International Journal of Knowledge Processing Studies
(KPS)
Homepage: http://kps.artahub.ir/

ORIGINAL RESEARCH ARTICLE

The Assessment of the Effect of Query Expansion on Improving the Performance of Scientific Texts Retrieval in Persian

Ahmadreza Varnaseri[1], Maryam Nakhoda[2,*], Sareh Karimi[3]
[1] *PhD student in Information Science and Knowledge, Faculty of Management, University of Tehran, Tehran, Iran.*
[2] *Assistant Professor, Department in Information Science and Knowledge, Faculty of Management, University of Tehran, Tehran, Iran.*
[3] *Graduate of Industrial Engineering, Saveh Azad University, Saveh, Iran.*

ARTICLE INFO

Article History:
Received: 2021/09/24
Accepted: 2021/09/28
Published Online: 2021/12/20

Keywords:
Query expansion
Persian language
Elmnet search engine
scientific texts

Number of Reference: 19
*Number of Figures:*1
Number of Tables: 3

DOI:
http://dx.doi.org/10.22034/kps.2021.141924

ABSTRACT

Purpose: This study aims to determine the effect of query expansion on scientific texts retrieval in Persian. Method: The present study was conducted using a quasi-experimental method. The results are obtained by analyzing 40 initial and expanded queries of postgraduate students in the Faculty of Management, University of Tehran. Query expansion was performed manually using primary research results. Findings: Query expansion of Persian scientific texts leads to an increase in the number of related retrieved documents, as well as the comprehensiveness and accuracy of retrieving scientific data in Elmnet search engine, which as a result, improves the overall performance of information retrieval. Results: Nowadays, automatic query expansion is on the agenda of databases. Given that Persian databases are not fully developed, and the existence of specific problems of writing in the Persian language, information literacy training and the method of defining and expressing information requirements and providing them to the information retrieval systems, can have a significant impact on postgraduate students and researchers, to retrieve the required information and save them time and money.

►**Citation (APA):** Varnaseri, A., Nakhoda, M., Karimi, S. (2021). The Assessment of the Effect of Query Expansion on Improving the Performance of Scientific Texts Retrieval in Persian. *International Journal of Knowledge Processing Studies, 1*(1): 39-51.

***Corresponding Author:**
Email: mnakhoda@ut.ac.ir
ORCID ID: 0000-0002-6558-7880

1. Introduction

The main objective of an information retrieval (IR) system is to retrieve documents that are relevant to a user's intentions from a large information space. Such systems calculate the similarity between a search query and documents and retrieve a list of documents that are arranged in descending order of similarity. The retrieved list of documents is sometimes large and contains many irrelevant documents, especially when searching the Web. The main issue that is encountered in the retrieval of documents that are not related to user needs is the vocabulary mismatch problem: the terms that the author has used to describe a concept in the document differ among users. This critical issue of vocabulary mismatch is further aggravated by short queries, which are becoming increasingly common in web search. Most of the prevalent web search queries contain no more than two or three words. Thus, the likelihood of encountering the severe issues of synonymy and polysemy is very low. Addressing the problem of vocabulary mismatch is essential for such short queries and important for effective information retrieval. (Reza et al., 2021).

To increase the Information Retrieval System's efficiency, there is a requirement to expand the native user query. There are many approaches to enhance the user query in which the primary method consists of semantic-based query expansion (QE). In a semantic-based QE approach, relevant documents are retrieved by considering all the similar terms of a given user query. The semantic-based QE approach helps in dealing with the limitation of low recall and low precision value of the Information retrieval system and deals with ambiguity and vagueness. Query Expansion techniques contain the semantic concepts that are relevant to semantic computing, computational intelligence, and information retrieval area. Computational intelligence technique is required during automatic query expansion for advanced information processing. (Sharma, Pamula and Chauhan, 2021). Due to the complexity of professional search tasks and their reliance on specialized terminology, query extension is a natural approach to help the searcher offer query extension is the process of transforming or increasing the user query to increase its effectiveness.

Knowledge workers (such as healthcare information professionals, patent agents, and recruitment professionals) undertake work tasks where search forms a core part of their duties. In these instances, the search task is often complex and time-consuming and requires specialist expert knowledge to formulate accurate search strategies. Interactive features such as query expansion can play a key role in supporting these tasks (Russell-Rose, Gooch and Kruschwitz, 2021).

Modeling the information-seeking patterns of academic researchers (Ellis, 1989), is different from modeling the behavioral patterns of other individuals who search for other general or specific purposes. To the extent that this diversity is different even in different scientific fields. These differences affect the research of academics and researchers (Amolochitis, 2013). Academic search tasks are inherently complex, uncertain, and multifaceted (Du and Evans, 2011).

Within the domain of academic search, it is unclear to what extent searchers employ the strategies specified in such models when faced with different types of information needs. They are also faced with a range of simple search tasks such as fact verification as well as complex tasks such as knowledge discovery (Hoeber; Pathel and Storie, 2019).

Academic queries and their constituent vocabulary can be offered to databases in order to search for general or highly specialized information, which is also a difficult task to judge the relevance of these documents because they are mostly at the knowledge discovery level. Although, the number of words in academic and scientific queries is relatively higher compared to general searches such as news texts (Pueyo and Redrado, 2003). However, the search results are still concise and do not meet the needs and expectations of users.

The main challenge that search engines currently face is that users' queries are usually short typically only two to three words. Short queries are usually the result of the difficulty users have in presenting their information needs and these queries inherently do not correctly represent users' information requirements. According to Belkin's well-known Anomalous State of Knowledge hypothesis, such difficulties are usually the result of gaps in users' knowledge. (Farhan, 2021).

Due to the complexity of professional search tasks and their reliance on specialized terminology, query extension is a natural approach to help the searcher offer query extension is the process of transforming or increasing the user query to increase its effectiveness.

This is due to the complexity of scientific articles and texts, which are mostly looking to solve problems, and also design and test scientific hypotheses, or they could be the scientific evidence of the part of the general and objective knowledge of specific science and could be presented following on a long-lasting scientific tradition design.

However, scientific queries submitted to databases are often short and ambiguous. The primary problem with queries of the users on the web is the low number of query terms. Research shows that users tend to submit short queries because they lack knowledge of the subject to identify their information needs, and the length of a web query is between 2 and 3 terms. The small number of terms in the query leads to the absence of some important terms that describe the information needs in the query (Spink et al., 2011), which as a result, does not provide adequate meaning for the desired query.

Another problem regarding query is its ambiguity. Query expansion is one of the most common approaches to solving the problem of short and obscure queries that has been used since the 1960s. Query expansion is an accepted and widely used approach which improves users' short queries by adding additional terms from the context, and in addition, it solves the problem of ambiguity in natural language by limiting the meaning of words by terms added to the query (Zhang, 2013).

Some people consider the query expansion to be a re-modification of the user query by adding additional terms and re-weighting the query terms by the system (Lavrenko and Croft, 2017). Some also focus only on re-weighting the terms of the query (Bendersky and Croft, 2008). Moreover, other people consider three approaches: adding additional terms, re-weighting; and a combination of adding additional terms and re-weighting them (Yates and Neto, 1999). Various methods, techniques, and algorithms have been used for query expansion of users in information retrieval systems. There are three methods for query expansion which include: manual, automatic, and interactive (Amin, 2008).

In the manual query expansion, the users determine the terms of expansion based on their experience and knowledge of the subject and the set of documents (Bhogal, McFarland and Smith, 2007). Manual query expansion is suitable for subject matter specialists and professionals. In the interactive query expansion process, the system identifies and presents a set of potential query expansion terms to the user, the user then determines which term or terms are suitable for the expansion.

The premise of the improvement of the interactive query is that people are more capable of judging whether the terms of useful and relevant than machines and this method shows the effectiveness in action and practice (Azad and Deepak, 2017). In automatic query expansion, the system automatically selects the expansion terms without any user intervention and thus, reformulates the query for the user. Automatic query expansion takes place in three ways: based on relevance, based on knowledge structures, and based on the web information.

Scholars routinely search-relevant papers to discover and put a new idea into proper context. Despite ongoing advances in

scholarly retrieval technologies, locating relevant papers through keyword queries is still quite challenging due to the massive expansion in the size of the research paper repository. To tackle this problem, we propose a novel real-time feedback query expansion technique, which is a two-stage interactive scholarly search process. Upon receiving the initial search query, the retrieval system provides a ranked list of results. In the second stage, a user selects a few relevant papers, from which useful terms are extracted for query expansion. The newly expanded query is run against the index in real-time to generate the final list of research papers. In both stages, citation analysis is involved in further improving the quality of the results. The novelty of the approach lies in the combined exploitation of query expansion and citation analysis that may bring the most relevant papers to the top of the search results list. The experimental results on the Association of Computational Linguistics (ACL) Anthology Network data set demonstrate that this technique is effective and robust for locating relevant papers regarding normalized discounted cumulative gain (nDCG) precision and recall rates than several states of the art approaches (Khalid, Wu, Alam and Ullah, 2021).

The Persian language is one of the most important spoken languages in the Middle East and Central Asia, and research in the field of information retrieval is expected to be developed in Persian. Query expansion is one of the fields that helps with data retrieval tasks. Query expansion methods and algorithms are implemented and used in different languages. Persian, which is a branch of the Indo-European language, is considered one of the most important languages in Asia. In Iran and Tajikistan, Persian is the official language, and it is one of the two official languages in Afghanistan, along with Pashto. Also, Persian was the official language of India before the arrival of British colonialism. Given the weaknesses in natural language in terms of semantics, synonyms, and spelling, more attention should be paid to the form of terms and the choice of appropriate words. Nowadays, a large volume of scientific content and web pages are produced in Persian. To solve these problems, the query expansion technique is applied, which will improve the retrieval accuracy through suggesting and adding appropriate words to the user query.

The major drawback is that information which is retrieved without understanding the meaning of the user's query, information is not much relevant and context is also missing in the retrieved documents. As there exist two types of ambiguity among the words synonym and polysemy while dealing with the semantic meaning of the words/concepts. Meaning/Context of words, relationship with other words is missing on current web search. The semantic knowledge of the keywords enhances the accuracy of the information retrieval process. Hence, to deal with the context of words, to find the semantic meaning of words, and to find their relationship with other words is the main aim of a research paper.

According to the cases mentioned in this article, we will expand the scientific queries by postgraduate students based on the results of previous searches, and furthermore, compare the search results of primary and expanded queries. Based on this comparison, we can determine the extent to which query expansion helps improve data retrieval results.

The importance and application of query expansion

Information retrieval is one of the main needs of users. Many users search the web and other sources daily to meet their information needs. Issues in natural language such as lexical inconsistency, polysemous words, short and vague query and incomplete knowledge of users about the subject matter in data retrieval lead to retrieval of irrelevant results which reduces user satisfaction with the retrieved results. Query expansion, by reviewing user queries and automatically adding relevant and valuable words, helps the users to search and retrieve documents related to the user's needs and purposes. If query expansion is not performed intelligently, it leads to deviation of the query which distances it from the

user's intent, which results in even more irrelevant results retrieval than the initial query results. Queries are created to express users' needs for information available on web pages, databases, and other resources.

2. Objective

This study aims to determine the effect of query expansion on retrieving scientific texts in Persian.

Questions

1. How much does the initial search affect the relevance of students' scientific information retrieval results?
2. How much does the query expansion affect the improvement of the performance of students' scientific information retrieval results?

The success of the query depends on two important factors:

1. Vocabulary selection method: manual, automatic, interactive
2. Vocabulary sources: Sources that provide expansion vocabulary: According to the retrieval results, based on the knowledge structures (such as ontology, thesaurus) given to the structure and features of the Persian language, the issue of ambiguity of the query words is more tangible which in turn, reduces the accuracy of retrieving documents.

Query expansion based on relevance: This type of query expansion takes place in two types of relevance feedback and pseudo-relevance feedback. Relevance feedback considers the search process as an interactive operation and selects documents that users identify as relevant to expand queries. In the pseudo-relevance feedback approach, documents with higher ranks in the list of primary query results, are considered as related documents. When relevance feedback of the user is missing, the approach also known as blind feedback will be used.

Numerous information retrieval systems use this approach for query expansion. After identifying sources based on various algorithms such as Rocchio, Probabilistic, and Local Context Analysis, automatic correction of expansion terms takes place.

Query expansion based on knowledge structures: In this method, expansion terms are extracted from knowledge structures using two approaches of corpus-based, and hand-built. The corpus-based approach provides terms from a variety of collections, such as clustering of terms and automatic corrections extracted from the text (Atwan and Mohd, 2017).

The hand-built approach uses external sources such as dictionaries, general thesauri, thesauruses for specific areas and ontologies (Efthimiadis, 1996).

Web-based query expansion: This approach uses web documents, online knowledge databases (such as Wikipedia), or query logs as a source of expansion terms. The motivation for using web information to expand queries is to enrich the collection with external information because external information is dynamic and reflects public opinion. Wikipedia is the largest multilingual web encyclopedia that, as a structured source of information, is the source of many natural language retrievals and processing tools such as DBpedia, Yago, and WEX ontologies (Mehdi et al., 2017). Moreover, query logs generated by web search engines record user interactions, including their experiences in re-formulating queries and their achievement of the desired results in the form of inquiry sessions.

Ontology-based query expansion: Ms. Farhoudi and her colleagues, using documents extracted from Wikipedia, created an automated ontology of the Persian language and used it to expand the query. In this method, the user's initial query is processed and the concepts in it are identified. Afterward, using the ontology graph and considering the relationships between ontology concepts, the concepts related to the initial query concepts are determined, and are considered for expansion after weighting. In this approach, the vector method is used to expand the query. The query vector is formed based on the concepts contained in the user query. The

query vector is an array whose length is equal to the number of concepts in the ontology. Each element of the array belongs to a specific concept of ontology. If the concept (s) in the query exists in the ontology, the corresponding element in the presentation will have a value of one, otherwise, it will be zero. Then, using ontology and the relationships between concepts, the elements of the array are weighted. At the end of the weighting process, the expanded query may contain many concepts with different weights, therefore, to achieve more optimal results, a threshold value is used, to remove concepts with a lower weight than the threshold. Finally, the original query is expanded by using the remaining concepts.

Instead of using the relationship of synonymous vocabulary classification, Navigli studied the possibility of using ontological information to extract the semantic domain of a word and expanded the query using the words in the "word definition". To clear up the ambiguity of the meaning of the word, Navigli did the following: each word of the user query obtains its synonym set from WordNet ontology, and next, he formed a semantic network based on the pair of "query word" and "a word synonymous with query word". This semantic network is formed based on the relationships between words in an ontology.

Automatic query expansion: in an automatic query expansion retrieval system, is responsible for determining and selecting the appropriate vocabulary for expansion without the user's participation. This system is designed in such a way to select words from the desired sources, afterwards, it ranks and weighted those words based on one of the ranking algorithms, and at last, it selects the most appropriate words for expansion.

Rocchio in the SMART system performs query expansion operations using normalized vectors. Queries and documents are represented by weighted vectors. Query vectors and documents are compared for retrieval operations. The retrieval system performs a relevance feedback mechanism using information about documents retrieved after the user query search. And the initial query changes in a repeat loop as follows: Using a threshold, vocabulary vectors of all related retrieved documents are added to the query vocabulary vector. In addition, again using a threshold, vocabulary vectors of all unrelated retrieved documents, are subtracted from the query vocabulary vector. In this way, alarge number of additional words are obtained, some of which have a negative weight. The weight of some query words also increases or decreases.

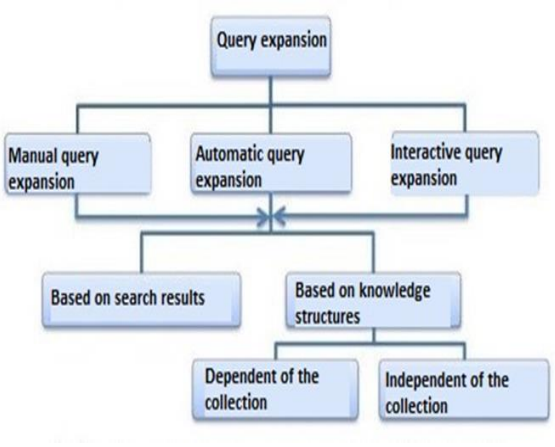

Figure1. Query expansion (methods and sources)

3. Research background

To collect the literature review and complete the content, the researchers have obtained the information needed in this paper by searching in some databases such as: Google Scholar, Science Direct, Pro Quest.

Authors	Title	Date	Results
Kulkarni and Kale	Information Retrieval based Improvising Search using Automatic Query Expansion	2021	Current search engines like Google, Yahoo give the search results, indeed users are facing problems in information retrieval. The main problem is because of word mismatch and availability of many resources. Petabytes of information is available because of Internet. From that huge available data, for a naïve user it becomes hectic to distinguish between relevant and irrelevant information to individual interest. Also, another reason of getting irrelevant information is incompatibility between terms that users are using and keywords present in documents. Query expansion is an adding keyword to the original query. The main issue in query expansion is selection of appropriate terms from user's original query. Vocabulary database helps to solve this issue. Identification of the similar words and language entities that are similar in meaning is done by vocabulary which is frequently incorporated in information retrieval system. Thesaurus has been used across a large area of in information retrieval also many applications and natural language processing. In this work, to improve performance of a search query, BM25 model is used for query expansion. Cosine similarity is used to determine similarity between two keywords. Rocchio algorithm is used to calculate the relevance feedback. Experimental result shows better results using Rocchio algorithm
Raza, and et.al.	User Interest Driven Semantic Query Expansion for Effective Web Search	2021	The mismatch between search query terms and documents affects the retrieval results of the existing IR system. Many semantic QE methods have been proposed to solve the term mismatch issue. These methods take advantage of the ontology knowledge source to expand search queries with terms semantically relevant to the original search query. However, semantic QE does not consider individual user interests, which can be extracted from the Web browser history. The use of ontology can help in obtaining domain semantics, whereas user preferences can be collected from the browsing history. Based on these ideas, we incorporate user interests into the process of semantic QE. Compared with existing semantic QE techniques, our system identifies the user query domain via ontology semantics (by exploiting ontology relationships) and captures user intents from history logs (via correlations). The retrieval results over Google for queries expanded by our system achieves better precision than the initial query results. The proposed technique achieves 81.2% and 86.4% average precision for top 50 and 100 Google results, respectively. Therefore, combining semantics and user interests can achieve substantial improvement in precision results. In this research, we focus on domain-specific ontologies for the identification of an initial query domain and utilize history logs for the generation of an expansion term set. In the future, we want to explore the effect of a large ontology (domain independent), as a domain-specific ontology contains limited terms. We further plan to exploit browsing history features, such as query session and document click time, in the process of expansion term extraction.
Jia, And Lin,	Using Query Expansion in Manifold Ranking for Query-Oriented Multi-Document Summarization	2021	It not only makes use of the relationships among the sentences, but also the relationships between the given query and the sentences. However, the information of original query is often insufficient. So we present a query expansion method, which is combined in the manifold ranking to resolve this problem. Our method not only utilizes the information of the query term itself and the knowledge base WordNet to expand it by synonyms, but also uses the information of the document set itself to expand the query in various ways (mean expansion, variance expansion and TextRank expansion). Compared with the previous query expansion methods, our method combines multiple query expansion methods to better represent query information, and at the same time, it makes a useful attempt on manifold ranking. In addition, we use the degree of word overlap and the proximity between words to calculate the similarity between sentences. We performed experiments on the datasets of DUC 2006 and DUC2007, and the evaluation results show that the proposed query expansion method can significantly improve the system performance and make our system comparable to the state-of-the-art systems
Jain And Seeja And Jindal	A fuzzy ontology framework in information retrieval using semantic query expansion	2021	To overcome the weaknesses of current web system and to utilize the strengths query expansion a novel framework based on fuzzy ontology is proposed for information retrieval. In the proposed framework, domain specific knowledge is utilized for ontology construction. In framework pre-defined domain ontologies and Global ontology, ConceptNet is used to construct a fuzzy ontology. Based on constructed fuzzy ontology most semantically related words for a query are identified and query is expanded. A fuzzy membership function is defined for different semantic relationships present among the Global ontology ConceptNet. Based on the proposed framework queries are expanded (Semantic query expansion) and evaluated on four popular search engines namely Google, Yahoo, Bing and Exalead. The performance metrics used are Precision, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), R-precision and Number of documents retrieved. The Web search engines are precision oriented. Based on the proposed framework all the metrics are improved approx. by 10%. Precision before the query expansion lies between 0.75-0.81 whereas after the query expansion lies between 0.85-0.89 on various search engines. The number of documents retrieved is almost improved 1/1000 after the query expansion

Naseri and et.al	Contextualized Embeddings for Query Expansion	2021	In this work we leverage recent advances in context sensitive language models to improve the task of query expansion. Contextualized word representation models, such as ELMo and BERT are rapidly replacing static embedding models. We propose a new model Contextualized Embeddings for Query Expansion (CEQE), that utilizes query-focused contextualized embedding vectors
Farhan, and et.al	Word-embedding-based query expansion: Incorporating Deep Averaging Networks in Arabic document retrieval	2021	This work has explored and modified WE expansion strategies with a sentence-embedding technique called DANs. It considers the average vector of the query term vectors in generating the candidate expansion terms based on the WE technique. The WE technique used in this study is Word2Vec. The probabilistic model Okapi BM25 and the EQE1 and V2Q approaches were used for comparisons. original query term vectors. Thus, if the DANs average vector is unable to retrieve useful candidate vectors, the two enhanced techniques will not be strongly affected because they also use their original query term vectors to find the candidate vectors
Esposito, and et.al	Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering	2020	The query expansion is created in a set of synonyms of the relevant terms in the user query,first extracted from a lexical source and then formed according to the documents used as the source of information in the QA system. A hybrid QE approach based on lexical sources and word embedding is proposed in this paper, which is mainly designed to work in QA systems based on IR information retrieval, in which a precise answer to a question is formulated by the user from the relevant documentation. Is extracted to a closed field. This approach is done with the aim of extracting the terms in the natural language questions and enriching them with words related to additional meaning to retrieve the relevant sentences.
Silva and Mendoza	). Improving query expansion strategies with word embeddings. In *Proceedings of the ACM Symposium on Document Engineering*	2020	In the study of improving query development strategies by embedding words, the usefulness of embedding the word to display queries and documents in query-document matching is pointed out and a re-ranking strategy is used. The re-ranking step is done by displaying questions and documents based on word embedding. And by restarting, they introduce IDF words as a new text display strategy based on word embedding, which allows us to create a query vector that has more relevance to informative terms throughout the process. Experimental results in the TREC benchmark data set show that our proposal consistently achieves the best results in terms of MAP.

4. Methodology

The research method is quasi-experimental and based on intervention in query expansion. In the present study, a total of 40 queries were conducted by postgraduate students of the Faculty of Management, University of Tehran, in the Persian scientific search engine of Elmnet. To conduct the research, the students who were willing to cooperate were asked to do a query or queries based on the subject matter and previous academic work and search in the Elmnet search engine. Then, based on the information in the title, abstract, and keywords, judge the relevance of the 10 retrieved documents. The relevance judgment was recorded at three levels (related, somewhat related, and unrelated). Furthermore, students were asked to add words to the original query based on these 10 evaluated results to expand the query to achieve better results. At this stage, the query expansion takes place manually. The search was repeated with an expanded query and the three-level relevance judgment was expanded queries as well as judgments of search results were recorded.

Suggested method: In terms of choosing the method of query expansion, the manual word expansion method is recommended.

After making a judgment about the relevance of the documents, data retrieval performance is measured by the evaluation metrics of data retrieval systems. The evaluation metrics used in the present study include related retrieved documents, retrieval, and accuracy.

Related retrieved documents: related documents recovered consist of related and somewhat related documents.

Accuracy (Prohibition): It refers to the ability of the indexing system to keep unrelated documents away from the user and it calculates the system's ability to select completely relevant options. Prohibition coefficient is the number of related documents retrieved to the total retrieved documents. In fact, the prohibition includes the number of related documents retrieved to the total number of retrieved documents in the searched documents. (Hariri et al., 2014)

Retrieval (comprehensiveness): Indicates the ratio of retrieved documents to all relevant documents in the database. However, in this study, it was not possible to access all related documents of the database. Thus, relative retrieval was applied, which is the ratio of the retrieved documents for one query to the total related documents of the two queries, deducted from the overlap of the retrieved documents (Hariri et al., 2014).

5. Findings

Table 1 presents the comprehensiveness and accuracy of the initial and expanded queries, along with the number of related retrieved documents.

Table 1.
Information retrieval metrics for initial and expanded queries

Row	Initial Queries				Expanded Queries				Overlap
	Query	Related documents	Retrieval	Accuracy	Query	Related documents	Retrieval	Accuracy	
1	Digital library development	7	0.75	1	Human factors in the development of the digital library	10	0.83	0.9	9
2	Children's libraries	0	0.44	0.9	The impact of children's libraries on their creativity	9	0.56	0.7	7
3	Fuzzy QFD	2	0.46	0.9	Fuzzy measurable house of quality matrix	9	0.69	0.6	6
4	Information needs of organizations	3	0.40	0.9	Investigating the information needs of managers of organizations	9	0.90	0.4	4
5	Information exchange protocol	2	0.38	1	Memorandum of Understanding of information exchange in digital libraries	10	0.77	0.5	5
6	Economic evaluation of the library	1	0.25	0.7	Economic evaluation of cost-benefit analysis library	7	0.88	0.2	2
7	Query expansion	1	0.33	0.7	Query expansion and search expansion	7	0.78	0.3	3
8	Electric power	2	0.57	0.8	Power electricity market indicators	8	0.57	0.8	8
9	Islamic securities	2	0.75	0.4	Securities Market Law of the Islamic Republic of Iran	4	0.50	0.6	6
10	Auditing	1	0.50	0.8	The importance of planning in auditing	8	0.57	0.7	7
11	Retrieval systems	0	0.41	1	Image retrieval systems	10	0.59	0.7	7
12	Organic Chemistry	1	0.43	0.9	Applied Organic Chemistry - Solvent	9	0.64	0.6	6
13	Genetics and indicator weights	2	0.40	0.8	Index overlap-Genetics and marker weights	8	0.80	0.4	4

Row	Initial Queries				Expanded Queries				Overlap
	Query	Related documents	Retrieval	Accuracy	Query	Related documents	Retrieval	Accuracy	
14	Commodity-Agriculture Exchange	2	0.44	0.7	Commodity Exchange Strategies - Cohesion in Agriculture	7	0.78	0.4	4
15	Geographical information system	0	0.43	0.8	Application of Geographic information system in crisis management	8	0.57	0.6	6
16	Application of knowledge	2	0.50	0.9	Knowledge application and knowledge evaluation	9	0.64	0.7	7
17	New communication services	1	0.73	0.4	Quality of modern communication and electronic services	4	0.36	0.8	8
18	Exceptional child disorder	1	0.36	0.8	Exceptional children and their behavioral and emotional disorders	8	0.73	0.4	4
19	Digital services	2	0.29	0.7	Digital services in agriculture	7	1.00	0.2	2
20	Handicrafts and employment	3	0.58	0.8	The role of handicrafts in employment and production	8	0.67	0.7	7
21	Philosophy of information	3	0.50	0.9	Information philosophy and the importance of information in today's society	9	0.75	0.6	6
22	Advanced statistics	0	0.11	0.8	Advanced Statistics Analysis	8	0.89	0.1	1
23	Social Networks	1	0.33	0.9	The role of social networks on people's thoughts	9	0.75	0.4	4
24	Medicine	1	0.33	0.7	Traditional Iranian-Islamic medicine	7	0.78	0.3	3
25	Urban infrastructure	1	0.46	0.8	Monitoring of urban infrastructure and its importance	8	0.62	0.6	6
26	SWOT model	0	0.36	0.7	Strategic analysis based on SWOT model	7	0.64	0.4	4
27	Strategic Management	1	0.25	1	Strategic management of cultural engineering	10	0.83	0.3	3
28	Human resources and productivity	0	0.33	0.8	The role of human resources and productivity in the country's oil and gas company	8	0.67	0.4	4
29	Electronic human resources	1	0.36	0.8	The role and impact of electronic human resource management on	8	0.73	0.4	4

Row	Initial Queries				Expanded Queries				Overlap
	Query	Related documents	Retrieval	Accuracy	Query	Related documents	Retrieval	Accuracy	
					human resources				
30	Index + scientometrics	1	0.42	0.8	Investigating the correlation between scientific effectiveness indicators + scientometrics	8	0.67	0.5	5
31	Search engine retrieval function	2	0.44	0.7	Comparison of retrieval performance of text search engines	7	0.78	0.4	4
32	Intelligent transportation systems	2	0.55	0.7	Implementation of intelligent transportation system	7	0.64	0.6	6
33	The relationship between industry and academia	1	0.50	0.7	Comparative evaluation of industry managers' views on establishing a relationship with the university	7	0.58	0.6	6
34	Science and technology policy making	3	0.55	0.8	Challenges facing science and technology policy and solutions	8	0.73	0.6	6
35	Mass media	1	0.44	0.6	Mass media planning and policy making	6	0.67	0.4	4
36	Automatic indexing	1	0.45	0.7	An overview of automated web indexing approaches	7	0.64	0.5	5
37	Systemic view and Quranic concepts	2	0.54	0.8	The supply chain of a systematic view and comprehensive Quranic concepts	8	0.62	0.7	7
38	Man's existential relationship with the world	1	0.45	0.7	A Comparative Study of the Existential Relationship between Man and the Universe from the Perspective of Philosophers	7	0.64	0.5	5
39	Spiritual sharing	2	0.40	0.8	Spiritual participation in mysticism and transcendence	8	0.80	0.4	4
40	Rock blast	1	0.36	0.8	Rock blast using FLAC3D software	8	0.73	0.4	4

Table 2 presents the Wilcoxon test data to examine the difference between the three metrics of related retrieved documents, retrieval and accuracy for the two groups of primary and expanded queries. Based on the significance level of the two metrics of related retrieved documents and accuracy are different for the two groups.

Table 2.
Wilcoxon test data

Test Data	Related retrieved documents	Retrieval	Accuracy
Test statistics	-4.921	-4.809	-5.542
Significance level	.000	.000	.000

According to the data presented in Table 3, and given the difference in the mean of rankings, the performance of the expanded queries in the two metrics of the related retrieved documents and the accuracy of the retrieved results are better than the initial queries. As a result, the query expansion has improved the performance of retrieving scientific information of postgraduate students in the Elmnet search engine.

Table 3.
Rankings and mean of rankings

		Number	Mean of Rankings	Total Rankings
Related retrieved documents	Negative ratings	37	20.03	741.00
	Positive ratings	2	19.50	39.00
	Equal	1		
Retrieval	Negative ratings	2	22.75	45.50
	Positive ratings	37	19.85	734.50
	Equal	1		
Accuracy	Negative ratings	0	.00	.00
	Positive ratings	40	20.50	820.00
	Equal	0		

6. Discussion and Conclusion

Providing short and vague queries leads to irrelevant information retrieval which causes the user to be dissatisfied with the search. One solution to this problem is to expand the query by people or machines. Manual query expansion is mostly performed by experts and often in the context of searching for scientific information. This is since the experts and researchers in specific fields and disciplines have a higher expectation of adding more relevant words than ordinary searchers of public or news information. In this study, by requesting manual query expansion from postgraduate students, we sought to determine the effect of the query expansion on the performance of retrieving scientific information from the Elmnet search engine. The obtained results indicated that the expansion of the query in the above conditions improves the performance of the information retrieval system. A case study of the primary and expanded queries reveals that most of the expansion words added to the query aimed to limit and specify the query.

This means that most students start the query with more general and broader subjects, and then seek to specify and limit the query. Some query words with a lower ratio, seek to add synonyms to the words of the initial query, which indicates that some topics are indexed in databases with synonymous or equivalent words. Nowadays, automatic query expansion is on the agenda of databases. However, given that Persian databases are not fully developed, and the existence of special problems of writing in the Persian language, information literacy training and the method of defining and expressing information requirements and providing to the information retrieval systems, can have a significant impact on postgraduate students and researchers, to retrieve the required information and save them time and money.

References

Abdelmgeid Amin, A. (2008). Using a query expansion technique to improve document retrieval. Information Technologies and Knowledge, 7(2), 343-345.

Amolochitis, E., Christou, I. T., Tan, Z. H., & Prasad, R. (2013). A heuristic hierarchical scheme for academic search and retrieval. Information Processing & Management, 49(6), 1326-1343. https://doi.org/10.1016/j.ipm.2013.07.002

Azad, H. K., & Deepak, A. (2017). Query Expansion Techniques for Information Retrieval: A Survey. arXiv preprint arXiv:1708.00247.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval: Addison-Wesley Harlow, England.

Bendersky, M., & Croft, W. B. (2008, July). Discovering key concepts in verbose queries. In Proceedings of the 31st annual international ACM SIGIR conference on Research and

development in information retrieval (pp. 491-498). ACM https://doi.org/10.1145/1390334.1390419.

Efthimiadis, E. N. (1996). Query Expansion. Annual review of information science and technology (ARIST), 31, 121-87.

Ellis, D. and Haugan, M. (1997). Modeling the Information Seeking Patterns of Engineers and Research Scientists in an Industrial Environment. Journal of Documentation, 53(4), pp. 38-403. https://doi.org/10.1108/EUM0000000007204

Ellis, D. (1989). A Behavioral Model for Information Retrieval System Design. Journal of Information Science 45(3), pp. 171-212. https://doi.org/10.1108/eb026843

Lavrenko, V., & Croft, W. B. (2017, August). Relevance-based language models. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 260-267). ACM. https://doi.org/10.1145/3130348.3130376

Lee, K. S., Croft, W. B., & Allan, J. (2008, July). A cluster-based resampling method for pseudo-relevance feedback. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 235-242). ACM. https://doi.org/10.1145/1390334.1390376

Li, X., Schijvenaars, B. J., & de Rijke, M. (2017). Investigating queries and search failures in academic search. Information processing & management, 53(3), 666-683. https://doi.org/10.1016/j.ipm.2017.01.005

Li, X., & de Rijke, M. (2019). Characterizing and predicting downloads in academic search. Information Processing & Management, 56(3), 394-407. https://doi.org/10.1016/j.ipm.2018.10.019.

Makri, Blandford and Cox (2008). Investigating the Information-Seeking Behavior of Academic Lawyers: From Ellis's Model to Design. Information Processing & Management 44(2), pp. 613-634. https://doi.org/10.1016/j.ipm.2007.05.001

Mehdi, M., Okoli, C., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2017). Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus. Information Processing & Management, 53(2), 505-529. https://doi.org/10.1016/j.ipm.2016.07.003

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information science, 27(3),129-146. https://doi.org/10.1002/asi.4630270302

Russell-Rose, T., Gooch, P., & Kruschwitz, U. (2021). Interactive query expansion for professional search applications. Business Information Review, https://doi.org/10.1177/02663821211034079

Sharma, D. K., Pamula, R., & Chauhan, D. S. (2021). Semantic approaches for query expansion. Evolutionary Intelligence, 1-16. https://doi.org/10.1007/s12065-020-00554-x

Wollersheim, D. (2005). Dynamic query expansion for information retrieval of imprecise medical queries (Doctoral dissertation, La Trobe University).

Zhang, H. (2013). Query enhancement with topic detection and disambiguation for robust retrieval (Doctoral dissertation, Indiana University).