Journal of Theoretical and Applied Information Technology
31st March 2024. Vol.102. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEMANTIC QUERY EXPANSION METHOD FOR DIGITAL RESOURCE OBJECTS RETRIEVAL

WAFA' ZA'AL ALMA'AITAH[1,*], ABDULLAH ZAWAWI TALIB[2], ALAA OBEIDAT[3], MOHD AZAM OSMAN[4], FATIMA N. AL-ASWADI[5], RAMI S. ALKHAWALDEH[6]

[1]Department of Intelligence Systems, Faculty of Artificial Intelligence, Al-Balqa Applied University, Al-Salt 19117, Jordan

[2, 4]School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang

[3]Faculty of Science, Basic Sciences Department, The Hashemite University

[5]Institute of Computer Science and Digital Innovation, UCSI University, 56000 Kuala Lumpur, Malaysia

[5]Faculty of Computer Science and Engineering, Hodeidah University, Al Hudaydah, Yemen
[6]Department of Computer Information Systems, The University of Jordan, 77110 Aqaba, Jordan

*Corresponding Author

E-mail: [1]wafaa_maitah@bau.edu.jo, [2]azht@usm.my, [3]alaaf@hu.edu.jo, [4]azam@usm.my, [5]Fatima.Nadeem@ucsiuniversity.edu.my, [5]fatima_aswadi@hoduniv.net.ye, [6]r.alkhawaldeh@ju.edu.jo

ABSTRACT

Digital resource objects (DRO) consider as one of the most useful resources for storing humanity's collected knowledge. Many organizations are now aiming to make this data available to individuals. The query provided to DROs by the non-expert user, on the other hand, is usually a brief and frequently confusing expression of his desire. In DROs, it is not enough to explicitly explain what the user requires. The reason for coming up with short user query is that the users usually have limited knowledge and terminologies of the specific domain area. The formative terms can be missing inside the user's query, leading to poor coverage of relevant documents. To cover the difference between the query of user and DROs, the semantic query expansion method (SQE) is proposed to improve the efficiency of DRO retrieval by enhancing the quality level of candidate terms to be inserted semantically to the entire query terms to enhance performance of DRO retrieval. The proposed SQE method comprises three steps namely query terms definition, candidate terms generation and the proposed correlation algorithm. The aim of the correlation algorithm is to extract the semantic terms to extend the query with related terms only. Results from experiment on CHiC2013 and ECHiC2013_EDE collections show that the proposed method can significantly outperform previous methods specifically in DROs.

Keywords: *Query expansion method, Information retrieval, Digital resource objects, Semantic query*

1. INTRODUCTION

DRO indicates to structured information that depicts, and simplifies the retrieval, consumption, and administration of knowledge resources. Aside from contents storage, DROs provide platforms for searching, retrieving, and organizing data from databases. Standardized resource descriptions aid in the search and retrieval of digital information resources by characterizing singular files, singular objects, or entire groups [1]. DRO content can be shared, integrated, and aggregated online, and digital file content can be quickly updated. These capabilities help users of digitized content by improving access to digital libraries and allowing them to be reused for research, learning, and producing new commercial contents. In the field of DRO's content, it is necessary to address the challenge of imprecise and succinct queries often posed by non-expert users. For example, users might enter queries such as "ancient vase" or "famous painting" when searching for information about artifacts and artwork. Similarly, queries about historical events may include terms such as "World War II" or "American Revolution", which indicates a tendency for users to search for summary results on important events. Cultural practices, such as "traditional wedding customs" or the "Japanese tea ceremony", can also raise vague inquiries. When exploring monuments and landmarks, users may use

queries such as "famous monuments in France" or "history of the pyramids", reflecting their preference for quick access to relevant information. In addition, queries such as "Greek Myths" or "Shakespeare's Plays" illustrate users' tendency to seek general information about literary works and folklore. By incorporating these specific query examples, a comprehensive understanding of the challenges posed by non-expert users in the area of heritage culture content emerges. This, in turn, lays the groundwork for a solution that can focus on query optimization, context-based suggestions, or improved query interpretation to improve user experiences in accessing and understanding cultural heritage information.

In general, query of user is one of several alternative formulations of an information demand that a user may have. As a result, user queries frequently do not accurately reflect the vocabulary of a document that meets the information demand. Even if a document has the exact information that a user requires, it cannot be retrieved if it lacks any of the keywords utilized by the user's query. Query expansion (QE) method is a method that attempts to improve poor queries of user by eliminating terms that decrease the retrieval efficiency [2], adding terms that aid retrieval [3], re-weighting old or new query phrases to change the focus of the inquiry [4] or employing a combination of those methods [5]. Since DROs suffer from the short query problem, this paper will focus on the QE method that is concerned with adding and re-weighing the terms to the user's query. Basically, QE methods use several sources to provide the candidate terms that will be added to the user's query. These sources can be identified by two main sources: the first is an external source such as Wikipedia and WordNet [6], and the second is a set of results obtained from the first-round retrieval [7]. During a typical retrieval session, a user will repeatedly develop a query that meets the information requirement. The initial attempt to formulate a query with a particular data requirement in mind is frequently inaccurate and can result in an answer group that fails to satisfy the user's information need [8]. The query may have been too broad, resulting in a wide range of documents in the answer collection. As a result, important materials get mixed in with documents which are irrelevant to the topic at hand. Alternatively, the query could have been overly specific, returning only a portion of the relevant items. A third possibility is that the user's query is utterly defective, with no overlap between the expected collection of documents and the true result set.

Query expansion (QE) method is a method used to increase the effectivity of information retrieval (IR) performance by reformulating the original user's query [9]. Basically, it is based on the assumption that the query written by the user usually retrieves results that are mostly irrelevant [2]. QE method has been widely studied as an effective way to solve the short query problem [10]. To cover the gap between the user's query and DROs various QE methods for DRO retrieval have been proposed [11]. Because the effectiveness of QE approaches is based primarily on selecting suitable candidates who are semantically connected to the query terms many of the recent studies deployed semantic terms on the QE method [12]. However, to the best the researcher's knowledge, a few studies in DROs turned to semantic terms such as by [13], and it has shown that the semantic similarity for query expansion can help to enhance the efficiency of the DRO retrieval. Similarly, the proposed semantic query expansion approach (SQE) aims to increase DRO retrieval performance by enhancing the level of quality of candidate keywords to be appended semantically to the full query terms. The semantic concepts are derived by combining the suggested correlation-algorithm, which is based on certain simple sensible Boolean heuristics, with Wikipedia as an external resource. The proposed SQE method has the same steps as the traditional QE method. Nevertheless, it differs from previous work in two important points. First, the source of candidate terms is the top k documents specifically extracted from the content of metadata units rather than the metadata title. Second, an extra step is involved which is concerned with correlation algorithm that is the extraction of semantic terms in the context of the query. DROs may struggle to capture the complexity and dynamic nature of user intent and needs. The limitations of relying solely on DROs are directly related to challenges faced by non-

expert users. Non-expert users often have difficulties expressing precise queries, lack familiarity with specialized terms, and experience evolving information needs.

Semantic query expansion entails understanding the contextual nuances and underlying meanings behind the user's query in addition to finding exact matches to search terms. By adding a deeper knowledge of language semantics, user intent, and the complicated links between words, this revolutionary technology goes beyond the limitations of standard keyword-based retrieval. This strategy tries to bridge the gap between what users express in their queries and the rich content available in digital resource repositories by employing modern natural language processing techniques, machine learning, and semantic analysis [14].

We will delve into the complexities of semantic query extension and its implications for improving digital resource object retrieval in this exploration. We'll look at how modern technology are enabling the development of smarter, more intuitive search systems capable of comprehending human language nuances. We will also examine the obstacles and opportunities that occur when adopting such methods, such as data privacy concerns, algorithmic complexity, and the ethical usage of AI in information retrieval [15].

2. RELATED WORK

Numerous studies have shown that a substantial portion of online searches involve non-expert users generating imprecise or brief queries. For instance, [16] claimed that the average query length is between two and three words. Not compatible query and document words, as well as brief searches, can have a significant impact on the process of locating suitable documents. Moreover, a study by [17], [18], [19] highlighted that users often face challenges when formulating queries that accurately convey their information needs, leading to less effective search outcomes. Several researches reported the progress of query expansion (QE) methods with some modifications, such as application of semantic analysis that leads to enhance efficiency of document retrieval via conventional QE method. Based on similarity thesauri approach, [20] proposed a probabilistic semantic QE method that is developed in an automatic manner. The similarity thesauri method depends on the knowledge domain regarding a certain collection in which it was developed from. The terms with most similar to the query concept are used to expand the query instead of choosing terms that are similar to the very query terms. [21] proposed a sematic QE algorithm in order to enhance document search in massive repositories wherein the algorithm restructures the query via WordNet. [22] proposed a semantic QE technique that employs terms that reflecting similar expression or similar semantic by using Wikipedia and WordNet3 functioned as external sources. An Indian Patent database (excel format) was analyzed and the similarity of terms were determined via Cosine similarity and Extended Jaccard coefficients. The outcomes from expanded queries and the cosine method outperformed non-expanded queried and Jaccard coefficient respectively. [6] presented a novel method in order to extract more term correlation via Markov network for QE. The extracted term correlation which is derived from Wikipedia is amalgamated to a pre-built fundamental Markov network via single local corpus. Because of the user's lack of content and terminology knowledge across DRO collection, either too many or too few results are obtained from retrievals. Some studies have tackled the issue of DRO collection accessibility using the traditional QE approach to improve collection content retrieval. [23] developed a semantic QE with three versions for picking keywords from Wikipedia that incorporates extra similarity metrics and integrates both external and internal evidences for QE. The investigation included two collections, CHIC-2012 and CHIC-2013, demonstrating that the proposed strategy improves retrieval performance. [24] described varied retrieval ways to test the relative quality of different QE and improvement in semantics procedures. A variety of strategies based on blind-QE were used, with Wikipedia serving as the external resource. Different methods were employed to choose the most suitable concepts to be incorporated to the initial query [25]. The experiments were conducted on a CH content test collection. The experimental outcomes show that expanding the combination of pseudo-relevant documents and external resources enhances retrieval performance, with the influence of outside resource expansion being more substantial. [26] To tackle the term-mismatch problem, a semantic model based on semantic associations between indexing terms was

proposed. The model alters documents based on a query of user and some knowledge of semantic term relationships. It adds to the document by incorporating query phrases that are not present in the document but are semantically correlated with a minimum of one document keyword. It then combines the updated document with two LM smoothing algorithms, Dirichlet and Jelinek-Mercer (JM). The test was conducted using a variety of CLEF corpora from the medical area. In terms of retrieval performance enhancement, the experimental results show a significant improvement over standard LMs and appear to be superior to translation models. [4] presented a semantic enhancement QE approach that includes Wikipedia term linkages into LM. It handles brief queries that cannot express a precise information requirement. This method seeks the optimal phrases for a query in order to semantically enhance the topic and guess the user's information needs or query intent [14]. The experiment was carried out on a collection of CH content test results. The experimental results demonstrate minimal change when using the Porter stemming approach on term links since the difference between both outcomes is extremely minor, however the outputs of the semantic enrichment approach suggest that utilizing links out is more effective than using a mix of in and out links.

3. PROPOSED SEMANTIC QUERY EXPANSION METHOD

The flowchart of the proposed SQE method is briefly presented in Figure 1. Also, the pseudocode of the SQE method is shown in Algorithm 1. When queries become lengthier, there is a greater chance that some significant terms will appear in both the query and the corresponding documents [27]. Furthermore, these searches frequently contain confusing phrases. As a result, using the initially entered user query to obtain relevant pages is nearly impossible [28]. The proposed SQE method comprises three steps namely query terms definition, candidate terms generation, forming top N results and the correlation algorithm. The aim of the correlation algorithm is to extract the semantic terms to extend the query with related terms only. Each stage will be demonstrated in the subsections that follow.

The "Semantic Query Expansion Method for Digital Resource Objects Retrieval" is an information retrieval approach that uses semantic analysis to improve search precision. It broadens user queries by employing pretrained algorithms to identify semantically relevant phrases. The approach discovers contextually comparable phrases by averaging semantic representations of query terms. These words are added to the original query, resulting in a set of enlarged Queries. When applied to a set of digital resources, expanded searches return relevant results that contain either the original or expanded terms. As a result, the approach provides a more comprehensive and context-aware set of results, improving the search experience for digital resource items.

3.1 Query Terms Definition

The first step of the SQE method is to get the definition for each user's query term. To accomplish this, a list of 571 stopwords must be used to remove all stopwords from the queries, such as "a", "about", "the" and "her". Identifying the "meaning" of every term in the query of user due to semantic purpose is a important step for semantic retrieval. In order to find the definition for each term belonging to the user's query, Wikipedia is used as an external resource for the purpose of providing articles related to each keyword in the query of user. After obtaining the terms from the user's query, each term is sent separately to Wikipedia to retrieve the related articles for each term. Thus, each term has its corresponding articles.

Let $Q = \{t_1, t_2 \dots t_n\}$ where Q is the query, t_i is the query term and n is the number of query terms. To get definition D for each $t \in Q$, each t is sent to Wikipedia to retrieve related articles. Then, from the top three articles, the first paragraph is considered from every retrieved article to obtain definition D. This work follows the work of [29] in which the first three articles are employed to determine the range of candidate terms.

The definition terms are listed in a set $\mathcal{C}$ that contain definitions as elements $\mathcal{C} = \left(D_{1_{t_1}}, D_{2_{t_1}}, D_{3_{t_1}}\right), \left(D_{1_{t_2}}, D_{2_{t_2}}, D_{3_{t_2}}\right), \left(D_{1_{t_n}}, D_{2_{t_n}}, D_{3_{t_n}}\right)$. Later, set $\mathcal{C}$ will be an input for the correlation algorithm.

3.2 Identifying the Candidate Terms

The aim of this step is to generate candidate terms to be inserted to the query of user. The core of this step is to extract the keywords from the top k documents obtained from the first-round retrieval results. It is built based on two assumptions. The first assumption says that top the k documents are the

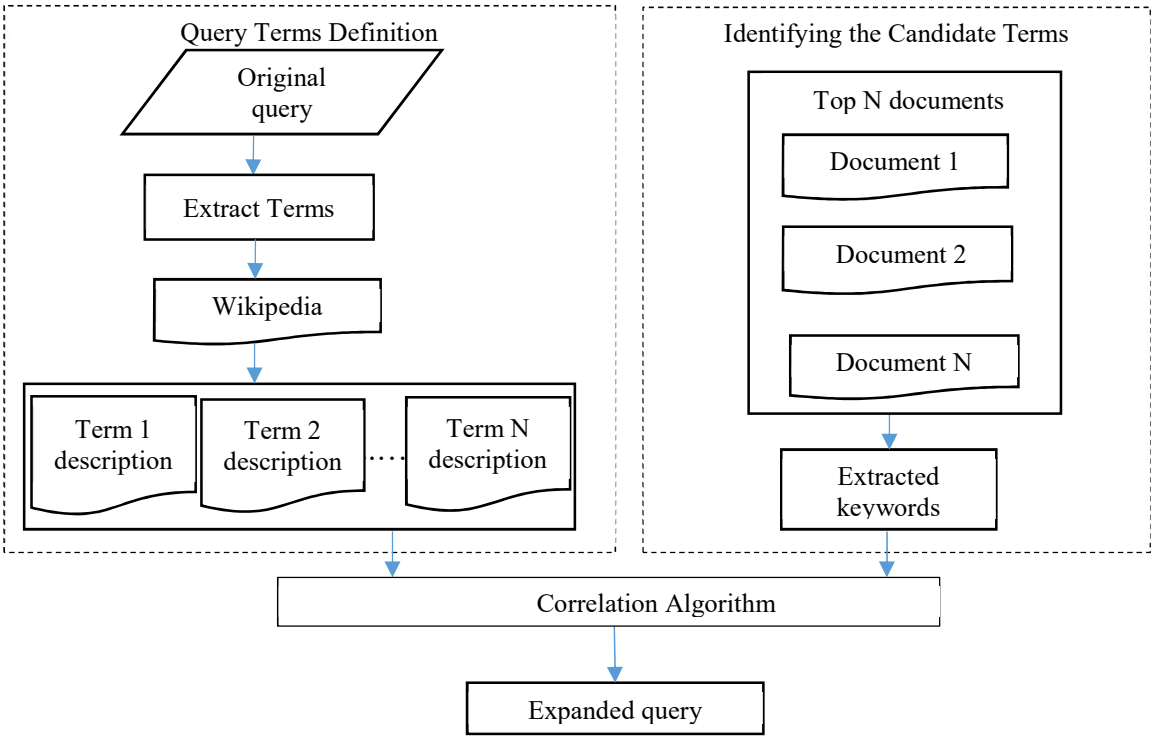

Figure 1: Semantic query expansion method

closest documents to the query [30] [31]. And the second assumption says that the frequency of the term in a single document determines how important the term is to the query [32].

To select the top k documents $D = \{d_1, d_2, \dots, d_k\}$, the k value needs to be determined where the best value for variable k is 10 and therefore, the first 10 retrieved documents from the results (first-round results) are the best range to extract the candidate terms [7]. Next, the domain of candidate terms is selected and the keywords (the degree of importance of a term in a document) have to be extracted from the selected domain (top k documents). The keyword is estimated based on the frequency of the term itself to the entire text of the document. The Frequency Inverse Document Frequency (TF-IDF) is a measure of informative terms in the documents. It is a basic measure that is used in almost all QE methods to calculate the weights of terms, i.e. user's query terms or candidate terms [33] and it is given as:

$$TF - IDF = tf(t) * log \frac{N}{df(t)} \quad (1)$$

where:

$tf(t)$: Number of times term t appears in document d,

N : Num of docs d in the collection,

$df(t)$: Num of docs d in corpus containing t and $df(t) \geq 1$.

Note that among the candidate keywords, as usual 100 keywords are taken to be candidate terms for each original query. The input of this step is the top k documents, and the output is the weight of the 100-candidate term.

3.3 Correlation Algorithm

Each user's query term definition obtained from the query terms definition step and the candidate terms coming from identifying the candidate terms step will be used as inputs to the proposed correlation algorithm (CA) to determine the important terms from each candidate terms that are to be added to the user's query. The proposed algorithm depends on some simple sensible Boolean heuristics.

Let $\mathcal{Q}$ be a set of query terms $t_i, \mathcal{C}$ be a set of definition $D, \mathcal{T}$ be a set of the weight of candidate terms, and $\mathcal{R}$ be a set of the relation describing which term belongs to which definition(s).

$\mathcal{Q} = \{t_1, t_2 \dots t_n\}$,

$\mathcal{C} = \left(D_{1_{t_1}}, D_{2_{t_1}}, D_{3_{t_1}}\right), \left(D_{1_{t_2}}, D_{2_{t_2}}, D_{3_{t_2}}\right) \dots \left(D_{1_{t_n}}, D_{2_{t_n}}, D_{3_{t_n}}\right)$,

$\mathcal{T} = \{t_{w1}, t_{w2}, t_{w3} \dots t_{wm}\}$

Algorithm 1: Proposed Semantic Query Expansion (SQE) Method

Input: User Query $\mathcal{Q} = \{t_1, t_2 \dots t_n\}$, Wikipedia, top 10 documents $D = \{d_1, d_2, \dots, d_3\}$
Step 1:
  For each $t_n \in \mathcal{Q}$ do
    Send t_n to Wikipedia // send it as query
    Return related articles
    Select the top 3 articles $\mathcal{C} = \{T_1, T_2, T_3\}$
    Extract the first paragraph $\mathcal{C} = \left\{\left(D_{1_{t_n}}, D_{2_{t_n}}, D_{3_{t_n}}\right)\right\}$ // read the first paragraph
  Return $\mathcal{C}$
Step 2:
  For each $d \in D$ do
    Tokenize/split the text into terms $\mathcal{T} = \{t_1, t_2, t_3 \dots t_m\}$
    For each $t_m \in \mathcal{T}$ do
      Compute $^{t}{}_{m}$ weight by using $TF - IDF = tf(t) * log\frac{N}{df(t)}$
      Rank the terms based on the weight $\mathcal{T} = \{t_{w1}, t_{w2}, t_{w3} \dots t_{wm}\}$
  Return $\mathcal{T} = \{t_{w1}, t_{w2}, t_{w3} \dots t_{w100}\}$ // top 100 terms
Step3:
  For each $t_{wm} \in \mathcal{T}$ do
    If $t_m \in \forall\{x: \mathcal{R}\}$ then
      $\mathcal{Q}_E = \mathcal{Q} \cup t_m$
    Else
      If $t_m \notin \forall\{x: \mathcal{R}\}$ then
        $\mathcal{Q}_E = \mathcal{Q}$
    Else
      If $t_m \in \exists\{x: \mathcal{R}\}$ then
        $\forall D_{t_n} \in \mathcal{C}$ do
          $\acute{\mathcal{C}} = \left(D_{1_{t_1}} \cup D_{2_{t_1}} \cup D_{3_{t_1}}\right) \cup \left(D_{1_{t_2}} \cup D_{2_{t_2}} \cup D_{3_{t_2}}\right) \cup \dots \cup \left(D_{1_{t_n}} \cup D_{2_{t_n}} \cup D_{3_{t_n}}\right)$
  Compute $SIM(t_m, \acute{C}) = \frac{\sum_{m=1}^{t}(t_m \cdot D_m)}{\sqrt{\sum_{m=1}^{t} t_m^2 \cdot \sum_{m=1}^{t} D_m^2}}$
  Rank the t_m based on the $SIM(t_m, \acute{C})$ value
Return $\mathcal{Q}_E$
Output: Expanded query $\mathcal{Q}_E$

$\mathcal{R} = \left(t_1, D_{1_{t_1}}, D_{2_{t_1}}, D_{3_{t_1}}\right), \left(t_2, D_{1_{t_2}}, D_{2_{t_2}}, D_{3_{t_2}}\right) \dots \left(t_n, D_{1_{t_n}}, D_{2_{t_n}}, D_{3_{t_n}}\right)$

It means that $D_{1_{t_1}}, D_{2_{t_1}}, D_{3_{t_1}}$ define t_1, $D_{1_{t_2}}, D_{2_{t_2}}, D_{3_{t_2}}$ define t_2 and so on.

The detailed steps of CA are as follows:

i. If the candidate term t_m occurs in all query terms' definitions, $t_m \in \forall\{x: \mathcal{R}\}$, where x is a pair of definitions in $\mathcal{R}$, then the query $\mathcal{Q}$ will be expanded to $\mathcal{Q}_E = \{t_1, t_2 \dots t_n, t_m\}$ with t_m directly.
ii. If the candidate term t_m does not occur in all query terms' definitions where $t_m \notin \forall\{x: \mathcal{R}\}$ where x is a pair of definitions in $\mathcal{R}$, then t_m will be discarded.
iii. Otherwise, when $t_m \in \exists\{x: \mathcal{R}\}$, where x is a pair of definitions in $\mathcal{R}$, then all query terms definitions in $\mathcal{C}$ are merged, and D which is a set of the common definitions is generated. The merging process starts from inside the pair of definitions in definitions in $\mathcal{C}$. It is taken as the union of definitions corresponding to specific term $(D_{1_{t_n}} \cup D_{2_{t_n}} \cup D_{3_{t_n}})$ in order to get rid of term duplicates, and then the union of the union definition terms together in set $\acute{\mathcal{C}}$ are as follows:

$\mathcal{C} = \left(D_{1_{t_1}}, D_{2_{t_1}}, D_{3_{t_1}}\right), \left(D_{1_{t_2}}, D_{2_{t_2}}, D_{3_{t_2}}\right) \dots \left(D_{1_{t_n}}, D_{2_{t_n}}, D_{3_{t_n}}\right)$

$$\acute{C} = \left(D_{1_{t_1}} \cup D_{2_{t_1}} \cup D_{3_{t_1}}\right) \cup \left(D_{1_{t_2}} \cup D_{2_{t_2}} \cup D_{3_{t_2}}\right) \cup \ldots \cup \left(D_{1_{t_n}} \cup D_{2_{t_n}} \cup D_{3_{t_n}}\right)$$

Then, the similarity $SIM(t_m, D)$ between t_m and common definition D is calculated using cosine similarity. It is a similarity measure that quantifies the cosine of the angle among two non-zero vectors in an inner product space. It is especially useful in positive space, where the result is neatly bounded in. The reason behind using it in this work is that it is the best measure compared to other similarity measures as reported in [34]. Also, cosine similarity measure works well when it cooperates with TF-IDF. In cosine similarity, each candidate term is handled as a single query, and set D that contains the terms' definition will be considered as a document. Therefore, cosine similarity will find the distance between each candidate term and D, and the candidate term with the shortest distance (the cosine value) is considered to be the closest to the document. Basically, the cosine equation is based on two coefficients: the weight of the candidate terms and the weight of the terms definition. For the first coefficient (weight of the candidate terms), the weights are found earlier as described in (Section 3.2) while the other coefficient is still not calculated. So, before using the cosine equation, we must find the second coefficient values. In the same way that is used to calculate the first coefficient, the weights of the second coefficient also calculated. The cosine similarity is mathematically expressed as:

$$SIM(t_m, D_m) = \frac{\sum_{m=1}^{t}(t_m \cdot D_m)}{\sqrt{\sum_{m=1}^{t} t_m^2 \cdot \sum_{m=1}^{t} D_m^2}} \quad (2)$$

where:

t : Length of definition set D,
D_m : Weight of the term m in definition set D,
t_m : Weight of the candidate term m.

Then, the candidate terms will be ranked based on its similarity with the closest terms are those having the highest similarity value, i.e. those with low angle and shortest distance.

iv. The last step in the SQE method is to add the closest candidate terms to the original query, and then automatically send them to the collection to retrieve the second-round results.

4. EXPERIMENTS, RESULTS AND DISCUSSION

Several experiments have been conducted to illustrate the efficacy of the suggested SQE approach. The major goal of the assessment is to demonstrate that the suggested SQE approach can affect the efficiency of the DRO retrieval. To handle the CHiC2013 and ECHiC 2013, typical mechanisms were created and implemented _EDE collections. For the evaluation purposes, the DS model is being used as the retrieval model to retrieve the CHiC2013 documents. Top 10 retrieved documents are to be obtained after query expansion. Figure 2 describes the experimental methods used to thoroughly assess the proposed SQE method system's performance against the existing methods. The first branch (IRS-1) is a normal IR system without QE, and the second branch (IRS-2) represents the IR system that applies QE, and both of them are considered as benchmarks. The third (IRS-3) and fourth (IRS-4) branches involve the suggested SQE method and they respectively use CHIC2013 English collection and its extended version ECHiC2013_EDE collection. The detail of the benchmarks will be discussed in the next section.

4.1 Experiment Setup

As a collection of tests, this study used CHIC-2013 English collection and its expanded version, ECHiC2013_EDE collection. The collection includes 1107 documents and 22 evaluation queries (on average, 1.6 terms per query). The evaluation queries represent each document's relevancy for each query. For all of the tests, the efficiency of the proposed method are reported using the three standard measures: Mean Average Precision (MAP), Precision at top 10 documents (p@10) and the Precision-Recall curve [35]. Two-tailed paired t-test is used to test whether the differences between the performances of the methods are statistically significant. An external resource such as Wikipedia (the English Wikipedia) was used to generate the related articles. Two IR system benchmarks were used. In both IR systems, the CHiC2013 collection is retrieved by the language model as described in [36].The first IR system (IRS_1) does not use the QE method and therefore only the initial results are subject to

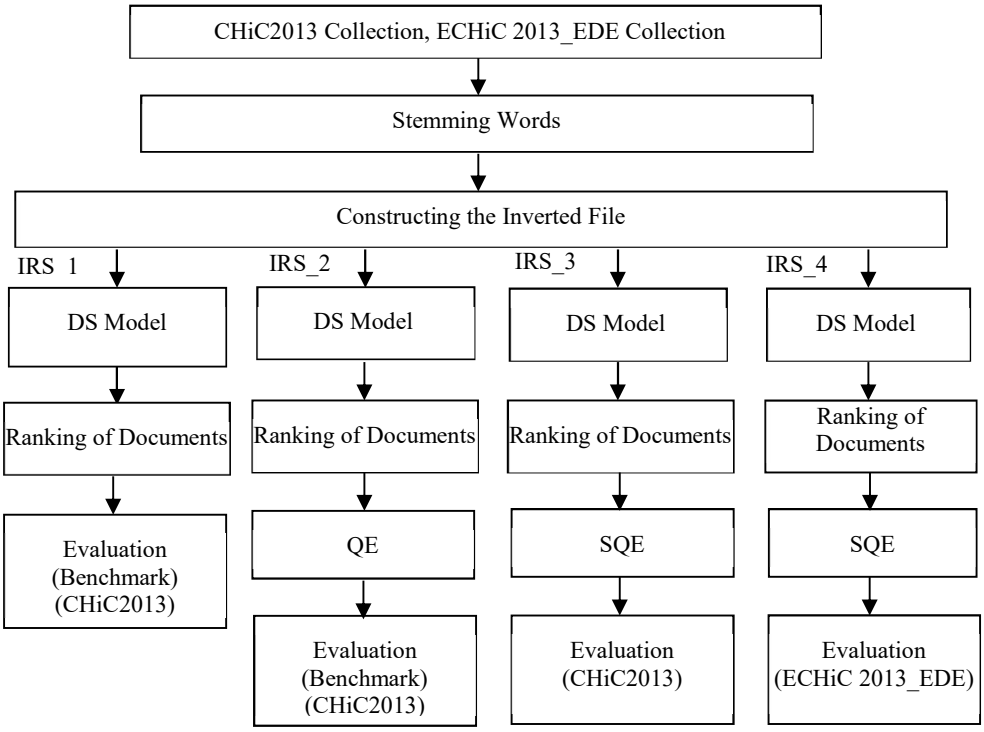

Figure 2: Methodology of the experiment (the labels refer to the branch number).

evaluation (the first-round results) while the second IR system (IRS_2) employs the QE method proposed by [37] and the second round results are subject to evaluation. The third IR system (IRS_3) employs the proposed SQE method. Furthermore, for the IR system handling ECHiC2013_EDE (IRS_4), documents were expanded using the SQE method. The experimental results for the proposed SQE method will be presented in detail in the next subsections. Table 1 presents some statistics of the test queries.

Table 1: Statistics of the test queries.

Parameter Name	Value
Number of testing queries	22
Number of Wikipedia articles	3
Average number of query terms	1.6
Number of candidate terms	100
Number of expanded terms	10

4.2 Experiment Results

The experiment rates the text documents based on 22 short inquiries. Table 2 displays the outcomes for IRS-1, IRS-2, IRS-3, and IRS-4.

According to this table, comparing MAP and P@10 for the suggested SQE approach to MAP and P@10 for the benchmarks results in a significant improvement in retrieval performance. In addition, Figure 3 shows the Precision-Recall curve for IRS-1, IRS-2, IRS-3 and IRS-4. Based on this figure, the precision of IRS-4 gains higher than IRS-3, IRS-2 and IRS-1 at different recall points, and it shows that the SQE method helps to improve the performance of DRO retrieval in both collections, CHiC2013 and ECHiC 2013_EDE. It is observed that the IR system for DROs achieves a significant performance when the QE method is employed especially when semantic terms work with the QE method as in the SQE method. As aforementioned, it is worth to highlight that the retrieval results on ECHiC 2013_EDE are better than the retrieval results on CHiC2013. The proposed SQE method depends on the top 10 results to determine the candidate terms. The effectivity of the candidate terms will be more effective and closer to the user's query. All at once, the results show that the recommended SQE approach is more suitable for DRO collection and better than the QE method.

Table 2: and P@10 MAP for IRS-1, IRS-2, IRS-3, and IRS-4

Information Retrieval System	MAP	P@10
IRS-1	0.502	0.419
IRS-2	0.540	0.488
IRS-3	0.590	0.531
IRS-4	**0.741**	**0.656**
Improvement (IRS-2, IRS-1)	3.800%	6.900%
Improvement (IRS-3, IRS-1)	8.800%	11.200%
Improvement (IRS-3, IRS-2)	5.000%	4.300%
Improvement (IRS-4, IRS-1)	23.900%	23.700%
Improvement (IRS-4, IRS-2)	20.100%	16.800%
Improvement (IRS-4, IRS-3)	15.100%	12.500%

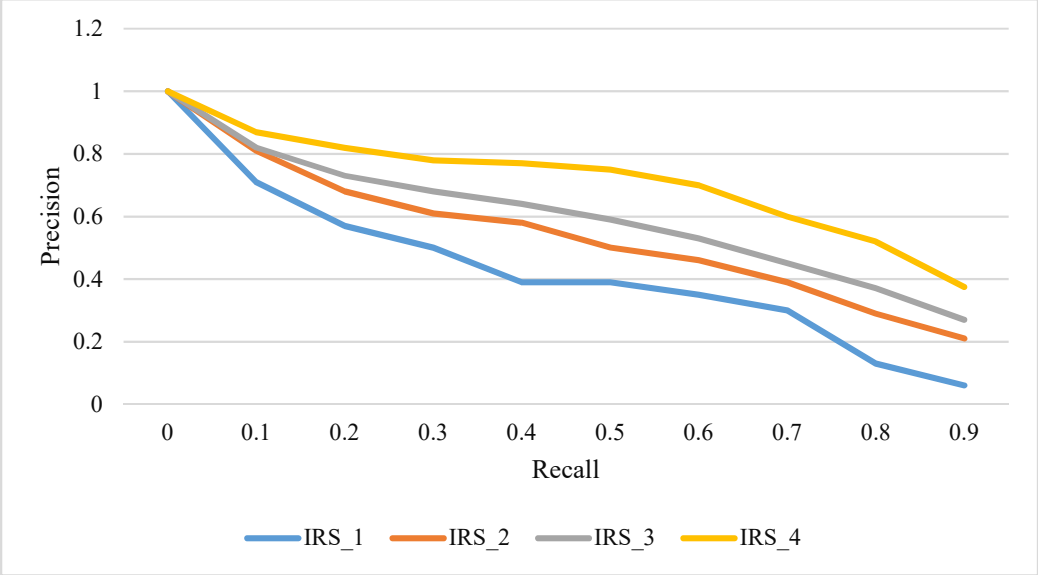

Figure 3: Comparison of IRS _1, IRS SI_2, IRS _3 and IRS_4 using averaged 9-point Precision Recall curve.

4.3 Statistical Significance Analysis

The outcomes of the t-test for all pairs of approaches are depicted in Table 3. It is worth noting that in all circumstances the p-value is less than 0.05 with respect to the MAP measure which implies that the retrieval performance of IRS_3 and IRS_4 are much superior to the corresponding comparator standards (IRS_1 and IRS_2).

As a result, because all of these p-values were less than 0.05 at a 95% confidence level, it is proven that the IRS_3 and IRS_4 (with the proposed SQE method) statistically outperform IRS_1 (without QE) and IRS_2 (with QE). Moreover, for the P@10 measure, it is clear from the table that not all cases have p-value less than 0.05. For example, the experiment for IRS_3 vs IRS_2 has p-value equals to 0.1344 and the experiment for IRS_4 vs IRS_3 has p-value equals to 0.1431.

Table 3: Paired t-test analysis on both measures MAP and P@10 for IRS_1, IRS_2, IRS_3, and IRS_4.

Information System	Measure			
	MAP		p@10	
	t-value	P-value	t-value	P- value
IR_2 vs IRS_1	3.953	7.263E-04	2.300	2.936E-02
IRS_3 vs IRS_1	9.985	1.989E-09	4.431	1.401E-04
IRS_3 vs IRS_2	9.442	3.365E-10	1.547	1.344E-01
IRS_4 vs IRS_1	12.280	8.624E-13	5.337	1.104E-05
IRS_4 vs IRS_2	11.623	7.309E-11	2.479	1.971E-02
IRS_4 vs IRS_3	5.347	2.653E-05	1.506	1.432E-01

Hence, there is no significant difference between the performance of IRS_3 and IRS_2, and between the performance of IRS_4 and IRS_3. Both IRS_3 and IRS_2 exhibit similar performance, the same is true for both IRS_4 and IRS_3. In addition, the p-values are less than the level (p 0.05), indicating that there is a variance between comparing techniques' averages and that the improvement in both MAP and P@10 did not happen by accident (with the exception of the prior two cases). A p-value of 0.0000000003365%, for example, indicates that there is only a (0.0000000003365%) chance that the findings of the IRS-3 vs IRS-2 experiment happened by coincidence. Because the P_values in all of the suggested approaches are low, it has been demonstrated that the results did not occur by coincidence and that enhancements are meaningful.

5. CONCLUSION

In this study, a SQE approach for improving DRO retrieval performance is proposed. The SQE approach aims to overcome the short user query problem, which has a detrimental impact on retrieval performance. The recommended SQE method enhances the level of quality of the potential terms to be included semantically to the entire query terms, and the resulting semantic terms are obtained through the use of the proposed correlation algorithm, which is based on some simple and effective Logical heuristics, and Wikipedia as an external resource. The SQE approach can improve the performance of DRO retrieval. There were two IR benchmarks used. Because the primary IR does not use the QE approach, just the initial results (the first-round results) are evaluated, whereas the second IR uses the usual QE technique and the second-round findings are evaluated.

The P@10, MAP, and Precision-Recall curves were employed for evaluation. As datasets, the CHiC2013 and ECHiC2013_EDE sets are used. The results suggest that the proposed technique can improve DRO retrieval performance when compared to other benchmarks.

REFERENCES:

[1] A. Kalisdha and C. Suresh, “Digital libraries: definitions, issues and challenges”, *Science and Humanities*, vol. 95, 2017.

[2] V. U. Thompson, C. Panchev, and M. Oakes, “Performance evaluation of similarity measures on similar and dissimilar text retrieval”, in *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, IEEE, 2015, pp. 577–584.

[3] M. T. Artese and I. Gagliardi, “Integrating, Indexing and Querying the Tangible and Intangible Cultural Heritage Available Online: The QueryLab Portal”, *Information*, vol. 13, no. 5, p. 260, 2022.

[4] M. A. Raza, M. Ali, M. Pasha, and M. Ali, “An Improved Semantic Query Expansion Approach Using Incremental User Tag Profile for Efficient Information Retrieval”, 2022.

[5] K. L. Tan and C. K. Lim, “Language model: Extension to solve inconsistency, incompleteness, and short query in cultural heritage collection”, in *AIP Conference Proceedings*, AIP Publishing LLC, 2017, p. 20138.

[6] L. Gan and H. Hong, “Improving query expansion for information retrieval using Wikipedia”, *International Journal of Database Theory and Application*, vol. 8, no. 3, pp. 27–40, 2015.

[7] S. Kumar, A. S. Ray, S. Kamila, A. Ekbal, S. Saha, and P. Bhattacharyya, “Improving

document ranking using query expansion and classification techniques for mixed script information retrieval", in *Proceedings of the 13th International Conference on Natural Language Processing*, 2016, pp. 81–89.

[8] H. ALMarwi, M. Ghurab, and I. Al-Baltah, "A hybrid semantic query expansion approach for Arabic information retrieval", *J Big Data*, vol. 7, no. 1, pp. 1–19, 2020.

[9] A. Krishnan, S. Ranu, and S. Mehta, "Leveraging semantic resources in diversified query expansion", *World Wide Web*, vol. 21, pp. 1041–1067, 2018.

[10] F. C. Fernández-Reyes, J. Hermosillo-Valadez, and M. Montes-y-Gómez, "A prospect-guided global query expansion strategy using word embeddings", *Inf Process Manag*, vol. 54, no. 1, pp. 1–13, 2018.

[11] S. Jain, K. R. Seeja, and R. Jindal, "A fuzzy ontology framework in information retrieval using semantic query expansion", *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100009, 2021.

[12] J. Singh and A. Sharan, "Co-occurrence and semantic similarity based hybrid approach for improving automatic query expansion in information retrieval", in *Distributed Computing and Internet Technology: 11th International Conference, ICDCIT 2015, Bhubaneswar, India, February 5-8, 2015. Proceedings 11*, Springer, 2015, pp. 415–418.

[13] B. Ahamed, R. M. S. Najimaldeen, and Y. Duraisamy, "Enhancement framework of semantic query expansion using mapped ontology", in *2020 international conference on computer science and software engineering (CSASE)*, IEEE, 2020, pp. 56–60.

[14] A. Obeidat and R. Yaqbeh, "Business Project Management Using Genetic Algorithm for the Marketplace Administration", *Journal of Internet Services and Information Security*, vol. 13, no. 2, pp. 65–80, 2023, doi: 10.58346/JISIS.2023.I2.004.

[15] W. Z. Alma'aitah, Z. T. Abdullah, and M. A. Osman, "Document expansion method for digital resource objects", in *2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT)*, IEEE, 2019, pp. 256–260.

[16] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang, "Query clustering using user logs", *ACM Trans Inf Syst*, vol. 20, no. 1, pp. 59–81, 2002.

[17] B. M. Wildemuth, D. Kelly, E. Boettcher, E. Moore, and G. Dimitrova, "Examining the impact of domain and cognitive complexity on query formulation and reformulation", *Inf Process Manag*, vol. 54, no. 3, pp. 433–450, 2018.

[18] K. Affolter, K. Stockinger, and A. Bernstein, "A comparative survey of recent natural language interfaces for databases", *The VLDB Journal*, vol. 28, pp. 793–819, 2019.

[19] H. Zamani, S. Dumais, N. Craswell, P. Bennett, and G. Lueck, "Generating clarifying questions for information retrieval", in *Proceedings of the web conference 2020*, 2020, pp. 418–428.

[20] M. A. Khedr, F. A. El-Licy, and A. Salah, "Ontology based semantic query expansion for searching queries in programming domain", *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021.

[21] N. Wu and Y. Pan, "Semantic query expansion method based on pay-as-yougo fashion for graph model", in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 12094.

[22] A. Allahim, A. Cherif, and A. Imine, "A Hybrid Approach for Optimizing Arabic Semantic Query Expansion", in *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 2021, pp. 1–8.

[23] F. Beirade, H. Azzoune, and D. E. Zegour, "Semantic query for Quranic ontology", *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 6, pp. 753–760, 2021.

[24] X. Cui, P. Zhai, and Y. Fang, "Semantic Query Expansion based on Entity Association in Medical Question Answering", in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 12022.

[25] D. K. Sharma, R. Pamula, and D. S. Chauhan, "Query expansion–Hybrid framework using fuzzy logic and PRF", *Measurement*, vol. 198, p. 111300, 2022.

[26] A. R. G. Purnama, I. N. Yulita, and A. Helen, "Search System for Translation of Al-Qur'an Verses in Indonesian using BM25 and Semantic Query Expansion", in *2021 International Conference on Artificial Intelligence and Big Data Analytics*, IEEE, 2021, pp. 1–7.

[27] A. Dhokar, L. Hlaoua, and L. Ben Romdhane, "Tweet Contextualization Approach Using a Semantic Query Expansion", *Procedia Comput Sci*, vol. 192, pp. 387–396, 2021.

[28] A. Obeidat and R. Yaqbeh, "Smart Approach for Botnet Detection Based on Network Traffic Analysis", *Journal of Electrical and Computer Engineering*, vol. 2022, pp. 1–10, 2022.

[29] N. Zhang, J. Wang, Y. Ma, K. He, Z. Li, and X. F. Liu, "Web service discovery based on goal-oriented query expansion", *Journal of Systems and Software*, vol. 142, pp. 73–91, 2018.
[30] A. Nehar, S. Bellaouar, D. Mahfoud, and F. Z. Daoudi, "A Hybrid Semantic Statistical Query Expansion for Arabic Information Retrieval Systems", in *2022 5th International Symposium on Informatics and its Applications (ISIA)*, IEEE, 2022, pp. 1–6.
[31] A. Obeidat and M. Al-shalabi, "An efficient approach towards network routing using genetic algorithm", *International Journal of Computers Communications & Control*, vol. 17, no. 5, 2022.
[32] H. K. Azad, A. Deepak, and K. Abhishek, "Query Expansion for Improving Web Search", *J Comput Theor Nanosci*, vol. 17, no. 1, pp. 101–108, 2020.
[33] A. Jain, A. Jain, N. Chauhan, V. Singh, and N. Thakur, "Information retrieval using cosine and jaccard similarity measures in vector space model", *Int. J. Comput. Appl*, vol. 164, no. 6, pp. 28–30, 2017.
[34] S. Lartariyatham, P. Wuttidittachotti, S. Prakancharoen, and S. A. Vallipakorn, "Comparative Weighting Methods of Vector Space Model", *ARPN Journal of Engineering and Applied Sciences*, pp. 1316–1323, 2017.
[35] P. Manning, *Drugs and popular culture*. Taylor & Francis London, UK, 2013.
[36] W. Z. Alma'aitah, A. Z. Talib, and M. A. Osman, "Language model for digital recourse objects retrieval", *J Theor Appl Inf Technol*, vol. 97, no. 11, pp. 2871–2881, 2019.
[37] M. Maryamah, A. Z. Arifin, R. Sarno, and Y. Morimoto, "Query expansion based on Wikipedia word embedding and BabelNet method for searching Arabic documents", *International Journal of Intelligent Engineering & System*, vol. 12, no. 5, pp. 202–213, 2019.