alpha_reductase_inhib...	0.000000	irrelevant
5	Doc6	Characterization of Angiotensin_Converting_Enz...	0.000000	irrelevant

Fig. 11.3 Ranked list of relevant documents that match the query using classical VSM

Now we present the results for word embedding-based models. Here as discussed in Sect. 11.2.2, the query and documents are represented as sequence of word embedding. This embedding model has been developed by tensorflow named as cord-19/swivel-128d (*TensorFlow Hub* 2023) trained on cord-19 dataset. Though this is a vector space-based model, the model actually represents the documents in the form of tensor as each word is represented via embedding. As the document embedding is high-dimensional numerical vectors, it is not worth presenting in form of table. With the advent of new technologies, one can use T-SNE or PCA to visualize the high-dimensional data in 2D space and for the details of these visualization one may refer Chap. 6.

After this representation, the query and document similarity can again be calculated using cosine similarity. Cosine similarity is calculated between the query vector and the document vectors and documents are ranked based on this similarity.

Figure 11.4 presents top5 documents, ranked based on cosine similarity between the query and the document.

When the results are compared with the benchmark dataset, it is observed that out of top 6 documents, 3 documents are relevant with a similarity threshold ≥ 0.7. The result is not perfect but better than classical VSM mode. It can be observed that all

	Doc_id	Documents	Cosine_Similarity	Relevancy
0	Doc1	Role of age comorbidity and renin angiotensin ...	0.847278	relevant
5	Doc6	Characterization of Angiotensin Converting Enz...	0.724236	irrelevant
4	Doc5	May patients receiving 5 alpha reductase inhib...	0.723748	irrelevant
1	Doc2	Bioinformatic characterization of angiotensin ...	0.682408	relevant
2	Doc3	COVID19 in diabetic patients related risks and...	0.551934	relevant
3	Doc4	Genetic diversity and evolution of SARSCoV2	0.162880	irrelevant

Fig. 11.4 Ranked list of relevant documents that match the query using the advanced VSM approach

the documents now have a nonzero similarity with the query (in spite of not having any keyword-based similarity with the query). This has been made possible due to semantic similarity captured by the word2vec. If the similarity threshold is kept more than 0.5, docs 1, 6, 5, 2, 3 are relevant. However, the model is identifying doc3 is identified incorrectly as it is not relevant according to benchmark. The similarity threshold plays an important role in selecting relevant documents and has to be set empirically. In fact there is no perfect model, so a scope of improvement always exists.

With advancements in IR tools, indexing and matching are embedded in a tool called ElasticSearch that was released in 2010. It is built on Apache Lucene that allows it to quickly store, search, analyze, and visualize a huge amount of data and return the results in milliseconds. It is represented as an Elasticsearch logstash and kibana (ELK), including Beats. Chakraborty et al. (2020) used ElasticSearch with BiomedBERT embeddings for question-answering and information retrieval tasks. Experiments were carried out to develop the new computational method. Companies like NetFlix, eBay, and Walmart use elastic stack to get inside information about their customers. Performance evaluation was done against pre-existing methodology used by the above-stated sub-components of IR systems.

11.3 Evaluation of Information Retrieval

Evaluating the performance of an IRS is not a straightforward task. People's disagreement about the relevance of a document to a query can be affected by their needs, preferences, expertise, and the collection from which the document is retrieved. Relevance is a subjective concept that depends on various factors (Salton and McGill 1983). However, the effectiveness of retrieval needs to be evaluated objectively. Various efficiency metrics have been defined to evaluate the system's ability to retrieve relevant documents, while minimizing the retrieval of non-relevant ones. Retrieval effectiveness is considered the primary performance indicator. Due to a lot of subjectivity involved, a notion of benchmark dataset is required to measure the effectiveness of an IRS. A benchmark dataset for information retrieval consists of three items:

- A document collection
- A test suite of information needs, expressed as queries
- A set of relevance judgments, a binary assessment of either relevant or non-relevant for each query document pair.

Recall and precision are two fundamental evaluation metrics for assessing the performance of an IRS. Recall measures the system's ability to retrieve all the relevant documents for a given query. It is calculated as the ratio of the number of relevant documents retrieved by the system to the total number of relevant documents in the collection. High recall means that the system retrieves a large proportion of relevant documents. Precision measures the system's ability to retrieve only relevant

documents for a given query. It is calculated as the ratio of the number of relevant documents retrieved by the system to the total number of documents retrieved. High precision means that the system retrieves mostly relevant documents and a few non-relevant ones. Both recall and precision are important for evaluating the performance of an information retrieval system. However, they are often in conflict with each other, and improving one may come at the expense of the other. A common way to balance recall and precision is to use a single evaluation metric that combines both, such as the $F1$ score, which is the harmonic mean of recall and precision. Table 11.1 displays the evaluation metrics that are frequently utilized.

The trade-off between recall and precision is important in information retrieval. Recall refers to the proportion of relevant documents retrieved by a search, while precision refers to the proportion of retrieved relevant documents. In general, higher recall implies that more relevant documents are retrieved, while higher precision implies that the retrieved documents are more relevant.

11.4 Information Retrieval in Biomedical Domain

The significance of information retrieval in the biomedical domain cannot be overstated (Lee et al. 2008). Access to accurate and relevant information is critical for biomedical researchers, clinicians, and other healthcare professionals to make informed decisions and advance medical knowledge. Effective information retrieval can facilitate the discovery of new treatments, improve patient outcomes, and accelerate progress in biomedicine. The field of biomedical information retrieval faces several key challenges, including:

- **The complexity of biomedical data**: Biomedical data is often complex and heterogeneous, with a wide range of data types and formats, including textual, numerical, and multimedia data. Additionally, biomedical data often includes domain-specific terminology and jargon, making it difficult for non-experts to search for and retrieve relevant information.
- **Vast amounts of data**: Biomedical research generates vast amounts of data, including scientific literature, electronic health records, genomic data, and imaging data. This presents a significant challenge in searching and retrieving relevant information from these large and complex datasets.
- **Rapidly evolving field**: Biomedical research is a rapidly evolving field, with discoveries and technologies emerging at a rapid pace. Keeping up with the latest research and trends can be challenging, particularly as new data sources and analysis techniques become available.

Biomedical information retrieval (IR) is a specialized application of the basic IR system (Ramampiaro and Li 2011). The basic IR system uses techniques such as indexing, query processing, and relevance ranking to retrieve relevant information from a collection of documents. Similarly, biomedical IR systems use these

Table 11.1 Evaluation metrics that are frequently utilized

Based on	Evaluation metric	Definition	Formula
Unranked retrieval metric	Precision (P)	The fraction of retrieved documents that are relevant	$P = \frac{\text{Number of relevant document retrieved}}{\text{Total number of documents}}$
	Recall (R)	The fraction of relevant documents out of all documents retrieved	$R = \frac{\text{Number of relevant document retrieved}}{\text{Total number of relevant documents}}$
	F-measure	Is the harmonic mean of precision and recall	$F = (2 * P * R)/(P + R)$ Where P represents precision, and R represents recall
Ranked retrieval metrics	Precision-recall curve	The plot summarizes the tradeoff between the true positive rates (aka. precision) and the positive rate (aka. recall) for the predicted model for all potential cult-offs (aka. thresholds) for a test	
	Precision at k	Measuring precision at a fixed level of retrieved results or, in other words, precision at k documents (P@k)	$P@k = \frac{\text{Number of relevant document retrieved}}{k \text{ number of documents retrieved}}$
	R-Precision	Is the ratio between all the relevant documents retrieved until the rank that equals the number of relevant documents you have in your collection in total (r) to the total number of relevant documents in your collection ®	R -Precision $= \frac{r}{R}$ Where R is # of relevant documents, i.e., used as the cutoff for calculation and varied query to query, and r is the count of the number of relevant returns. For example, suppose there are 100 documents in the document collection, 30 of which are relevant ($R = 30$), the rest irrelevant. So you retrieve the first 30 documents (because 30 are relevant in total in your collection) and, say, 10 are relevant ($r = 10$). Your R-Precision is then

(continued)

Table 11.1 (continued)

Based on	Evaluation metric	Definition	Formula
	ROC Curve	It plots the true positive rate (aka. sensitivity/recall) against the false positive rate (aka. 1-specificity)	
	Normalized discounted cumulative gain (nDCG)	Measures the quality of the search results by ordering documents as very relevant, somewhat relevant, and then irrelevant at each position for the value of p	$\text{nDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p}$ Where DCG_p is discounted cumulative gain and IDCG_p is ideal DCG at position p. And $\text{IDCG}_p = \sum_{i=1}^{\lvert\text{REL}_P\rvert} \frac{2^{\text{rel}_i}-1}{\log_2(i+1)}$ Here, REL_P represents the list of relevant documents in the corpus up to position p and rel_i is graded relevance of result at the position i
	Mean average precision (mAP)	Measure that combines recall and precision for ranked retrieval results	$\text{mAP} = \frac{1}{n}\sum_{k=1}^{k=n} \text{AP}_k$ Where AP_k = the Average precision of class k and n is the number of classes

techniques to retrieve relevant biomedical information from various data sources, including scientific literature, electronic health records, and genomic data.

There are several key differences between the basic and biomedical IR systems. Biomedical IR systems require specialized knowledge and techniques to handle the complexity and heterogeneity of biomedical data, including use of controlled vocabularies and ontologies and integrating multiple data sources. Biomedical IR systems often require advanced techniques such as natural language processing, machine learning, and image analysis to retrieve and analyze biomedical data effectively.

The significance of biomedical IR lies in its ability to provide accurate and relevant information to biomedical researchers, clinicians, and other healthcare professionals, which can improve patient outcomes and accelerate progress in biomedicine. By using specialized techniques and strategies to overcome biomedical data challenges, biomedical IR systems can provide a valuable tool for accessing and analyzing the vast amounts of biomedical data generated daily. Although there are differences between biomedical information retrieval (IR) and general IR, the fundamental principles of IR still apply to both.

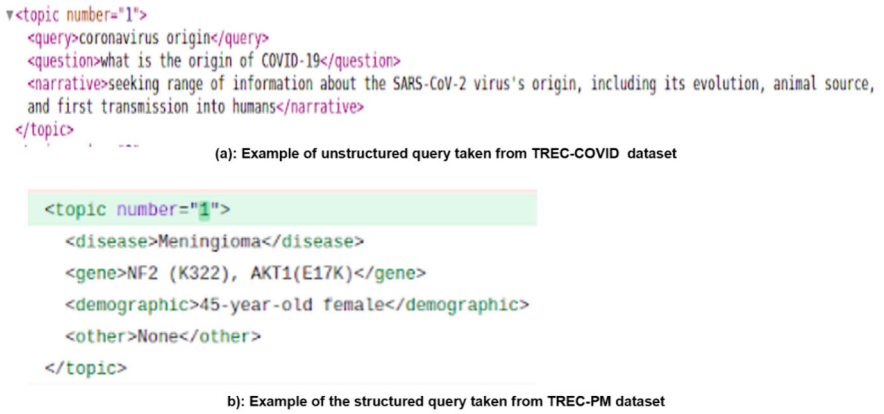


Fig. 11.5 Example of unstructured and structured query

11.4.1 Benchmark IR Datasets in Biomedical IR

The benchmark dataset for biomedical IR is generally collection of scientific articles, mostly PuMmed abstracts along with a set of queries with relevance judgments. The relevance judgements are manually annotated by experts. Communities like TREC introduced these datasets.

11.4.1.1 Types of Queries in Biomedical IR

Queries are the formal statements of the required information, like a search string in a Google search that has several keywords/phrases that match a document set with different degrees of relevance. Generally speaking, IR queries are two types:

(a) Unstructured queries—An unstructured query is a free text-type query classified as a keyword-based, question-based, or Boolean-based query.
(b) Structured queries—A structured query is the labeled keyword/concept query used to drive patient-specific information.

Figure 11.5 represents an example of an unstructured query taken from the Text REtrieval Conference (TREC)-Covid dataset (TREC-COVID Home 2002) and a structured query from the TREC-PM dataset (TREC Precision Medicine Track 2018).

11.4.1.2 Some Popular Benchmark Datasets for Biomedical IR

A lot of biomedical IR datasets are available for experiments. These datasets range from genomic research dataset, question-answering dataset, clinical datasets, etc. Some of the popular datasets are presented in Tables 11.2 and 11.3.

Table 11.2 Unstructured query-related biomedical IR datasets

Dataset	Description
TREC Genomic track https://dmice.ohsu.edu/trec-gen/ **Covid Track** https://ir.nist.gov/trec-covid/	• Focused on genomics research seeking biomedical literature Focused on COVID -19, reliable information-seeking biomedical articles
BioASQ http://bioasq.org/	Organizes challenges on biomedical semantic indexing and question-answering (QA) tasks relevant to hierarchical text classification, Information retrieval, QA from structured and unstructured data, and multi-document summarization on a large collection of MEDLINE documents
Ohsumed Hersh (Hersh et al. 1994) https://davis.wpi.edu/xmdv/datasets/ohsumed.html	It is a collection of 348,566 abstracts taken from Medline covering 270 journals from 1987 to 1991. Each query contains a brief statement about the patient followed by the information required
n2c2 • https://n2c2.dbmi.hms.harvard.edu/ • https://n2c2.dbmi.hms.harvard.edu/2022-track-1 • https://n2c2.dbmi.hms.harvard.edu/2022-track-2 • https://n2c2.dbmi.hms.harvard.edu/2022-track-3 • Wang Y, Fu S, Shen F, Henry S, Uzuner Ö, Liu H. 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. JMIR Med Inform 2020;8(11):e23375. https://doi.org/10.2196/23375	Is National NLP Clinical Challenges outgrowth of former i2b2 Center hosted challenges over various tasks like • Contextualized Medication Event Extraction • Extracting Social determinants of health (SDOH) using Social history annotation Corpus(SHAC) • Progress note understanding: Assessment and plan reasoning using the MIMIC-II dataset • Clinical semantic textual similarity, family history extraction, and clinical concept normalization

Table 11.3 Structured query related to biomedical IR datasets

Dataset	Description
TREC • Clinical decision support (CDS) http://www.trec-cds.org/2020.html • Clinical trials (CT) Train dataset: http://www.trec-cds.org/2021.html • Test dataset: https://trec.nist.gov/data/trials2021.html • Precision Medicine (PM) http://www.trec-cds.org/2016.html	• Focused on clinicians for evidence-based full-text literature to support diagnosis, treatment, and testing decisions • Focused on matching patients to relevant clinical trials • Focused on oncologists looking for evidence-based literature and clinical trials by giving the topic /query in structured format as disease, gene, and demographic component, which gave information about the type of cancer, biomarker, age, and sex of patient, respectively. Here, the existing literature is searched for the desired disease w.r.t a specific genetic profile and then matched to a clinical trial if the relevant results are achieved

11.4.2 CLEF eHealth Lab Series

The CLEF eHealth lab series provides benchmark datasets for evaluating information retrieval, text mining, and knowledge management systems in the context of electronic health records (EHRs) and related medical documents. These datasets are carefully curated to ensure that they represent real-world scenarios. There are a series of documents of various types such as clinical notes, discharge summaries, and radiology reports, among others. The benchmark datasets provided by the CLEF eHealth lab series are typically multilingual, meaning that they contain documents in multiple languages, which allows for the evaluation of systems that can handle different languages. The datasets also contain gold-standard annotations, that are used to evaluate the performance of the participating systems against a common standard. The benchmark datasets provided by the CLEF eHealth lab series are an important resource for researchers and practitioners in healthcare information retrieval and related areas. They allow for developing and evaluating new methods and algorithms and help advance the state of the art in this important field. Table 11.4 illustrates tasks related to extracting, managing, and retrieving health-related information.

11.5 Need for Query Expansion (QE) in IR

The primary goal of the IRS is to maximize the retrieval of relevant documents with minimal irrelevant documents retrieved. However, there is often a problem of term mismatch in IR systems, as the systems usually compare query and document terms on a lexical rather than a semantic level (Egozi et al. 2011; Guo et al. 2006). Even if the semantics is captured in representation, it is very limited. This can result in retrieving irrelevant documents when the user's query is too specific enough, especially since the average length of user queries is usually less than two words. (Rijsenbreg 1979). Researchers have investigated query expansion (QE) techniques to improve the performance of information retrieval systems (Rijsenberg 1979). QE

Table 11.4 CLEF eHealth lab series

Domains	Tasks
Information extraction https://sites.google.com/site/clefehealth2017/task-1 https://quaerofrenchmed.limsi.fr/	Mono/multilingual IE from death reports, NER from English/French biomedical articles, and acronym normalization
Information management	eHealth data visualization, nurse's handover reports management
Information retrieval https://sites.google.com/site/clefehealth2017/task-3 https://sites.google.com/site/clefehealth2017/task-2	Patient-centered IR and technologically assisted reviews in empirical medicine

has been widely studied in information retrieval (Sakai and Robertson 2001). QE aims to overcome the problem of term mismatch in information retrieval systems that occurs when systems compare query and document terms at a lexical level rather than a semantic level, and when user queries are short and not specific enough. Although query expansion has been shown to improve the results of information retrieval systems in many cases, it is not always the case, and the robustness of query expansion needs to be investigated to ensure that the expanded terms do not severely degrade the effectiveness of some queries. Furthermore, QE introduces the risk of query drift, which may take the search in a different direction by changing the topicality of the original query.

11.6 Query Expansion: An Overview

Query expansion, sometimes also known as term expansion, is a technique that enhances the original query with additional terms to improve retrieval performance. One of the most basic ways to accomplish this is to use a dictionary or a general thesaurus (Baeza-Yates and Ribeiro-Neto 1999; Imran and Sharan 2009). A thesaurus is a collection of words and/or phrases structured to help articulate ideas. Thesauri has been used in information retrieval to aid users in expressing or broadening their informational needs in a query. However the expansion terms need to be selected very carefully, improper selection may lead to noisy terms and thus leading to retrieval of non-relevant documents. Many questions may arise for selecting the terms suitable for query expansion. Some of the research questions that emerge are:

- What are good terms?
- Which are the best terms for query expansion?
- Where can we get the query expansion terms?
- How useful can the query expansion be?
- How can we present the terms? How can we rank the terms?
- Which ranking algorithm or method should we use for automatic query expansion, and which for interactive query expansion? Do we need different algorithms for the different types of query expansion?
- Are searchers able to recognize the good terms?
- How do the searchers select terms? What criteria do they use?
- What kind of relationships do the users there between the original query terms and the terms select?
- Is there a difference between what the user selects and what the system suggests as a good term?

11.7 Types of Query Expansion

Researchers have employed numerous query expansion approaches, yet there is no consistent system for categorizing them in the query expansion literature. In this regard, an effort is made to classify query expansion approaches into two groups: those based on the expansion mechanism and those based on the source of selection. Figure 11.6 illustrates that query expansion can be accomplished through manual, interactive, or automatic means.

Manual Query Expansion: When a user modifies a query without assistance from the system, it is referred to as manual query expansion. This approach does not involve any support from the system.

Interactive Query Expansion: Interactive query expansion involves two parties responsible for selecting expansion terms: the retrieval system and the user. The system selects and ranks terms from predetermined fields, while the user chooses which terms to add to the search. While this method is expected to be effective, it is difficult to pinpoint success or failure due to increased uncontrollable variables.

Automatic Query Expansion: Automatic query expansion (AQE) involves query modification without user control. Many laboratory experiments have been conducted on systems that incorporate AQE techniques. The process of AQE is often concealed within the larger information retrieval process, making it difficult to determine how it occurs. However, AQE is argued to be beneficial because the system can access statistical information on the usefulness of expansion terms, allowing for better term selection for the user's query. The source for selecting the term for expanding the query may be external or internal (from the corpus itself).

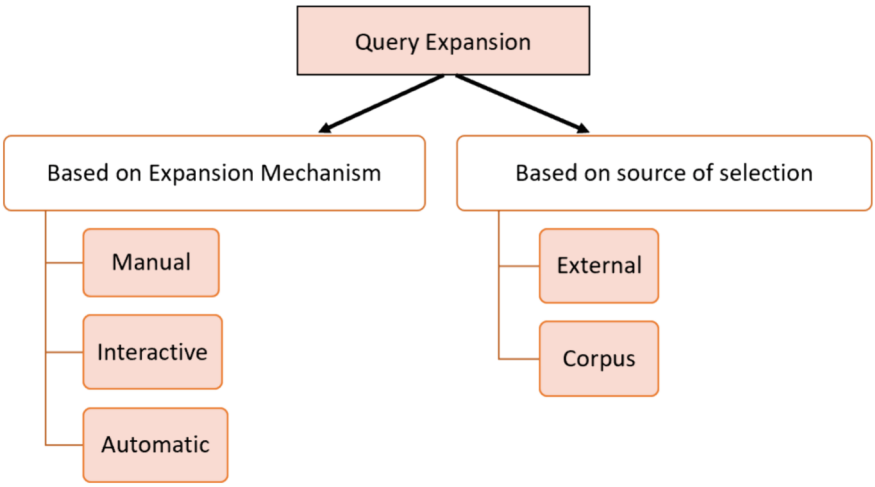


Fig. 11.6 Methods and sources used for query expansion

Query Expansion Using External Sources

In this approach, external sources, e.g., domain-specific or global thesauri, dictionaries, and lexicons are used. Gong (Gong et al. 2010) utilized WordNet as the basis for query expansion, implementing WordNet Lexical Chains and semantic similarity to group terms in the same query by semantic similarities (Mubaid and Nguyen 2006). Hersh et al. (2000) accessed thesaurus-based expansion term using UMLS Metathesaurus explained by an example below:

1. Sample query from Oshumed test collection (Hersh et al. 1994).

```
Acute tubular necrosis due to aminoglycosides, contrast dye, outcome and treatment
```

2. Sample query expanded with Metathesaurus.

```
acute tubular necrosis due
L0085410| Acute tubular necrosis| C0022672|
Kidney Tubular Necrosis, Acute
aminoglycosides
L0002556| Aminoglycosides| C0002556|
Aminoglycosides
contrast dye
L0116993| contrast| C0110625| contrast
L0013343| Dyes| C0013343| Dyes
outcome
treatment
L0040807| Treatment| C0087111| Treatment
```

3. Sample query expanded with one-level child expansion giving concepts like Aminoglycosides (first 10 words selected based on similarity with query terms)

```
Aminoglycosides
Amikacin
Amikacin Sulfate
Butirosin Sulfate
Framycetin
Genticin
Gentamicins
Hygromycin B
Kanamycin
Kantrex
```

Corpus-Based Query Expansion

Terms for expansion are selected from document collection only. The terms may be selected by forming term clusters, an automatically constructed thesaurus, an association thesaurus, or pseudo classification.

11.8 Approaches for Query Expansion

This section will discuss different methods and strategies for query expansion. We will explore various sources of information that can be used for query expansion, including controlled vocabularies, ontologies, and text corpora, as well as the different techniques for selecting and weighting expansion terms.

Several techniques have been proposed for query expansion. These techniques have been classified into various categories mainly based on data sources used for selecting the expansion terms and the mechanism used for selecting the terms. Broad categories are: global and local analyses. Global and local analyses are further divided into two subclasses. Figure 11.7 represents various query expansion techniques.

11.8.1 Global Analysis

In global analysis, query expansion techniques select the expansion terms that are semantically similar to original query terms from some hand-built knowledge sources such as ontology, thesaurus, or large corpora. Based on the source used, global analysis is further divided into two subclasses: ontology-based/knowledge source-based and corpus-based.

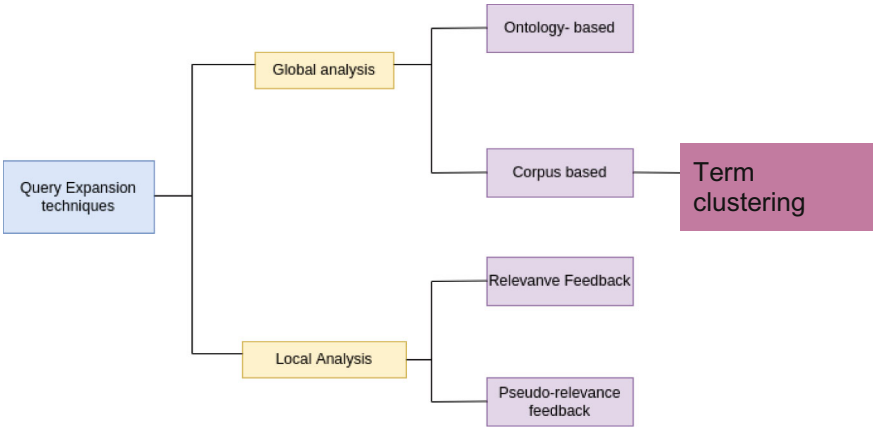


Fig. 11.7 Various query expansion techniques used in the biomedical domain

11.8.1.1 Ontology-Based Approach

This method involves searching new terms w.r.t to query terms in the ontologies. The ontologies can be domain-specific or domain-independent ontologies. Domain-independent ontologies are general purpose ontologies like linguistic ontology, WordNet that include terms like singular/plural, creating language understanding, whereas the domain-specific ontologies help to identify the terms that are semantically related to a given term based on the domain-knowledge and fail when it reaches boundary cases. Also with new advancements, new portals have been developed that provide access to different ontologies in the same domain. These portals allow you to extract terms from different ontologies in the same domain and also provide mechanisms to combine and merge results for the same input from various sources, e.g., bioportal (Whetzel et al. 2011) is an open-source, comprehensive repository of biomedical ontologies encompassing various entities from dictionaries to controlled vocabularies or knowledge structures and processes. Bioportal automatically passes the user input to different ontologies such as MeSH and SNOMED-CT and returns a combined result. When an ontology format changes, the user's query is submitted to the new version and access is provided by a web browser. It can also help to extract the entity's synonyms or related terms from specific or all ontologies. Figure 11.8 shows the example of extracting related terms of 'ACE' and 'BRCA1' from various ontologies using bioportal.

```
BRCA1 ['BRCC1', 'IRIS', 'PNCA4', 'BRCAI', 'PSCP', 'RNF53', 'BROVCA1', 'PPP1R53',
BRCA1 ['BRCAI', 'PNCA4', 'IRIS', 'BRCC1', 'RNF53', 'BROVCA1', 'PPP1R53', 'PSCP']
BRCA1 ['BRCC1', 'IRIS', 'PNCA4', 'BRCAI', 'PSCP', 'RNF53', 'BROVCA1', 'PPP1R53',
BRCA1 ['BRCC1', 'IRIS', 'PNCA4', 'BRCAI', 'PSCP', 'RNF53', 'BROVCA1', 'PPP1R53',
BRCA1 ['BRCAI', 'PNCA4', 'IRIS', 'BRCC1', 'RNF53', 'BROVCA1', 'PPP1R53', 'PSCP']
BRCA1 ['BRCC1', 'RNF53', 'FANCS', 'PPP1R53', 'BRCA1 DNA repair associated']
Ace ['CD143', 'AW208573']
ACE ['CD143', 'MVCD3', 'ACE1', 'DCP1', 'DCP', 'ICH']
Ace ['StsRR92', 'Dcp1']
ACE ['MVCD3', 'ACE1', 'CD143', 'DCP1', 'ICH', 'DCP']
Ace ['AW208573', 'CD143']
ACE ['CD143', 'MVCD3', 'ACE1', 'DCP1', 'DCP', 'ICH']
Ace ['CD143', 'AW208573']
Ace ['StsRR92', 'Dcp1']
Ace ['CD143', 'AW208573']
ACE ['CD143', 'MVCD3', 'ACE1', 'DCP1', 'DCP', 'ICH']
ACE ['MVCD3', 'ACE1', 'CD143', 'DCP1', 'ICH', 'DCP']
Ace ['AW208573', 'CD143']
Ace ['AcChE', 'ACE', 'CG17907', 'l(3)87Ed', 'ace-2', 'ache', 'AChE', 'dmAChE', 'l
ACE ['ACE1', 'CD143', 'angiotensin I converting enzyme']
```

Fig. 11.8 Synonyms extracted from various ontologies

Example of Ontology-Based Query Expansion

To improve retrieval performance, various knowledge sources (KS) are available for query expansion. The most commonly used KS are MeSH and UMLS metathesaurus (McInnes et al. 2009). MeSH is a hierarchical controlled vocabulary that the domain experts use to assign MeSH Headings (MH) to documents in PubMed, allowing users to retrieve documents that primarily discuss specific topics. Each MH is associated with several Entry Terms (ET), which are synonyms, alternate forms, or closely related terms used for indexing and retrieval. MeSH provides two methods for extracting synonyms: Exact term matching and partial/any term matching. Exact term matching techniques extract synonyms for biomedical term extracted from query with the ontology term that exactly matches to it, whereas partial term matching includes narrower/broader terms in the ontology referring the biomedical term. For exact term matching coverage may be less, however partial match may result in query drift due to presence of irrelevant terms. We present an example of ontology-based query expansion. Here we opt for exact term matching to extract terms for query expansion. UMLS metathesarus has been used here to extract query terms.

To extract synonyms from the UMLS metathesaurus, it is necessary to download and install the UMLS database and run queries using relational MRREL and MRCONSO. MRREL extracts terms with synonyms represented as 'SY' relation, and MRCONSO extracts the term names of terms with 'SY' relationship. Below is the example for query expansion using MeSH ontology as well as the process for expanding a query using MeSH/UMLS metathesaurus.

Query: Are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19? *(Text REtrieval Conference (TREC)TREC-COVID Home 2002)*

Step 1: Convert a query into a baseline query by removing stop words, lowercase, etc. Above query is converted to baseline query as below-

'patients angiotensin-converting enzyme inhibitors ace increased risk COVID-19'.

Step 2: Extract the biomedical named entities using the Metamap (Aronson and Lang 2010) tool.

Furthermore, store results like unique id, concept name preferred name in dictionary-named keywords as given below:

```
keywords=\
{
   'covid-19':
              {'id':['C5203670'],

               },
   'ACE Inhibitors':
              {
               'id':['C0003015'],
               'pref_name':['Angiotensin-Converting Enzyme Inhibitors']
              },
   'increased':
            {
              'id':['C0205217','C0442805'],
               'pref_name':['increase']
            },
   'risk':
     {
       'id':['C0035647','C4552904'],
        'pref_name':['Subject Risk']
     }
}
```

Step 3: Extract the Synonyms:

3(a) If the keyword matches the Exact Term in MeSH using bioportal, retrieve the synonyms.

```
# Enter your API by SignIn Bioportal
REST_URL = "http://data.bioontology.org"
API_KEY = "Enter your API-key here"
```

```
# Import packages

import warnings
warnings.filterwarnings('ignore')
from __future__ import print_function
import json
import argparse
import lxml.html as lh
from lxml.html import fromstring
import urllib.request, urllib.error, urllib.parse
import json
import os
from pprint import pprint
```

```
# Defining Function to extract entity from bioportal

def get_json(url):
    opener = urllib.request.build_opener()
    opener.addheaders = [('Authorization', 'apikey token=' + API_KEY)]
    return json.loads(opener.open(url).read())
```

```
def term_normalization(pref_name):
 if "," in pref_name:
   trm=pref_name.split(",")
   ent_term=trm[1].strip()+" "+trm[0]
   return ent_term.lower()
 elif "(" in pref_name:
   ls1=[]
   b=pref_name.split("(")[0].strip()
   a=re.findall("\((.*)\)",pref_name)
   ls1.append(a[0].lower())
   ls1.append(b.lower())
   return ls1
 elif "Genus: " in pref_name:
   ls1=[]
   b=pref_name.split("Genus: ")[-1].strip()
   return b
 else:
   ent_term=pref_name
   return ent_term.lower()
def get_mesh_synonyms(mshterm):
 m=[]
 print((mshterm))
 mshterm=mshterm.replace(" ","%20")
 meta_data=get_json(REST_URL +
"/search?q="+mshterm+"&include=prefLabel,definition,synonym,semanticType&ontologies=MESH
")["collection"]
 for i in range(len(meta_data)):
```

```
      a=meta_data[i].get("prefLabel").replace(" ","%20")
      if a.lower()== mshterm.lower():
        q=meta_data[i].get("synonym")
        if q != None  :
          for item in q:
            item=term_normalization(item)
            if type(item)==list:
              m.extend(item)
            else:
              m.append(item)

return list(set(m))
```

Here get_mesd_id () matches the key:

if key is available:
  Retrieve synonyms
Elif replace the key with the preferred term if available:
  Retrieve synonyms

```
# Synonyms extraction

mesid=[]
for i in keywords:#keywords.keys():
 i=i.strip()
 concept_mshid=get_mesd_id(i,"concept")
 if len(concept_mshid)!=0:
  print(i,concept_mshid)
  mesid.extend(concept_mshid)
  # get synonyms
  symn=get_mesh_synonyms(i)
  print("   Symn:\n",symn)
 else:
```

```
  #print(i)
  try:
   concept_mshid=get_mesd_id(keywords[i]['pref_name'][0],"concept")
   i=keywords[i]['pref_name'][0]
   print(i,concept_mshid)
   mesid.extend(concept_mshid)
   print("-->",mesid)
   symn=get_mesh_synonyms(i)
   print("   Symn:\n",symn)
   break
  except:
   pass
```

3(b) elif Match the Concept Unique Identifier (CUI) of the term to the UMLS metathesaurus MRREL table and use MRCONSO to retrieve its name:

Query: select STR from MRCONSO where CUI in (select CUI2 from MRREL where SAB = 'MSH' and CUI1 = 'term cui' and REL = 'SY');

Step 4: Expand the query with the relevant terms.

```
are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk
for COVID-19 \'covid 19 pandemic\' \'sars coronavirus 2 infection\' \'2019-ncov
infection\' \'sars-cov-2 infection\' \'2019-ncov diseases\' \'sars cov 2 infection\' \'covid
19 virus disease\' \'2019 novel coronavirus disease\' \'covid-19 pandemics\' \'2019-
ncov disease\' \'covid 19 virus infection\' \'covid-19 virus infection\' \'covid-19 virus
infections\' \'covid-19 virus diseases\' \'2019 novel coronavirus infection\' \'2019-ncov
infections\' \'sars-cov-2 infections\' \'covid19\' \'2019 ncov disease\' \'severe acute
respiratory syndrome coronavirus 2 infection\' \'coronavirus disease-19\' \'coronavirus
disease 19\' \'coronavirus disease 2019\' \'2019 ncov infection\' \'covid-19 virus
disease\' \'covid 19\' \'covid-19 pandemic\' \'angiotensin i-converting enzyme
inhibitors\' \'angiotensin i-converting enzyme inhibitor\' \'angiotensin i converting
enzyme inhibitor\' \'angiotensin i converting enzyme inhibitors\' \'angiotensin-
converting enzyme antagonists\' \'angiotensin converting enzyme antagonists\'
\'angiotensin-converting enzyme inhibitor\' \'ace inhibitors\' \'kininase ii antagonists\'
\'ace inhibitor\' \'angiotensin-converting enzyme inhibitors\' \'angiotensin converting
enzyme inhibitor\' \'angiotensin converting enzyme inhibitors\' \'kininase ii inhibitor\'
\'kininase ii inhibitors\' \'client\' \'patient\' \'clients\' \'increased risk\'
```

Step 5: Run the query using Elasticsearch and store results in a text file, e.g., final_query_20_expansion.txt.

Step 6: Evaluate the precision result using trec_eval *(Text REtrieval Conference (TREC) Trec_eval, 2008)* using the below syntax:

Syntax:./trec_eval -q -m P.5,10,15,20,25 <judgement_file_path> <result_file_path>

E.g.:

./trec_eval -q -m P.5,10,15,20,25 qrels final_query_20_expansion.txt.

Here, the judgment file path is "qrels" and the Elasticsearch result file is "final_query_20_expansion.txt".

Output:

```
covid-19 ['D000086382']
covid-19
  Symn:
 ['sars cov 2 infection', 'covid19', 'sars-cov-2 infections', '2019 novel coronavirus disease',
'2019 ncov infection', '2019 novel coronavirus infection', 'covid-19 pandemic', 'sars-cov-2
infection', 'coronavirus disease 19', '2019-ncov infection', '2019-ncov infections', 'covid-19
virus diseases', 'coronavirus disease-19', 'coronavirus disease 2019', 'covid-19 virus disease',
'covid 19 virus disease', 'covid-19 virus infection', '2019-ncov diseases', 'covid 19 pandemic',
'2019-ncov disease', 'covid-19 virus infections', 'sars coronavirus 2 infection', 'severe acute
respiratory syndrome coronavirus 2 infection', 'covid 19', 'covid-19 pandemics', 'covid 19 virus
infection', '2019 ncov disease']
Angiotensin-Converting Enzyme Inhibitors ['D000806']
--> ['D000086382', 'D000806']
Angiotensin-Converting Enzyme Inhibitors
  Symn:
 ['ace inhibitor', 'angiotensin i-converting enzyme inhibitor', 'angiotensin-converting enzyme
inhibitors', 'angiotensin converting enzyme inhibitor', 'angiotensin i-converting enzyme
inhibitors', 'angiotensin-converting enzyme inhibitor', 'kininase ii inhibitor', 'angiotensin
converting enzyme inhibitors', 'kininase ii inhibitors', 'angiotensin converting enzyme
antagonists', 'ace inhibitors', 'angiotensin i converting enzyme inhibitor', 'kininase ii
antagonists', 'angiotensin i converting enzyme inhibitors', 'angiotensin-converting enzyme
antagonists']
['D000086382', 'D000806']
```

11.8.1.2 Corpus-Based Approach

For corpus-based approaches, the content of the entire corpus plays a crucial role in selecting expansion terms. Early research in this area employed statistical techniques, such as the co-occurrence-based approach, which seeks to establish correlations between terms within a given corpora. This method operates under the assumption that words co-occurring in the same document are likely to share some semantic relationship. This semantic relation proves helpful in effectively distinguishing between relevant and irrelevant documents. Individual query terms are assigned co-occurring terms, a weight may be assigned to these terms based on some co-occurrence score. These assigned weights serve as indicators of the relevance of each term within the expanded query, which is subsequently used to rank the retrieved documents. Following this, each term is evaluated based on its similarity to the query term, resulting in the assignment of a similarity score. Only those terms with higher weights are included in the expansion query. Moreover, the expanded terms can be further categorized into term clustering and concept-based terms, facilitating a more refined expansion process.

Term Clustering: It is a significant aspect of co-occurrence-based query expansion to facilitate the selection of terms. Here the terms in the corpus, generally represented as word embedding, are organized into clusters. The resulting clusters offer a more

coherent representation of the underlying semantic structure within the corpus. Due to the inherent semantic relatedness of embeddings, terms that share similar meanings or are conceptually related tend to be grouped together, allowing for a more nuanced understanding of the context. By utilizing term clustering in query expansion, initially a cluster is selected and then as a second step, and the expansion term is selected from the cluster. This process facilitates a computationally efficient method with a more focused term selection. Furthermore, the incorporation of term clustering helps overcome vocabulary mismatch by allowing the related terms to be candidate expansion terms. This may result in enhancing the recall of the result.

In the biomedical domain, term clustering becomes particularly valuable due to the presence of specialized terminology and domain-specific concepts. Biomedical term clustering techniques often take into account domain-specific knowledge, such as medical ontologies or thesauri, to improve the accuracy and relevance of the clusters. Within the biomedical domain, several embedding methods have been suggested to facilitate this process. Table 11.5 provides an overview of various embedding techniques and models proposed specifically for the biomedical domain. These techniques contribute to a more comprehensive understanding of term relationships, enabling more effective query expansion strategies.

Table 11.6 summarizes various embedding visualization techniques developed for exploring and interpreting biomedical data.

Example of Corpus-Based Query Expansion

To demonstrate corpus-based query expansion, we utilized the CORD-19 dataset and employed word2vec to create word embeddings. The following steps were taken to achieve this goal:

1. Preprocess the dataset by:
   (a) Removestopwords, URLs and punctuation marks
   (b) Perform Case conversionConverta number into words
   (c) Convertspecial symbol like plus to meanigful word such as ca+ to caplus
   (d) Tokenize the word
   (e) Lemmatize the word to know the actual word
2. Use scispacy (https://allenai.github.io/scispacy/) to extract biomedical entities and concatenate entities with more than two words using hyphens. For example, if the entity '2019-ncov infections' is extracted from the dataset, it would be concatenated into a single entity as '2019-ncov-infections.'
3. Train the model with a 100-dimension vector and 100 k vocabulary size. See Annexure 1 for the top 20 similar words for biomedical entities like COVID-19, coronavirus, and treatment, extracted from sample queries.
4. Select the top 5 terms for each biomedical entity extracted from the query for query expansion. For example, if the query is related to COVID-19, the expanded query terms would be COVID-19, COVID-19 infections, COVID-19 disease, COVID-19, and COVID-19 infections. Hyphens are used to concatenate biomedical phrases as entities and possible combinations are used for query

Table 11.5 Various embedding techniques/models proposed in the biomedical domain

Embedding models	Embedding type	Description	Source
BioWordVec	Word embedding	Used Pubmed and Clinical notes from the MIMIC-III Clinical database with a vector of 200 dimensions for BioWordVec and 700 dimensions for BioSentVec	https://github.com/ncbi-nlp/BioSentVec
BioSentVec	Sentence embeddings		
CORD-19 Swivel Embeddings	Word embeddings	Google introduced the text embedding module for TF-HUB to support the researchers in analyzing the natural language text related to COVID-19. These embeddings were trained on the CORD-19 dataset (*Allenai/cord19: Get Started With CORD-19*, 2020) using titles, abstracts, body text, references, and authors of these articles	https://colab.research.google.com/github/tensorflow/hub/blob/master/examples/colab/cord_19_embeddings.ipynb#scrollTo=9VusdTAH0isl
BioBERT Lee (Lee et al. 2019)	Word embedding	It is a BERT Devlin (Devlin et al. 2018) model and pre-trained on PubMed abstracts and full PMC text articles for biomedical text mining tasks such as named entity recognition, question-answering, relation extraction, model, and its code are publicly available	https://github.com/dmis-lab/biobert
ClinicalBERT	Word embedding	It is a pertained model on clinical notes publicly available in the MIMIC-III (Johnson et al. 2019) database having ~ 2M notes. Various versions of ClinicalBERT are presented as 1. Clinical BERT 2. Clinical BioBERT 3. Discharge summary BERT 4. Discharge summary BioBERT	https://github.com/EmilyAlsentzer/clinicalBERT
SciBERT	Word and sentence embedding based on the version	It is a trained model on semantic scholar [13] full-text papers where 18% are computer science, and 82% are biomedical papers. Currently, there are 4 versions of SciBERT publicly available, and they are as follows: 1. Cased (vocabulary contains both lowercase and uppercase) 2. Uncased (vocabulary converted into lowercase) 3. Models using BaseVocab (fine-tuned on BERT-base model) 4. Model using SciVocab (scientific vocabulary built on sentences)	https://github.com/allenai/scibert

Table 11.6 Various embedding visualization techniques

Visualization technique	Description
T-distributed stochastic neighbor (t-SNE) (Maaten and Hinton 2008)	A probabilistic approach describes the high-dimensional vector to a low-dimensional space by preserving their neighborhood identities
Principle Component Analysis (PCA) (Wold et al. 1987)	The reduction technique depends on finding orthogonal linear combinations of initial feature dimensions that extract the most variance and project data points onto them
Unified Manifold Approximation and Projection (UMAP) (Mclnnes et al. 2018)	It is a reduction technique based on manifold learning techniques and ideas from Riemannian geometry and topological data analysis. In contrast with t-SNE, UMAP preserves more of the global structure with high runtime performance and has better visualization quality
K-Means (Li and Wu 2012)	It is an unsupervised clustering algorithm that can be applied to high-dimensional datasets entities/sentences to a d-dimensional vector assigned to a predefined number of clusters. These clusters can be used to draw a group of similar elements based on their embedding representation and can be visualized with the help of Word Cloud

expansion (e.g., 'COVID-19-infection' can be written as 'COVID-19 infection' or 'COVID-19 infection').

5. Pass the query using Elasticsearch, and evaluate the results using trec_eval as stated above in the case of knowledge sources.

This is a very simple experiment on query expansion in biomedical domain. The objective is to introduce the readers to the basic steps in query expansion, observe the terms obtained and their relation to the orignal query terms and how they may effect the retrieval in positive or negative way. The experiment will help users in understanding the basics of query expansion that has now become a part of modern IRS. We compared the above-stated approaches on a set of sample queries. Table 11.7 illustrates the expanded query using knowledge sources represented as EQ_KS and word embedding techniques Word2Vec on the corpus represented as EQ_Corpus.

Tables 11.8, 11.9 and 11.10 present the precision results for the original and expanded queries using knowledge sources and corpus-based word embeddings at ranks 5, 10, 15, 20, and 25. Annexure 2 displays precision graph plots for each query.

Query expansion is still a field of research and should be carried out very meticulously. Thus, the empirical results may or may not show an improvement in the result. In many cases, filtering of expansion terms needs to be done to eliminate the irrelevant expansion terms.

Table 11.7 Expanded query for knowledge base and embedding using the corpus

Query	Baseline query	EQ_KS	EQ_Corpus
How does the coronavirus respond to changes in the weather	Coronavirus change weather	How does the coronavirus respond to changes in the weather \'rabbit coronaviruses\' \'rabbit coronavirus\' \'coronaviruses\' \'coronavirus\' \changes\' \'weather\' \'fog\' \'fogs\'	How does the coronavirus respond to changes in the weather \'coronavirus\' \'weather\' \'corona virus\' \'coronavirus infection\' \'betacoronavirus\' \'coronavirus-2019\' \'zoonotic-coronavirus\' \'changes\' \'weather\' \'climate\' \'meteorological\' \'climatic\' \'air pollution\' \'rainfall\' \'alteration\' \'altered\' \'difference\' \'shift\' \'variation\'
How long can the coronavirus live outside the body	Coronavirus lives outside the body	How long can the coronavirus live outside the body \'rabbit coronaviruses\' \'rabbit coronavirus\' \'coronaviruses\' \'coronavirus\' \'live\' \'extrinsic\' \'external\' \'outer\' \'outside\' \'body\' \'human body\' \'human bodies\'	How long can the coronavirus live outside the body \'corona virus\' \'coronavirus\' \'coronavirus infection\' \'betacoronavirus\' \'coronavirus-2019\' \'zoonotic-coronavirus\' \'live\' \'outside\' \'inside\' \'traveled\' \'confined\' \'body\'

(continued)

Table 11.7 (continued)

Query	Baseline query	EQ_KS	EQ_Corpus
What type of hand sanitizer is needed to destroy COVID-19?	Hand sanitizers destroy Covid-19	What type of hand sanitizer is needed to destroy Covid-19 \'covid 19 pandemic\' \'sars coronavirus 2 infection\' \'2019-ncov infection\' \'sars-cov-2 infection\' \'2019-ncov diseases\' \'sars cov 2 infection\' \'covid 19 virus disease\' \'2019 novel coronavirus disease\' \'covid-19 pandemics\' \'2019-ncov disease\' \'covid 19 virus infection\' \'covid-19 virus infection\' \'covid-19 virus infections\' \'covid-19 virus diseases\' \'2019 novel coronavirus infection\' \'2019-ncov infections\' \'sars-cov-2 infections\' \'covid19\' \'2019 ncov disease\' \'severe acute respiratory syndrome coronavirus 2 infection\' \'coronavirus disease-19\' \'coronavirus disease 19\' \'coronavirus disease 2019\' \'2019 ncov infection\' \'covid-19 virus disease\' \'covid 19\' \'covid-19 pandemic\' \'hand sanitizers\' \'destroy\' \'hand antiseptics\' \'hand disinfectants\' \'needed\'	What type of hand sanitizer is needed to destroy covid-19 \'covid-19 infection\' \'covid 19 disease\' \'covid19\' \'covid 19 infections\' \'covid 2019\' \'sars-cov-2 infection\' \'wearing face masks\' \'eye protection\' \'hand washing\' \'hand hygiene\' \'alcohol based hand rubs\' \'hand sanitizer\' \'needed\' \'destroy\' \'resist\' \'compete\' \'kill\' \'degrade\' \'disrupt\'

(continued)

Table 11.7 (continued)

Query	Baseline query	EQ_KS	EQ_Corpus
Are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19?	Patients Angiotensin-converting enzyme inhibitors increased the risk of COVID-19	Are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19 \'covid 19 pandemic\' \'sars coronavirus 2 infection\' \'2019-ncov infection\' \'sars-cov-2 infection\' \'2019-ncov diseases\' \'sars cov 2 infection\' \'covid 19 virus disease\' \'2019 novel coronavirus disease\' \'covid-19 pandemics\' \'2019-ncov disease\' \'covid 19 virus infection\' \'covid-19 virus infection\' \'covid-19 virus infections\' \'covid-19 virus diseases\' \'2019 novel coronavirus infection\' \'2019-ncov infections\' \'sars-cov-2 infections\' \'covid19\' \'2019 ncov disease\' \'severe acute respiratory syndrome coronavirus 2 infection\' \'coronavirus disease-19\' \'coronavirus disease 19\' \'coronavirus disease 2019\' \'2019 ncov infection\' \'covid-19 virus disease\' \'covid 19\' \'covid-19 pandemic\' \'angiotensin i-converting enzyme inhibitors\' \'angiotensin i-converting enzyme inhibitor\' \'angiotensin i-converting enzyme inhibitor\' \'angiotensin i-converting enzyme inhibitors\' \'angiotensin-converting enzyme antagonists\' \'angiotensin converting enzyme antagonists\' \'angiotensin-converting enzyme inhibitor\' \'ace inhibitors\' \'kininase ii antagonists\' \'ace inhibitor\' \'angiotensin-converting enzyme inhibitors\' \'angiotensin converting enzyme inhibitor\' \'angiotensin converting enzyme inhibitors\' \'kininase ii inhibitor\' \'kininase ii inhibitors\' \'client\' \'patient\' \'clients\' \'increased risk\'	Are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19 \'covid-19 infection\' \'covid 19 disease\' \'covid19\' \'covid 19 infections\' \'covid 2019\' \'sars-cov-2 infection\' \'ace\' \'angiotensin-converting enzyme 2\' \'angiotensin-ii\' \'ace 2\' \'angiotensin-converting enzyme inhibitors\' \'increased\' \'risk\' \'high risk\' \'hazard\' \'likelihood\' \'incidence\' \'patient\' \'case\' \'hospitalized\' \'child\' \'diagnosed\'

(continued)

Table 11.7 (continued)

Query	Baseline query	EQ_KS	EQ_Corpus
What evidence is there for the value of hydroxychloroquine in treating Covid-19?	Evidence hydroxychloroquine treating Covid-19	What evidence is there for the value of hydroxychloroquine in treating Covid-19 \'covid 19 pandemic\' \'sars coronavirus 2 infection\' \'2019-ncov infection\' \'sars-cov-2 infection\' \'2019-ncov diseases\' \'sars cov 2 infection\' \'covid 19 virus disease\' \'2019 novel coronavirus disease\' \'covid-19 pandemics\' \'2019-ncov disease\' \'covid 19 virus infection\' \'covid-19 virus infection\' \'covid-19 virus infections\' \'covid-19 virus diseases\' \'2019 novel coronavirus infection\' \'2019-ncov infections\' \'sars-cov-2 infections\' \'covid19\' \'2019 ncov disease\' \'severe acute respiratory syndrome coronavirus 2 infection\' \'coronavirus disease-19\' \'coronavirus disease 19\' \'coronavirus disease 2019\' \'2019 ncov infection\' \'covid-19 virus disease\' \'covid 19\' \'covid-19 pandemic\' \'treating\' \'plaquenil\' \'oxychloroquine\' \'oxychlorochin\' \'2-((4-((7-chloro-4-quinolinyl)amino)pentyl)ethylamino)-ethanol\' \'hydroxychlorochin\' \'hydroxychloroquine sulfate',’Hydroxychloroquine Sulfate (1:1) Salt',’2-((4-((7-chloro-4-quinolyl)amino)pentyl)ethylamino)ethanol\'	What evidence is there for the value of hydroxychloroquine in treating Covid-19 \'covid-19 infection\' \'covid 19 disease\' \'covid19\' \'covid 19 infections\' \'covid 2019\' \'sars-cov-2 infection\' \'hydroxychloroquine\' \'chloroquine\' \'azithromycin\' \'remdesivir\' \'lopinavir\' \'treating\' \'treat\' \'treatment\' \'managing\' \'alternative treatment\' \'therapy\'

(continued)

Table 11.7 (continued)

Query	Baseline query	EQ_KS	EQ_Corpus
Is remdesivir an effective treatment for COVID-19	Remdesivir effective treatment COVID-19	Is remdesivir an effective treatment for COVID-19 \'covid 19 pandemic\' \'sars coronavirus 2 infection\' \'2019-ncov infection\' \'sars-cov-2 infection\' \'2019-ncov diseases\' \'sars cov 2 infection\' \'covid 19 virus disease\' \'2019 novel coronavirus disease\' \'covid-19 pandemics\' \'2019-ncov disease\' \'covid 19 virus infection\' \'covid-19 virus infection\' \'covid-19 virus infections\' \'covid-19 virus diseases\' \'2019 novel coronavirus infection\' \'2019-ncov infections\' \'sars-cov-2 infections\' \'covid19\' \'2019 ncov disease\' \'severe acute respiratory syndrome coronavirus 2 infection\' \'coronavirus disease-19\' \'coronavirus disease 19\' \'coronavirus disease 2019\' \'2019 ncov infection\' \'covid-19 virus disease\' \'covid 19\' \'covid-19 pandemic\' \'gs-5734\' \'13c3-gs-5734\' \'13c3 gs 5734\' \'13c3 gs-5734\' \'gs 5734\' \'gs-465124\' \'2-ethylbutyl (2S)-2-(((2R, 3S, 4R, 5R)-5-(4-aminopyrrolo(2,1-f) (1,2,4)triazin-7-yl)-5-cyano-3,4-dihydroxytetrahydrofuran-2-yl) methoxy)(phenoxy) phosphoryl) amino) propanoate\' \'gs-829143',\'L-alanine, N-((S)-hydroxyphenoxyphosphinyl)-, 2-ethylbutyl ester, 6-ester with 2-C-(4-aminopyrrolo(2,1-f)(1,2,4)triazin-7-yl)-2,5-anhydro-d-altrononitrile\' \'effective\' \'therapeutic procedure\' \'therapeutic\' \'treatment\' \'therapy\' \'therapies\' \'tr eatments\'	Is remdesivir an effective treatment for COVID-19 \'covid-19 infection\' \'covid 19 disease\' \'covid19\' \'covid 19 infections\' \'covid 2019\' \'sars-cov-2 infection\' \'hydroxychloroquine\' \'chloroquine\' \'favipiravir\' \'remdesivir\' \'lopinavir\' \'umifenovir\' \'supportive treatment\' \'antiviral therapy\' \'treatment\' \'treat\' \'supportive care\' \'therapy\' \'effective\' \'efficacious\' \'cost effective\' \'appropriate\' \'efficient\' \'successful\'

Table 11.8 Precision results for original queries

Query	Org_ P@5	Org_ P@10	Org_ P@15	Org_ P@20	Org_ P@25
How does the coronavirus respond to changes in the weather	0.8	0.4	0.46	0.5	0.4
How long can the coronavirus live outside the body	0.4	0.2	0.13	0.15	0.12
What type of hand sanitizer is needed to destroy COVID-19?	0.2	0.3	0.2	0.15	0.12
Are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19?	0.8	0.9	0.86	0.9	0.92
What evidence is there for the value of hydroxychloroquine in treating COVID-19?	1	0.8	0.8	0.85	0.88
Is remdesivir an effective treatment for COVID-19	1	1	1	0.95	0.96

Table 11.9 Precision results of expanded queries using knowledge sources techniques

Query	EQ_ KS_ P@5	EQ_ KS_ P@10	EQ_ KS_ P@15	EQ_ KS_ P@20	EQ_ KS_ P@25
How does the coronavirus respond to changes in the weather	0.2	0.1	0.6	0.5	0.4
How long can the coronavirus live outside the body	0.4	0.4	0.2667	0.2	0.16
What type of hand sanitizer is needed to destroy COVID-19?	0.6	0.3	0.2667	0.2	0.2
Are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19?	1	1	1	0.95	0.92
What evidence is there for the value of hydroxychloroquine in treating COVID-19?	1	0.7	0.6667	0.65	0.68
Is remdesivir an effective treatment for COVID-19	0.6	0.4	0.4667	0.4	0.32

11.8.2 Local Analysis

Local analysis is based on the selection of expansion terms from some relevant documents retrieved in response to the user's query. The assumption being that the relevant documents may contain some terms related to the query, if selected for expansion may improve the retrieval efficiency of the IRS. However, this requires careful selection of candidate documents and candidate terms to be included in the process. Local analysis methodology can be categorized in two ways: (1) Relevance Feedback (RF) and (2) Pseudo-Relevance feedback.

Table 11.10 Precision result of expanded queries using word embedding techniques

Query	EQ_ Corpus_ P@5	EQ_ Corpus_ P@10	EQ_ Corpus_ P@15	EQ_ Corpus_ P@20	EQ_ Corpus_ P@25
How does the coronavirus respond to changes in the weather	0.8	0.5	0.5333	0.55	0.48
How long can the coronavirus live outside the body	0	0.1	0.1333	0.1	0.12
What type of hand sanitizer is needed to destroy COVID-19?	0	0.1	0.2	0.2	0.1667
Are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19?	1	1	1	1	0.96
What evidence is there for the value of hydroxychloroquine in treating COVID -19?	1	1	0.8667	0.9	0.88
Is remdesivir an effective treatment for COVID-19	0.8	0.8	0.73333	0.7	0.6

11.8.2.1 Relevance Feedback (RF)

In this approach, the user's feedback about the documents are collected in response to the original user query by stating whether the document is relevant or not. The set of relevant documents become the candidate documents for selecting expansion terms. Rocchio (1971) introduced the first relevance feedback algorithm, aka. Rocchio's method. This approach is further categorized into two types: explicit feedback and implicit feedback. Harman (1992), Salton and Buckley (1990) have done this with explicit feedback. The user explicitly evaluates the relevance of retrieved documents, whereas Zhoun et al. (2012) have done implicit feedback, the user activity on the set of documents retrieved in response to the original query is used to infer the user's preferences indirectly. Sankhavara (2018) has seen that RF-based techniques outperform biomedical document retrieval, including the cost of human judgments.

11.8.2.2 Pseudo-Relevance Feedback (PRF)

(Aka) Blind feedback, the document feedback is automatically generated by directly using the top-ranked documents retrieved in response to the original user query. The assumption being that if we observe the result of a reasonable may be not a perfect IRS, at least top k documents are expected to be relevant. Croft and Harper (1979) were the first to propose it by employing this technique on probabilistic models. Pan et al. (2018) used pseudo-relevance feedback based on proximity information for clinical decision support (CDS) retrieval systems. Xu and Croft (2017) studied query expansion techniques using global and local context analyses. In some cases,

the top retrieved document might not be the best strategy for expansion terms as they might contain similar contents that lead to similar expansion terms and will not be useful.

11.9 Summary

Biomedical information retrieval (IR) has become increasingly important in recent years due to the explosive growth of biomedical data and the need for efficient and effective access to this information. In this chapter, we have provided an introduction to the field of biomedical IR, including an overview of its challenges, opportunities, and key techniques. We have discussed the unique characteristics of biomedical data, such as its heterogeneity, complexity, and diversity, and how these characteristics pose challenges for information retrieval systems. We have also described some key techniques and approaches to address these challenges. The field of biomedical IR offers exciting opportunities for researchers and practitioners to make significant contributions to developing new methods and technologies that can improve access to biomedical data and enhance healthcare delivery. As the volume and complexity of biomedical data continue to grow, there is a pressing need for advanced IR systems that can effectively and efficiently search, manage, and retrieve this information that too in real time.

Annexure 1

Display the most-similar words extracted from word2vec model trained on Cord-19 dataset with 100 dimensions for the entities extracted from query mentioned in Sect. 10.2.1.

```
1 word_vectors.most_similar("covid-19", topn=20)
```

```
[('covid-19', 0.8536357283592224),
 ('covid-19-infection', 0.7515257596969604),
 ('covid-19-disease', 0.7425595552192688),
 ('covid19', 0.7316215038299956),
 ('covid-19-infections', 0.7061384320259094),
 ('covid-2019', 0.6736037731170654),
 ('sars-cov-2-infection', 0.6532856225967407),
 ('covid-19-pandemic', 0.6490296125411987),
 ('sars-cov-2', 0.6273252964019775),
 ('sars-cov-2-pandemic', 0.6220694780349731),
 ('covid-19-epidemic', 0.6186789274215698),
 ('covid-19-pneumonia', 0.6028560400009155),
 ('covid-19-outbreak', 0.5918027758598328),
 ('sars-cov-2-epidemic', 0.5776214599609375),
 ('2019-ncov', 0.5767941474914551),
 ('covid-19-virus', 0.5653988122940063),
 ('sars-cov2', 0.5627263188362122),
 ('coronavirus-disease-2019', 0.56017005443573),
 ('2020', 0.5579708218574524),
 ('covid', 0.5559819340705872)]
```

```
1 word_vectors.most_similar("coronavirus", topn=20)
```

```
[('corona-virus', 0.735944926738739),
 ('coronavirus-infection', 0.7029961943626404),
 ('betacoronavirus', 0.6740383505821228),
 ('coronavirus-', 0.670659065246582),
 ('coronavirus-2019', 0.6651095151901245),
 ('zoonotic-coronavirus', 0.6487312316894531),
 ('coronavirus-2', 0.6184579730033875),
 ('coronavirus-sars-cov-2', 0.6116166710853577),
 ('coronavirus-disease', 0.5941445827484131),
 ('beta-coronavirus', 0.5933305025100708),
 ('sars-cov-2-coronavirus', 0.5717233419418335),
 ('coronavirus-strain', 0.5708651542663574),
 ('human-coronavirus', 0.5661829113960266),
 ('coronaviruses', 0.5621484518051147),
 ('virus', 0.5608371496200562),
 ('coronavirus-disease-2019', 0.5466081500053406),
 ('severe-acute-respiratory-syndrome', 0.5417728424072266),
 ('coronavirus-diseases', 0.5407667756080627),
 ('coronavirus-infections', 0.5370599031448364),
 ('-cov', 0.5329594016075134)]
```

```
1 word_vectors.most_similar("remdesivir", topn=20)
```

```
[('favipiravir', 0.7558713555335999),
 ('hydroxychloroquine', 0.7403241395950317),
 ('lopinavir', 0.7269766330718994),
 ('chloroquine', 0.6878768801689148),
 ('umifenovir', 0.6749991178512573),
 ('ritonavir', 0.6666966676712036),
 ('chloroquine-phosphate', 0.6570385694503784),
 ('tocilizumab', 0.6456430554389954),
 ('arbidol', 0.6443508863449097),
 ('gs-5734', 0.6383938789367676),
 ('fda-approved', 0.6327259540557861),
 ('azithromycin', 0.6317389607429504),
 ('tenofovir', 0.6205436587333679),
 ('repurposed-drugs', 0.6185039281845093),
 ('mycophenolic-acid', 0.6068700551986694),
 ('sarilumab', 0.6053802967071533),
 ('nitazoxanide', 0.6053063273429871),
 ('ivermectin', 0.6032578945159912),
 ('t-705', 0.6021537780761719),
 ('derivative-hydroxychloroquine', 0.5990582704544067)]
```

```
1 word_vectors.most_similar("evidence", topn=20)
```

```
[('clinical-evidence', 0.6651832461357117),
 ('clinical-observations', 0.5804057717323303),
 ('empirical-evidence', 0.5781129598617554),
 ('literature', 0.5701152682304382),
 ('information', 0.5466405153274536),
 ('knowledge', 0.5409126281738281),
 ('data', 0.5115118622779846),
 ('epidemiological-evidence', 0.5044851303100586),
 ('finding', 0.5011138319969177),
 ('genetic-evidence', 0.4991692304611206),
 ('empirical-support', 0.49516838788986206),
 ('rationale', 0.4827805575522583),
 ('experimental-evidence', 0.4791420102119446),
 ('research', 0.4768190979575806),
 ('link', 0.4701811969280243),
 ('scientific-evidence', 0.4696579873561859),
 ('summary', 0.46750617027282715),
 ('insight', 0.4652464687824249),
 ('observation', 0.45829904079437256),
 ('study', 0.4579997062683105)]
```

```
1 word_vectors.most_similar("outside", topn=20)
```

```
[('inside', 0.5810935497283936),
 ('travelled', 0.5360012054443359),
 ('hubei', 0.5293998718261719),
 ('confined', 0.5234470963478088),
 ('hubei-province', 0.5206879377365112),
 ('away', 0.5014082193374634),
 ('wuhan', 0.4984980523586273),
 ('elsewhere', 0.4978253245353699),
 ('traveled', 0.48851120471954346),
 ('throughout', 0.4762432873249054),
 ('overseas', 0.47358644008636475),
 ('middle-east', 0.47140824794769287),
 ('room', 0.4677923023700714),
 ('travel', 0.4665108323097229),
 ('nearby', 0.463786780834198),
 ('africa', 0.46168264746665955),
 ('within', 0.46132946014404297),
 ('visited', 0.46129897236824036),
 ('abroad', 0.45992299914360046),
 ('around', 0.4585128724575043)]
```

```
1 word_vectors.most_similar("effective", topn=20)
```

```
[('efficacious', 0.793621301651001),
 ('efficient', 0.7387054562568665),
 ('cost-effective', 0.6182395219802856),
 ('appropriate', 0.5935278534889221),
 ('successful', 0.5811464190483093),
 ('reliable', 0.5752586722373962),
 ('effectiveness', 0.5678553581237793),
 ('feasible', 0.5533955097198486),
 ('promising', 0.5465052127838135),
 ('proper', 0.5444860458374023),
 ('optimal', 0.524872899055481),
 ('suitable', 0.5236702561378479),
 ('ineffective', 0.5099552869796753),
 ('useful', 0.509791374206543),
 ('stringent', 0.508504331111908),
 ('adequate', 0.508224606513977),
 ('vaccine', 0.5063185691833496),
 ('reasonable', 0.4995389282703397),
 ('effectively', 0.4936229884624481),
 ('efficacy', 0.49065935611724854)]
```

```
1 word_vectors.most_similar("live", topn=20)
```

```
[('live-vaccines', 0.6154671907424927),
 ('vaccine-viruses', 0.6145454049110413),
 ('attenuated', 0.5392789840698242),
 ('arkdpi-derived', 0.5325135588645935),
 ('ib-vaccines', 0.5254157185554504),
 ('virus-vaccines', 0.5221821665763855),
 ('inactivated', 0.5215327143669128),
 ('h120-vaccine', 0.5178202986717224),
 ('purchased', 0.5141024589538574),
 ('killed', 0.5057281255722046),
 ('vector', 0.5052645802497864),
 ('vaccinia-virus-ankara', 0.5020899772644043),
 ('animal-markets', 0.5007162094116211),
 ('vaccinated', 0.498425155878067),
 ('commercial', 0.4977818429470062),
 ('live-virus', 0.49401360750198364),
 ('animal-market', 0.48633021116256714),
 ('carrying', 0.48520395159721375),
 ('legal-trade', 0.47734832763671875),
 ('live-attenuated', 0.4766719341278076)]
```

```
1 word_vectors.most_similar("hand-sanitizer", topn=20)
```

```
[('wearing-face-masks', 0.6330808401107788),
 ('eye-protection', 0.6090339422225952),
 ('hand-washing', 0.6032581925392151),
 ('hand-hygiene', 0.5986652374267578),
 ('alcohol-based-hand-rubs', 0.5866068601608276),
 ('alcohol-based-hand-sanitizers', 0.5841511487960815),
 ('hygiene', 0.5674364566802979),
 ('personal-hygiene', 0.5653126239776611),
 ('wearing-masks', 0.5646619200706482),
 ('hand-sanitizers', 0.5642377138137817),
 ('sanitization', 0.5633803606033325),
 ('face-masks', 0.5617189407348633),
 ('sanitizers', 0.5610991716384888),
 ('antiviral-use', 0.5600476264953613),
 ('handwashing', 0.5550966858863831),
 ('mask-wearing', 0.5506600141525269),
 ('dental', 0.5448571443557739),
 ('soap', 0.5419381856918335),
 ('washing-hands', 0.5397700667381287),
 ('goggles', 0.5392211079597473)]
```

```
word_vectors.most_similar("increased", topn=20)
```

```
[('decreased', 0.8684967756271362),
 ('increase', 0.8016087412834167),
 ('reduced', 0.8004354238510132),
 ('higher', 0.7620439529418945),
 ('greater', 0.7592437267303467),
 ('elevated', 0.7063256502151489),
 ('lowered', 0.705357015132904),
 ('decrease', 0.6944239139556885),
 ('diminished', 0.688621461391449),
 ('increasing', 0.6824566125869751),
 ('lower', 0.6672701835632324),
 ('declined', 0.6301380395889282),
 ('decreasing', 0.6211043000221252),
 ('enhanced', 0.5869908332824707),
 ('improved', 0.5640379190444946),
 ('decline', 0.5627540349960327),
 ('high', 0.5571213960647583),
 ('impaired', 0.5553809404373169),
 ('dropped', 0.5540823936462402),
 ('declining', 0.5518118739128113)]
```

```
word_vectors.most_similar("risk", topn=20)
```

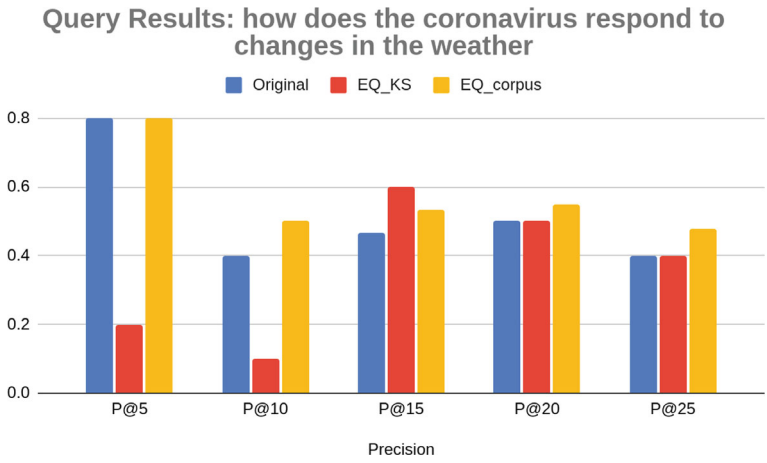
```
[('likelihood', 0.7570781707763672),
 ('high-risk', 0.6449357271194458),
 ('hazard', 0.6242806911468506),
 ('higher-risk', 0.6026164293289185),
 ('burden', 0.6003137230873108),
 ('incidence', 0.5721932649612427),
 ('odds', 0.564045786857605),
 ('chance', 0.5323513150215149),
 ('concern', 0.5302383899688721),
 ('risk-factor', 0.51320481300354),
 ('awareness', 0.511039137840271),
 ('occupational-risk', 0.5017816424369812),
 ('danger', 0.4995952248573303),
 ('elevated-risk', 0.4969768822193146),
 ('mortality-risk', 0.4965442419052124),
 ('vulnerability', 0.49485042691230774),
 ('acquiring', 0.48827314376831055),
 ('rate', 0.48712992668151855),
 ('health-risks', 0.48255687952041626),
 ('older-individuals', 0.4825335741043091)]
```

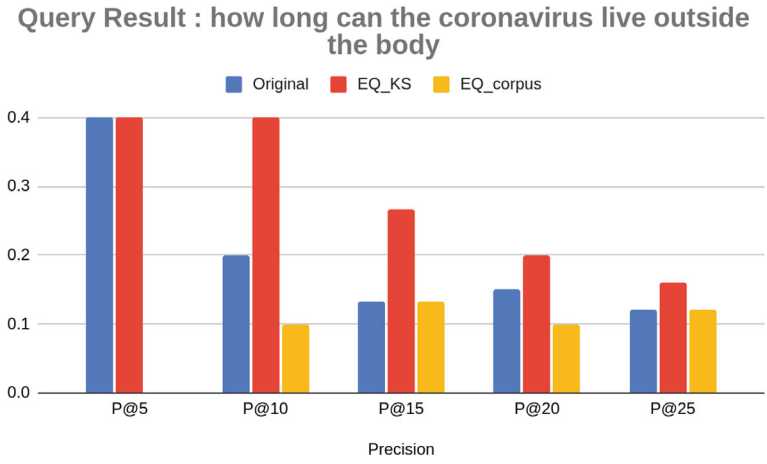
```
word_vectors.most_similar("angiotensin-converting-enzyme", topn=20)
```

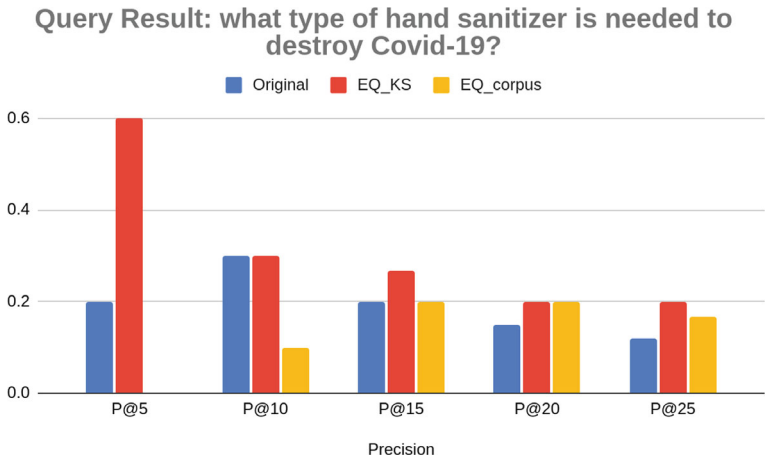
```
[('ace', 0.7953211665153503),
 ('angiotensin', 0.7547454833984375),
 ('angiotensin-1–7', 0.7505214214324951),
 ('angiotensin-converting-enzyme-2', 0.7184836864471436),
 ('angiotensin-ii', 0.7065269947052002),
 ('renin-angiotensin-system', 0.6996192932128906),
 ('angiotensin-converting-enzyme', 0.6905388832092285),
 ('angiotensin-converting', 0.6875545978546143),
 ('ace-2', 0.6866792440414429),
 ('angiotensin-i', 0.6852770447731018),
 ('ang', 0.6782753467559814),
 ('receptor-axis', 0.6782747507095337),
 ('ang-ii', 0.6733224391937256),
 ('angiotensin-converting-enzyme-2', 0.6729364395141602),
 ('angiotensin-converting-enzyme-', 0.6686009168624878),
 ('angiotensin-ii-receptor-blockers', 0.6676031351089478),
 ('angiotensin-ii-type-1-receptor', 0.6593121290206909),
 ('angiotensin-1-7', 0.6510974764823914),
 ('angii', 0.6494097113609314),
 ('angiotensin-', 0.6470130085945129)]
```

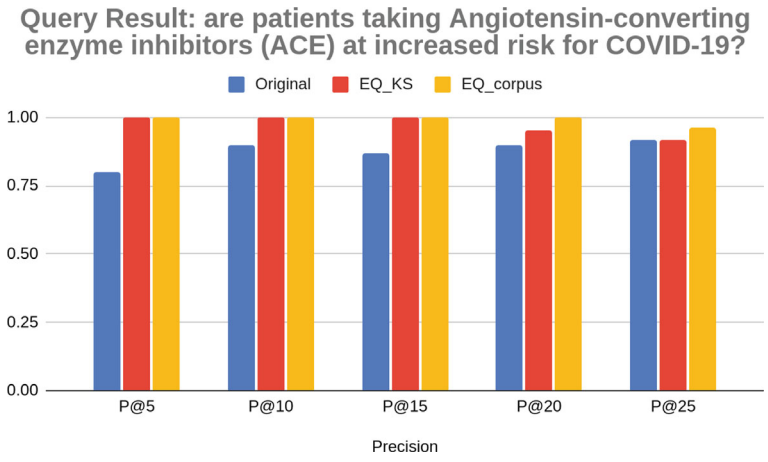
Annexure 2

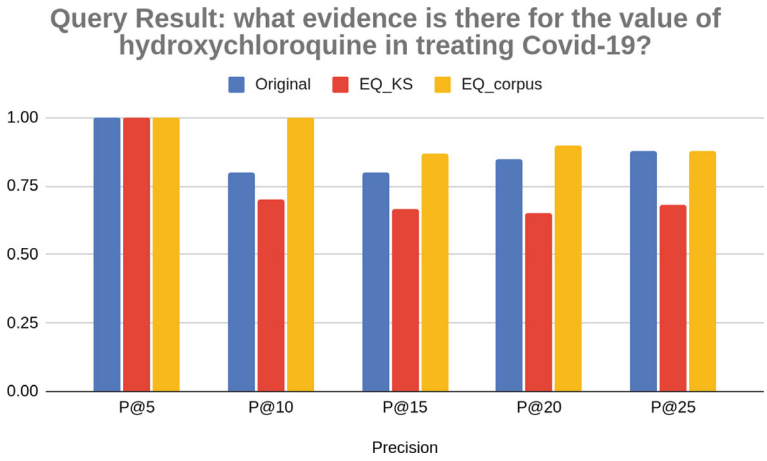
Displays the comparison of Precision@5, 10, 15, 20, 25 for original, expanded queries using knowledge sources and corpus strategies.

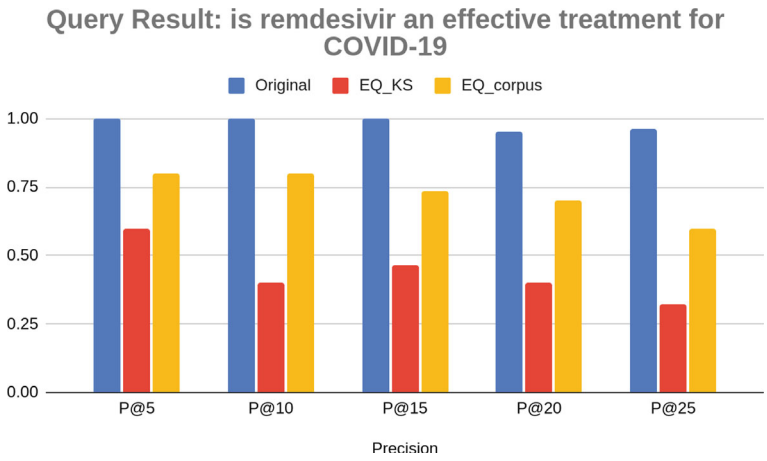

Query Results: how does the coronavirus respond to changes in the weather
Original
EQ_KS
EQ_corpus
0.8
0.6
0.4
0.2
0.0
P@5
P@10
P@15
P@20
P@25
Precision


Query Result : how long can the coronavirus live outside the body
Original
EQ_KS
EQ_corpus
0.4
0.3
0.2
0.1
0.0
P@5
P@10
P@15
P@20
P@25
Precision


Query Result: what type of hand sanitizer is needed to destroy Covid-19?
Original
EQ_KS
EQ_corpus
0.6
0.4
0.2
0.0
P@5
P@10
P@15
P@20
P@25
Precision


Query Result: are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19?
Original
EQ_KS
EQ_corpus
1.00
0.75
0.50
0.25
0.00
P@5
P@10
P@15
P@20
P@25
Precision





References

Aronson AR, Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. J Am Med Inf Assoc 17(3):229–236. https://doi.org/10.1136/jamia.2009.002733

Baeza-Yates RA, Ribeiro-Neto BA (1999) Modern information retrieval. ACM Press/Addison-Wesley. ISBN 0-201-39829-X

Chakraborty S, Bisong E, Bhatt S, Wagner T, Elliott R, Francesco (2020). BioMedBERT: A pre-trained biomedical language model for QA and IR. In: International conference on computational linguistics, pp 669–679. https://doi.org/10.18653/v1/2020.coling-main.59

Croft WB, Harper DJ (1979) Using probabilistic models of document retrieval without relevant information. J Documentation 35(4):285–295. https://doi.org/10.1108/eb026683

Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Egozi O, Markovitch S, Gabrilovich E (2011) Concept-based information retrieval using explicit semantic analysis. ACM Trans Inf Syst 29(2):1–34. https://doi.org/10.1145/1961209.1961211
Gong Z, Muyeba M, Guo J (2010) Business information query expansion through semantic network. Enterp Inf Syst 4(1):1–22
Guo X, Liu R, Shriver CD, Hu H, Liebman MN (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics (oxford, England) 22(8):967–973. https://doi.org/10.1093/bioinformatics/btl042
Hao M, Fan K (2017) A method for calculating the similarity of TF-IDF texts for synonyms in biomedical domains. In: Advances in engineering research (AER), p 130
Harman D (1992) Relevance feedback and other query modification techniques. In: Information retrieval: data structures and algorithms, pp 241–263. https://doi.org/10.5555/129687.129698
Hersh W, Buckley C, Leone TJ, Hickam D (1994) OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, pp 192–201. https://doi.org/10.5555/188490.188557
Hersh W, Price S, Donohoe L (2000) Assessing thesaurus-based query expansion using the UMLS Metathesaurus. In: Proceedings/AMIA ... Annual Symposium, pp 344–348
Johnson AEW, Pollard TJ, Berkowitz SJ et al (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6:317
TREC-COVID Home (n.d.) IR. Retrieved 20 Oct 2022, from https://ir.nist.gov/trec-covid/
TensorFlow Hub (n.d.) TensorFlow Hub. Retrieved 8 June 2023, from https://tfhub.dev/tensorflow/cord-19/swivel-128d/3
Imran H, Sharan A (2009) Thesaurus and query expansion. Int J Comput Sci Inf Technol (IJCSIT) 1(2):89–97
Lee W-N, Nigam S, Karanjot, Musen M (2008) Comparison of ontology-based semantic-similarity measures. In: AMIA ... Annual symposium proceedings/AMIA symposium. AMIA Symposium, pp 384–388
Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arxiv https://doi.org/10.48550/arXiv.1901.08746
Li Y, Wu H (2012) A clustering method based on K-means algorithm. Phys Procedia 25:1104–1109. https://doi.org/10.1016/j.phpro.2012.03.206
Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html
McInnes B, Pedersen T, Pakhomov S (2009) UMLS-interface and UMLS-similarity: open-source software for measuring paths and semantic similarity. In: AMIA Annual Symposium Proceedings, pp 431–435
Mclnnes L, Healy J, Melville J (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arxiv https://doi.org/10.48550/arXiv.1802.03426
Mooers C (1951) Making information retrieval pay. Zator Company, Michigan.
Mubaid HA, Nguyen HA (2006) A cluster-based approach for semantic similarity in the biomedical domain. In: Conference proceedings—IEEE engineering in medicine and biology society, pp 2713–2717. https://doi.org/10.1109/IEMBS.2006.259235
Pan M, Zhang Y, He T, Jiang X (2018) An enhanced HAL-based pseudo relevance feedback model in clinical decision support retrieval. In: International conference on intelligent computing. Springer, pp 93–99. https://doi.org/10.1007/978-3-319-95933-7_12
Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG (2007) Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 40:288–299
Pedersen T, McInnes BT, Liu Y, Melton GB, Pakhomov SV (2009) UMLS: similarity: measuring the relatedness and similarity of biomedical concepts. In: AMIA ... Annual symposium proceedings. AMIA symposium, pp 431–435
Piwowar H, Priem J, Larivière V (2018) The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ

Ramampiaro H, Li C (2011) Supporting biomedical information retrieval: the biotracer approach. In: Transactions on large-scale data- and knowledge-centered systems IV: special issue on database systems for biomedical applications. Springer, Berlin, Heidelberg, pp 73–94. https://doi.org/10.1007/978-3-642-23740-9_4

Rawal S (2020) Multi-perspective semantic information retrieval in the biomedical domain. arxiv, v1. https://doi.org/10.48550/arXiv.2008.01526

van Rijsbergen C (1979) Information retrieval: theory and practice. In proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems, vol 79, pp 1–14

Rocchio JJ (1971) Relevance feedback in information retrieval

Sakai T, Robertson SE, Walker S (2001) Flexible pseudo-relevance feedback via direct mapping and categorization of search requests. BCS-IRSG ECIR 2001 Proceedings, pp 3–14

Salton G, Buckley C (1990) Improving retrieval performance by relevance feedback. J Am Soc Inf Sci 41(4):288–297. https://doi.org/10.1002/(SICI)1097-4571(199006)41:4%3c288::AID-ASI8%3e3.0.CO;2-H

Salton G, McGill MJ (1983) Introduction to modern information retrieval. McGraw-Hill Book Co., New York

Sankhavara J (2018) Biomedical document retrieval for clinical decision support system. In: Proceedings of ACL 2018. In the student research workshop. https://aclanthology.org/P18-3012/

Savino, P., & Sebastiani, F. (1998). Essential bibliography on multimedia information retrieval, categorization and filtering. In: Slides of the 2nd European digital libraries conference tutorial on multimedia information

Text REtrieval Conference (TREC) Precision Medicine Track. (2018). Text Retrieval Conference. Retrieved 20 Oct 2022, from https://trec.nist.gov/data/precmed.html

Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39(suppl_2):W541-W545.

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometrics Intell Lab Syst 2(1–3):37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Xu J, Croft WB (2017) Query expansion using local and global document analysis. ACM Sigir Forum 51:168–175. https://doi.org/10.1145/243199.243202

Zhoun D, Lawless S, Liu J, Wade V (2012) Improving search via personalized query expansion using social media. Inf Retrieval 15(3–4):218–242. https://doi.org/10.1007/s10791-012-9191-2