Semantic Approaches for Query Expansion: Taxonomy, Challenges, and Future Research Directions

Azzah Allahim*1,2, Asma Cherif2,3, and Abdessamad Imine4

1Jouf University, College of Computer and Information Sciences, Sakaka, Saudi Arabia.
2King Abdulaziz University, Faculty of Computing and Information Technology, Jeddah, Saudi Arabia, acherif@kau.edu.sa
3The Center of Excellence in Smart Environment Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia
4Lorraine University and LORIA-CNRS-INRIA Nancy Grand Es, Nancy, France, abdessamad.imine@loria.fr

Abstract
As the internet is flooded with an enormous amount of information, information retrieval systems are struggling to deliver ideal results to users. However, query expansion techniques have emerged as a game-changer, significantly enhancing information retrieval. Lately, semantic query expansion techniques have garnered more attention from researchers, as they offer more relevant and practical outcomes to users. By utilizing these techniques, users can retrieve more meaningful and useful information from the web. At present, there exist only a limited number of research works that comprehensively examine semantic query expansion. It is imperative to have a lucid understanding of the various semantic query expansion methods in order to ascertain their capabilities and drawbacks. This article presents a thorough overview of query expansion methods, with a particular emphasis on semantic approaches. It reviews recent frameworks developed between 2015 and 2022 and assesses the limitations of each approach. The challenges inherent in the semantic query expansion field are discussed, along with future research directions. The study highlights the linguistic approach as the most efficient and flexible path for researchers to pursue, while the ontology approach is preferred for domain-specific search applications. Therefore, developing the ontology field can pave the way for future advancements in semantic query expansion. Additionally, leveraging AI and focusing on the query context rather than user intervention can lead to more optimal expanded queries.

Keywords: Query Expansion, Semantic Query, Semantic Web, Ontology

1 Introduction
The internet has become a significant source of information, and prominent companies in industries like healthcare, advertising, and trading are using it to publish their services. Furthermore, scholars and researchers use the internet to share their thoughts and experiences. To make the most of this vast amount of information, the semantic web was developed. Its main idea is to connect information based on its meaning. As a result, the number of internet users is rapidly increasing, and people are using search engines and other tools to find information. Because of the diverse cultural and educational backgrounds of internet users, they have different ways of searching for information, using various formats and structures for their queries. Some prefer using many statements, while others use short statements.

The primary link between users and web information is formed by Information Retrieval Systems (IRS). It is crucial that these systems cater to the needs of users regardless of the query structure. When searching for information, users typically enter their query into the IRS and receive relevant documents in response. With the rapid growth of web information, retrieval systems face the challenge of accommodating diverse queries. They must be flexible yet

*Corresponding author: E-mail(azzah.allahim@gmail.com)
effective in serving both information seekers and providers. A major challenge arises when users submit short queries, which is increasingly common [1]. Short queries expand the search scope, resulting in the retrieval of more irrelevant documents.

Commercial search engines use keyword-based mechanisms to retrieve search results that match the exact terms in the user’s query. However, this approach can be limiting for users who struggle to articulate their needs, resulting in an "intention gap" [2]. To address this issue, researchers and developers have introduced several techniques, including query enhancement, suggestion, relaxation, and expansion [3]. Rather than relying on syntactic matching, these techniques aim to improve the contextual meaning of the user’s query. [4]. Query expansion (QE), for instance, involves adding related keywords to the original query to broaden the search area and obtain more results. For example, a search for "study supplies" could be expanded to include terms like “student”, “students” and “supply”. The expanded terms are integrated into the original query to yield more accurate results.

To produce better query results, the IRS is now incorporating the semantic vision of the query. This involves using semantic query expansion, which involves adding related terms to the query that are semantically related to the original terms. For example, if the original query included terms like "student" and "education," the expanded query might also include terms like “schoolboy”, “learner”, “provide”, “supply”, “stock” and “equipment”. By adding these related terms, the IRS can generate more relevant search results.

Query expansion (QE) can be classified into three main categories based on the source used: local, global, and semantic. Local-source QE involves using user feedback and query logs as an expansion source. Global-source QE uses a language model corpus as the primary expansion source. Knowledge-based or semantic expansion involves using semantic corpuses. Semantic query expansion involves reformulating the query with different words that have the same meaning to produce more relevant results. Three approaches can be used: linguistic-based, ontology-based, and hybrid [5]. The linguistic-based approach uses linguistic techniques to extrapolate the semantic relationship between words with the help of linguistic dictionaries. The ontology-based approach uses a semantically constructed tree or ontology that connects different topics based on their meaning. The hybrid approach combines both of these approaches.

Despite the importance of query expansion, there are few comprehensive works that delve into the topic. For example, in their paper [6], the authors attempted to clarify the concept of query expansion by examining a wide range of previous works. While this may be useful for advanced researchers, it can be confusing for those who are new to the field. On the other hand, in [5], the authors provide a comprehensive taxonomy of query expansion approaches, but they only present and analyze a few frameworks. Furthermore, the early steps of the query expansion process are not detailed enough. Finally, in [3], the authors give a brief and general overview of query reformulation definitions without discussing them in relation to recent and effective frameworks.

To contribute to the advancement of the field of query expansion and assist researchers in further developing this area of study, the main contributions of this paper are:

- Presenting semantic query expansion along with its definition, approaches, and techniques used.
- Providing a comprehensive review of related works based on the approach used.
- Discussing the most used techniques and extrapolating the main challenges.
- Browsing some open issues in the field to help researchers construct future directions for query expansion.

The remainder of the paper is organized as follows. Section 2 present some recent works that summarizes QE aspects. Section 3 discusses the background and refinement techniques. Section 2 presents some related work. Section 5 discusses some recent frameworks based on the Linguistic-based approach. Section 6 presents the Ontology-based approach, and Section 7 discusses the Hybrid approach. Section 8 summarizes the challenges and gives an overview on open issues. Finally, Section 9 concludes the paper.

2 Related work

It is important to compare and discuss the existing works surveying QE research to establish the novelty and significance of the present research.

In their research, Selvaretnam et al. [2] explored the challenges of query expansion, specifically addressing word ambiguity and query-document vocabulary mismatch. To overcome these challenges, the authors suggested studying
the linguistic characteristics of the query terms, including their morphological and syntactical features. Additionally, identifying the relationships between query terms can help in understanding the direct and indirect connections between a document and the query, thus mitigating the issue of query-document vocabulary mismatch. The authors approached the query expansion process analytically, breaking it down into a set of modules and describing the key concepts of each one. While their work may be more beneficial for advanced researchers, it provides valuable insights into the complexities of query expansion.

Ooi et al. [3] conducted a study that compared different query expansion techniques. They provided a comprehensive overview of query expansion methods, including relevance feedback, language model, and corpus-based models. Additionally, they analyzed these techniques in detail and highlighted their differences. The study also discussed the dissimilarities between query expansion, query refinement, and query suggestion. The authors concluded by outlining the benefits and drawbacks of each technique.

Then, Raza et al. [5] made a significant contribution to the discussion of semantic QE and its importance. They discussed various approaches to semantic QE and the methods used to evaluate QE. The researchers concluded their work by briefly outlining two future directions in the field. The study sheds light on the importance of semantic QE and the different ways it can be approached, as well as the need for continued research in this area.

As query expansion has proven to be one of the most effective solutions to challenges in information retrieval systems [5], this paper explores the different types and main steps of query expansion. Additionally, it provides an in-depth examination of the semantic perspective of query expansion by defining and discussing various semantic query expansion methods. It also addresses the gap in understanding the concepts and main steps of semantic query expansion for researchers in this field. Furthermore, it summarizes recent high-quality works in the field, covering a diverse range of techniques and input types, including those for image search and different languages including Arabic. A historical perspective of the techniques used and the most common dictionaries is presented in tabular form. We gathered frameworks from digital repositories such as ScienceDirect, IEEEExplore, SDL, and Web of Science, covering publications from 2015 to 2022. To accurately search for relevant sources, multiple keywords were used, including “query expansion”, “semantic expansion”, “query recommendation”, “query similarity”, “semantic web”, “cosine similarity”, and “terms rank”.

A table comparing the present paper with recent existing related works in terms of various QE aspects is presented in table 1. The aspects include QE pipeline, approaches, challenges, open issues, most recent LR, and LR datatype. This table provides a brief comparison between the present paper and the existing work.

<table>
<thead>
<tr>
<th>Ref</th>
<th>Pipeline</th>
<th>Approaches</th>
<th>Challenges</th>
<th>Open issues</th>
<th>Up-to-date LR</th>
<th>LR datatype: text, images</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[5]</td>
<td>✓</td>
<td>Partially</td>
<td>Partially</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our paper</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: A brief comparison between our paper and the recent existing related work in terms of different QE aspects

According to the table, previous research on QE has covered certain aspects but has not fully addressed all of them. For example, in [2], the authors discussed QE approaches, challenges, and open issues, but did not provide a complete QE pipeline, and only focused on text-based works. On the other hand, [3] and [5] only presented QE approaches without covering other important aspects. Thus, previous works have offered a technical perspective on QE, highlighting its approaches, challenges, and open issues. In contrast, our paper aims to present a comprehensive view of QE, catering to both beginners and researchers. It offers an introduction to QE through its complete pipeline and approaches, as well as an in-depth analysis of challenges and future directions by presenting the latest works with various datatype frameworks.

### 3 Background

Traditionally, when a user passes a query to the IRS, documents with the related terms will be retrieved and ordered based on statistical computations. However, with the rapid expansion of information, which has left us with an endless number of documents, this approach loses its efficiency. Therefore, the query should be considered with different yet effective perspectives such as query expansion, in which the query is reformulated with more related terms.
The query expansion philosophy has been introduced and examined to improve the query results, thereby increasing its effectiveness. The technique relies on expanding the query terms by adding additional ones that are related to the original ones, then exploring the results based on the expanded query [7, 8, 9]. This can help overcome some vocabulary mismatch and query ambiguity problems. These problems appear because of either polysemy or synonyms [3]. Polysemy refers to different definitions of a word such as the word “python”, which can mean a snake or a programming language. Synonyms, on the other hand, are different words with a similar meaning such as “words” and “lyrics”.

Generally, the query expansion technique is carried out through a pipeline structure as illustrated in Figure 1. First, in the preprocessing step, the query is filtered to maintain only the useful terms. Then, in the feature extraction, the terms are transferred into word vectors to prepare them for the expansion step. Consequently, the candidate terms are selected by measuring their similarities with the original query terms. Furthermore, the selected ones are ranked based on their similarity scores. Finally, the query is reformulated based on the top-ranked terms. These steps are detailed in the following.

### 3.1 Preprocessing

The first step is the preprocessing of the query. This is done by removing unwanted terms, such as stop words, and keeping only the important ones. For example, if the user searches using the query “What is the capital of Saudi Arabia?”, then the words “is”, “the”, and “of” do not play an important role in the search process. Instead, they can affect the efficiency of the search application by adding extra and unneeded work. Preprocessing can be accomplished using many steps based on the application needs.

The main preprocessing procedures are tokenization, data cleaning and stemming as shown in Figure 2.

**Tokenization** In this step, the text is transformed into tokens, which are useful and meaningful elements. For this, we split the text into words or phrases based on the application’s workflow.

**Data cleaning** After breaking the text into tokens, the unnecessary ones will be eliminated. Unwanted tokens could be stop words, symbols, URLs, extra white spaces, or words in different languages if the application only deals with a
### Table 2: Example of features presentation: children features which are their: age, Math grades and English grades.

<table>
<thead>
<tr>
<th>Child name</th>
<th>Age</th>
<th>Math</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>5</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Sara</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Noor</td>
<td>6</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Mike</td>
<td>3</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Suen</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

specific language and the text is written in more than one language.

**Stemming**  
In this step, all the words are changed to their root word. Using the stem form is much more useful than the entered text since most of the external knowledge sources (i.e., dictionaries, lexicons, thesaurus, and ontology) are structured using the stem form of the words.

Many tools and libraries can be used in the preprocessing phase such as NLTK\(^1\), spaCy\(^2\) and Gensim\(^3\), which can be used for cleaning data and stemming.

### 3.2 Features extraction

The second step in the query expansion process is the extraction of the expansion features. This step is the main procedure that differentiates between the query expansion approaches and can be done in several ways.

After gathering useful words from the query to be expanded, they are transformed into vectors. The main idea of presenting the text as vectors is to determine the weight and dimensions of each word for further comparison. As illustrated in Figure 3, each word will be transformed into a vector with a direction and a magnitude that represents its meaning. Consequently, identifying and manipulating the linguistic similarity of the words will be conceivable. To construct a vector, its dimensions (i.e. characteristics or features) need to be identified. For example, if we would like to present a group of children at school, we should first define the features of each child. As is shown in Table 2, we have chosen their Math grades, English grades and age as the desired features and dimensions. For instance, the word "Jack" will be represented by the values of its features, which are 5,10 and 7 in the table 2. Therefore, "Jack" will be located in the vector space as the point (5,10,7). Consequently, each record in the table represents a vector of a child. Now, by using the vectors, we can present the data in the space as a point in the vector space as illustrated in Figure 4. From the representation, we can conclude that Sara and Suen have similar features since they are close to each other in the vector space. On the other hand, Mike and Noor have different feature since their points are far away from each other in the vector space.

![Figure 3: General approach to measure the similarity in texts.](https://example.com/fig3.png)

---

\(^1\)Natural Language Toolkit. [https://www.nltk.org](https://www.nltk.org)

\(^2\)Industrial-Strength Natural Language Processing. [https://spacy.io/](https://spacy.io/)

\(^3\)Gensim:Topic modelling for humans. [https://radimrehurek.com/gensim/](https://radimrehurek.com/gensim/)
To determine the features, several useful natural language processing (NLP) algorithms can be used. In the following, some of the widely used algorithms are defined (e.g., term frequency-inverse document frequency (TF-IDF), bag of words (BOW), continuous bag of words (CBOW) and Word2vec). These algorithms are shown in Figure 5.

**TF-IDF: term frequency-inverse document frequency.** TF-IDF is a statistical model that aims to identify the importance of each word to a specific document in a collection of documents. In doing so, it computes the weight of each word based on its appearance in the document; however, it is offset by the number of documents that contain the word. The main logic behind TF-IDF is that it gives high attention to words that rarely appear in the corpus [10].

**BOW (Bag of Words).** BOW is an approach that takes advantage of the occurrences of the most frequently used words in a text, which are collected as a bag of words. Indeed, it builds the vector based on the presence of the words in that bag. It assigns 1 to a word if it appears in the bag and 0 otherwise [11].

**CBOW: Continuous bag of words.** CBOW is an approach that applies BOW to the context around an input word to predict a target word. That target word, indeed, has a co-occurrence relationship with the input word. CBOW is mainly used in neural nets for vector predictions [5, 12].

**Skip-Gram.** Skip-Gram is also a supervised learning method used in neural nets. It is used for the same purpose as CBOW; however, it is used to predict the context given the word [12].

**Word2vec.** Mikolov et al. introduced Word2vec as an open-source project in [13] and [14]. It is a neural net-trained model with two layers. Without human intervention, Word2vec processes text by converting the words to vectors. It generates a full numerical representation of words by considering the text features. It is based on either predicting the word by using the context (CBOW) or predicting the context by using the word (skip-gram) [12]. As a result, Word2vec gathers vectors that are related to similar words in a vector space. One of the most widely used pre-trained models is Google’s Word2Vec model. It has word vectors for around 3 million words that are trained using 100 billion words [12]. In fact, some frameworks, such as in [15], have used the advantage of the Word2vec mechanism to determine the initial similarities between the entered query and a set of documents. Moreover, the GloVe model [16], which is a model for the unsupervised learning of word representations, used Word2vec for training and outperformed some models on word analogy, word similarity, and named entity recognition tasks.
3.3 Terms selection

In this phase, the candidate terms are selected. The selection is performed based on the relationship between the original query terms and the candidate terms. The similarity between them is the key to finalizing the relevant candidate terms list. To measure the closeness of the terms with the original query terms, a similarity measurement method is used that represents the next phase in dealing with the word vectors as illustrated in Figure 3. In the ontology-based approach, for example, the depth of the term in the ontology is used as a measurement criterion. In the linguistic structure approach, methods such as Cosine similarity, Jaccard similarity, Euclidean distance, and Latent Semantic Analysis can be used. After defining the similarity weight of each candidate term, a threshold is selected to choose the most relevant ones. In the following, we briefly discuss the main similarity measures shown in Figure 6.

**Cosine similarity.** Regardless to their size, cosine similarity shows the similarity between two documents based
on the cosine angle between their vectors in the vector space. The similarity range is from 1 to -1, where 1 means the documents are very similar and almost identical. The narrower the angle, the more similarity between the documents [17].

**Jaccard similarity.** Jaccard similarity measures the similarity between two vectors using one feature at a time by taking the intersection of both and dividing it by their union, as shown in the following formula.

\[
J(A,B) = \frac{A \cap B}{A \cup B}
\]

Indeed, it computes the similarity between the features in both vectors. The intersection here represents the number of times the feature equals 1 in both vectors. The union is the addition of three parts. The first part is the total number of times the first vector value of that feature equals 1 while the other one equals 0. The second part is the total number of times the first vector value of that feature equals 0 while the other one equals 1. The third part is their intersection [18, 19].

**Euclidean distance.** Euclidean distance represents the shortest path between two vectors. It is the true straight line distance between two vectors. It measures how close two word vectors are in the vector space [20].

**Singular Value Decomposition (SVD).** Latent semantic analysis (LSA) learns latent topics by performing a matrix decomposition on the document-term matrix using singular value decomposition (SVD). LSA is typically used as a dimension reduction or noise-reducing technique. Mainly, it aims to reduce the dimensions of a document by using the topics that present the underlying information in the document [21]. It decomposes the m*n documents/terms matrix, where m represents the number of matrix dimensions, into three matrices. The first one is the m*n singular matrix, which represents the terms/topics matrix that assigns the terms with their topics. The second one is the n*n diagonal matrix, where n represents the number of the matrix dimensions, which represents the topics/topics matrix that determines the importance of the topic. The third one is the n*m singular matrix, which represents the topics/documents matrix that shows the distribution of the topics across the documents. After that, it rearranges the matrices based on the descending order of the topics/topics matrix. Finally, it chooses the topics with the highest importance values [9].

### 3.4 Terms Ranking

To show the search results to the users, they must be ordered by how close they are to the user's original query. Since each of the selected candidate terms has a similarity value, the results will be ordered based on that value. Thus, the terms will be ranked based on how similar they are to the user query. To manage the closeness of the terms to the original query terms, a threshold could be established that is used to define the minimum value of the needed similarity score. The selection of the threshold should not be random. Since it has a direct effect on the results, it should be chosen thoughtfully. Indeed, several experiments could be conducted to determine the best threshold value.

### 3.5 Query Reformulation

After manipulating the query and exploring the related terms, the selected ones will be used to reformulate the query by replacing the original query terms with the new expanded ones. The replacement may concern one term at a time or multiple terms, depending on the application. Consequently, the documents will be returned based on the newly expanded queries.

### 3.6 Evaluation

Retrieved documents plays the main role in evaluating QE framework. This is could be measured by examining the portion of retrieved documents that are actually relevant. Recall and precision are the metrics used to test IRS [22], where they can measured the desired improvement. Recall metric is used if the tests include searching in s specific database, for instance TREC⁴ datasets. It measures the ability of the framework to retrieve all relevant documents in the database, see equation 2. On the other hand, precision metric measures the relevance among retrieved documents form limitless database, such as: the web, see equation 3. Another way of precision metric have been used, which is

---

⁴TREC, Text Retrieval Conference: https://trec.nist.gov/
mean average precision (MAP). Generally, average precision is used with ranked retrieval results, see equation 4. It is considered if more relevant documents are required to be returned earlier [23]. MAP is the mean value of average precision for multiple runs of the framework. Maximising precision can be accomplished by using precise terms in the expansion process, while using general terms can maximize the recall [2].

\[
\text{Recall} = \frac{\text{Total number of documents retrieved that are relevant}}{\text{Total number of relevant documents in the database}} \tag{2}
\]

\[
\text{Precision} = \frac{\text{Total number of documents retrieved that are relevant}}{\text{Total number of documents that are retrieved}} \tag{3}
\]

\[
\text{Average precision} = \frac{\sum_{n} \text{precision of the top-}n \text{-retrieved documents}}{\text{Total number of relevant documents}} \tag{4}
\]

Furthermore, to evaluate QE framework, a baseline model should be chosen to use its results for comparison. The baseline model could be a state-of-art model, pseudo-relevance feedback model, Rocchio model, or (no expansion) model. The later one is running the framework with and without expansion and compare their results [5].

It is worth mention that user evaluation could be considered in the evaluation process to seek his satisfaction. However, it can increase the computation time since it can add turn-around time along with its additional computations.

### 4 Query expansion categories

Query expansion categories are organized based on the method they use to extract the related terms. The main categories are: local-source-based, global-source-based, and knowledge-based (semantic-based), as illustrated in Figure 7.

#### 4.1 Local-source based category

This approach depends completely on user feedback. For queries that had been discovered before, the feedback on the relevance of the retrieved documents is the main factor involved in choosing the terms. Pseudo-Relavence Feedback (PRF) is a feedback approach that chooses related terms from only top k relevant documents [1, 24].

Furthermore, Query Logs is also a local-source-based approach. It takes advantage of the interactions between the user and IRS and maps the users’ logs with the new query [9]. This approach is simple and effective in some cases. However, it depends on the user’s past behavior. Hence, it is hard to perform an automated solution with the integration of this approach.
4.2 Global-source-based category

Corpus such as WordNet [25] play the main role in this category. The terms are extracted based on statistical probabilities of the terms that exist on the corpus, which is done by establishing a language model. All the retrieved documents are included in the extraction phase. Candidate terms are chosen based on their relationship with the corpus terms [1]. Unfortunately, this category increases the generality of the terms. Indeed, since it uses an external corpus, the comparison between the query terms and the candidate terms will be based on their co-occurrence or other statistical computations. Therefore, it will not be guaranteed that the extracted terms and truly related, based on the meaning, to the query terms. In addition, due to its use of all retrieved documents, it is computationally expensive [9]. Both the local-source-based category and global-source-based category are useful paths to take in expanding a query. Nevertheless, they can cause query drifting due to the absence of the semantic linking between the terms. Using semantic consideration in finding the new terms can minimize the candidate terms. Therefore, there will be a better chance to choose the optimal ones.

4.3 Knowledge-based category (semantic expansion)

Semantic query expansion is part of the new generation of query expansion techniques. It improves the reformulation of the query by making the query more intelligent. This category is a key to a semantic web where all the information is linked. It structures the query in such a way that allows it to better understand the meaning of the query. This category relies on the knowledge structure. This structure could be a linguistic structure or ontology. Consequently, this category can be organized into approaches based on the knowledge structure being used. Therefore, semantic query expansion has three approaches: the linguistic structure approach, ontology approach, and hybrid approach.

4.3.1 Linguistic structure approach

In this structure, predefined sources are used, such as dictionaries, lexicons, and thesauruses. It depends on applying linguistic relationship methods between the query initials and terms in the structure. Mainly, it depends on weighting the original and expansion concepts to capture the search goal of a query [26]. Part of speech (POS), stemming, and BOW are a few examples of techniques that have been used with this approach. They can reduce the semantic ambiguity of the query terms [2]. This can be achieved with the help of the chosen predefined source to determine the semantic distance between the terms. The linguistic structure approach is flexible and does not require any computations. Also, it uses available sources. However, it is hard to build a domain specification with this structure.

Below are some of the well-known sources that can be used to find the candidate terms.

WordNet. WordNet is one of the most trustworthy and reliable resources for researchers in linguistics analysis. Many psycho-linguistic concepts, which focus on the interrelation between linguistic factors and psychological concepts, and computational theories are applied in its design. It contains different forms of words, including nouns, verbs, adjectives, and adverbs. The key design of WordNet is the synonym sets and relationships between them. A synonym refers to constructing a set of synonymous words that stands for a single word sense. Also, wordNet includes the definition and example of each word sense [25].

Wikipedia. Wikipedia is a very popular website that is based on a collaborative online encyclopedia. It is available in 285 languages and has more than 55 million articles. Each article explains a particular word, subject, or even phrase. In addition, each article includes links to related topics within Wikipedia.

4.3.2 Ontology approach

The ontological structure is a well-defined representation of lexical terms. Ontology is considered one of the conceptual aspects of the semantic web. By using specific tools such as the Resource Description Framework (RDF), an ontology can construct a semantic map between the concept terms. It defines their meaning and the actual relationships among them. Ontologies can be built for a specific domain or for multiple domains. Consequently, if the ontology has been structured in a careful manner, it can serve NLP applications [8]. However, the integration of ontologies is a difficult task and requires complex computations. SPARQL is used as a query language to manipulate data that are stored in RDF format. DBpedia and CYC represent examples of a well-known ontology. The ontological structure is a well-defined representation of lexical terms. Ontology is considered one of the conceptual aspects of the semantic

5Wikipedia, the free encyclopedia: https://en.wikipedia.org/
Approach | Computationally complex | Source Availability | Source Integration | Support domain specific search
--- | --- | --- | --- | ---
Linguistic structure Approach | No | Yes | Easy | No
Ontology Approach | No | Only domain specific | Difficult | Yes
Hybrid Approach | Yes | Only domain specific | Difficult | Yes

Table 3: Semantic expansion approaches.

web. By using specific tools such as RDF, an ontology can construct a semantic map between the concept terms. It defines their meaning and the actual relationships among them. Ontologies do not work well when there is “vagueness” in information [27]. They require certain and clear information to be stored, scored and used.

**DBpedia**. DBpedia is a crowd-sourced community that works to create structured content out of the information available in Wikimedia projects. It stores knowledge in a graph representation in a machine-readable form and represents a way for information to be collected.

**CYC**. CYC is an inference engine developed by Cycrop. It is a knowledge base that contains a formal inference tree of fundamental human knowledge including facts, rules of thumb, and heuristic reasoning about everyday life events [7].

**CRISP**. Computer Retrieval of Information on Scientific Projects (CRISP) is a biomedical ontology which contains research projects conducted at universities, hospitals, and other research institutions [9].

**ConceptNet**. ConceptNet is an ontology that contains 300000 words and constructed semantic relationships among them. It is built by using well-known semantic source such as DBpedia and OpenCyc. [27].

### 4.3.3 Hybrid approach

This approach merges the advantages of both linguistic structure and ontology to enhance query expansion. In this approach, two levels of discovery are considered. In the first level, the terms will be extracted from the linguistic structure. Then, the resulting terms will be fetched with the ontology to explore its parent and children. This approach can produce terms that are strongly related. However, it requires high computation, and there is a possible chance for query drifting to occur.

Table 3 illustrates a brief comparison between these three structures based on their complexity, whether they are computationally expensive or not, and source aspects such as their availability, integration ability and whether they support domain specification.

Based on the above semantic expansion approaches, all the frameworks have been categorized. Figure 8 presents a proportional distribution of the frameworks based on the semantic approach they use. In the following three sections, we examine the frameworks based on the approach they use.

## 5 Linguistic-Based Approaches

In this type of approach, the extraction of the related terms is based on the linguistic structure of the query. It relies on predefined linguistic resources (i.e. dictionaries, lexicons and thesauruses) such as Wordnet. The relationships between the query terms and their relevant terms are computed using a suitable NLP algorithm. In this section, we introduce, in historically ascending order, some recent frameworks that are based on use of the linguistic approach to expand the query.

---

6[https://wiki.dbpedia.org/](https://wiki.dbpedia.org/)
7[https://www.cyc.com](https://www.cyc.com)
8[https://biportal.bioontology.org/ontologies/CRISP](https://biportal.bioontology.org/ontologies/CRISP)
9[https://www.nber.org/research/data/computer-retrieval-information-scientific-project](https://www.nber.org/research/data/computer-retrieval-information-scientific-project)
Meili et al. [28] proposed a method to help software maintainers perform updates on projects. Their work mainly focuses on expanding the query to improve the search for code. The main source used is WordNet [25]. The authors’ basic methodology is to apply the POS with each word in the query. Then, all the resulting tags will be matched with their related synonyms in WordNet [25]. Moreover, the methods identifiers are collected to be used in the query matching phase. Then, the results will be chosen based on their similarity with the natural language phrases that exist in these identifiers. The similarity score is calculated by the total number of common words between the expanded query and the identifier divided by the total number of words on both. The candidate selection criteria depend on how far the tag is from the threshold set by the authors. To test their approach’s effectiveness, they performed two main evaluation stages. First, they compared the result of the expansion method versus the original one. The main comparison was based on the number of relevant and irrelevant suggested terms. The results show an improvement in precision and recall with 40% and 24%, respectively. Second, they compared their approach with the result of Conqure [29]. Their approach outperforms its precision by 5% and its recall by 8%. The approach is simple yet showed improvement in the code query’s effectiveness due to the consideration of methods identifiers. However, the performance of the approach depends indirectly on the quality of the codes. Hence, using a specific domain ontology would be more useful and comprehensive.

Pawan et al. [10] provided a semantic expansion technique to overcome the delay in searching for patents by using patent abstracts as a query and the Inter-national Patent Classification (IPC) as metadata to reduce the search time. Their method is based on three main stages. Firstly, all the relevant abstracts from the IPC are extracted. The extraction is performed by calculating the Term Frequency – Inverse Document Frequency (TF-IDF) feature. Secondly, the expansion phase occurs on two levels: extracting relevant words from WordNet and extracting the reliable page titles from Wikipedia. With WordNet, a vector is established. It contains weights ranging between 0 and 1 for each word. The closer the new word is to the original word, the higher weight it holds. After that, a vector of incoming and outgoing links is calculated using the Page Rank algorithm on Wikipedia. Each page has from 10 up to 100 links, which means there is limited expansion. This was determined to avoid failure. Finally, the final semantic similarity will be based on the expanded vectors. Its measurement is conducted by using cosine similarity and Jaccard coefficient. To test their work, the authors performed experiments with and without the suggested expansion. The precision and recall for the former showed better results. The integration of using different resources is a sufficient improvement. Moreover, the usage of IPC can provide more accurate and faster results. However, in the semantic similarity calculation, the Jaccard coefficient method showed poor performance. Also, they did not clarify a threshold to accept the term based on its semantic similarity score.

Xiang et al. [30] introduce a real-time personalized Twitter search method. The main idea of their work is to consider users’ interests and preferences and semantic features. The approach goes into four steps: feature extraction, feature representation, candidate generation and ranking. In the feature extraction phase, a tweet stream is obtained. After that, the tweets will be preprocessed to get rid of unwanted ones. Also, the tweets will be filtered to eliminate redundant ones. Then, for each tweet, nouns and verbs will be extracted as semantic features. In addition, the poster
information, hashtags, URL and number of comments will be extracted as social attributes and will be used to test the
tweet quality. For the query, it will be expanded by using the TF-IDF algorithm. In the feature representation phase,
every word will be represented as a vector using word2vec. In the candidate generation phase, two main factors are
considered to determine the candidate list: the result of the rule filter of the tweet and its quality. The rule filter is a
Boolean logic keyword model that uses a threshold of the TF-IDF value to specify if the tweet is relevant or not. For
testing the tweet quality, social attributes will be used to train the quality model; then, the tweets will be tested by the
model. Then, the semantic score and the quality score will be merged for every tweet. After that, in the ranking phase,
a threshold adjustment method is used. It works dynamically based on recent historical data of tweets. Based on that,
only the most relevant and qualified tweets are returned. The framework was tested using 10 days of a Twitter sampling
stream. It was compared with other similar frameworks. It had the best score in many queries, and its performance is
similar to the optimal one. The approach introduced a consideration of the tweet’s quality with respect to its relevance.
However, the threshold adjustment method is based on making random assumptions on the tweets’ distribution time.

Bihal et al. [9] tried to eliminate the ambiguity of a query. They took the environment of the query into consider-
ation by building the context around it with the help of the language models (LM) and applying the LSA method.
Their work was based on two main phases. The first phase is the query recommendation phase. This step is based on
the user’s past searches. By using LM, the past queries are measured and ordered by their correlations with the new
query. It relies on the presence of a term in a query and the presence of a document among the clicked documents.
The ordered list of past queries will then be handled by the next phase. The second phase is the LSA method for
query expansion. Basically, it considers the conceptual meaning of the words. It constructs a word-document matrix
to determine the word usage among documents by using singular value decomposition (SVD). SVD assigns a vector
to each term. It focuses on the occurrence of the terms in similar documents and queries, even if they never co-occur in
the same query. After that, the vectors will be ranked using the similarity score for each query term by using the Cosine
similarity measure. To validate their work, the authors used the CISI text database from SRT. They used the original
queries results as a baseline. Their system improves the precision by 24% and F-measure by 7.76%. The method took
advantage of the user’s query log in order to define the semanticity based on the user’s preferences. However, the log
size should be considered since it could affect the time efficiency.

Kyle et al. [31] used a states-based approach to generate the candidate terms. It is based on providing n sub-state
framework where the surviving related candidate terms for the current state will be moved to the next state to extract
more related ones. First, it constructs the root query, which is generated to be treated as the basic reference throughout
the states. After preprocessing of the query, if its length is 3 tokens, then a connection is made with Wikipedia pages
corresponding to it. Otherwise, the N-grams feature is applied to longer queries. For each token, the term frequencies
are stored. Then, each term will be passed to six data utilization modules to gather the corresponding weight of the
terms. The six data utilization modules are developed linguistic modules. Consequently, a stem query is generated to
be passed to the next state. The stem query is selected based on the terms with the best weights from the last step.
To prevent query drift, the original query is passed to the next state along with the stem query. The final selected
terms are the output of the final state. To test the approach, they conducted a comparison experiment with five existing
algorithms. 50 different web topics were used. Their approach had an overall mean average precision of 80%, which
outperformed the other algorithms by 27%. Overall, the approach highlights the importance of using states to examine
the term importance. Furthermore, query drift was prevented by appending the original query in each state. However,
time-complexity-wise, the algorithm is based on revisiting Wikipedia, which could be computationally expensive.

Quynh et al. proposed a method for retrieving images by using an image as a query. It relies on the user to
determine the semanticity of the images [32]. In addition, it differentiates the feature by weighting their semantic
importance to gather more accurate results. The main idea of the proposed algorithm is transferring the user image
query into a multi-point query. Each query point will be considered as a cluster in the image database for further search.
After the user sends their image for a search, the query image will initially be changed to four representations, which
are multiple versions of the image with different color layers. Each representation is a cluster. The feature vectors
will be extracted for each of them. Then, the distance between multi-point query and the images in the database will
be calculated and assigned based on the minimum distance between the image, the feature importance weight and
the query points. The feature importance weight is used to increase the semanticity of the image. Indeed, the feature
importance is calculated as the inverse of the variance of the feature in each axis in the multi-feature space. This
distance is combined with the semantic weight of each query point, which is calculated using the number of related
images in each cluster. Based on the user selection, the selected images will be the clusters’ new centroids, and the
above algorithm will be repeated. If the user is satisfied, then the algorithm stops. Otherwise, the whole procedure
will be repeated with the user intervention until they are satisfied. The method was tested with a database with 34
categories with a total of 3400 images. It produced a higher precision compared to the other four methods. This method takes advantage of the user’s judgment to determine the semanticity. Also, it brings attention to the feature importance and how it can be used to get accurate results. However, when calculating the semantic weight of each query point, the system relies on a traditional content-based retrieval algorithm. Indeed, building the clusters around the centroids lacks any definition of real semanticity between them; this can affect the actual semanticity since the images relatedness is not clearly identified.

In the work of Jagendra et al. [22], they focused more on the ranking phase of the expansion. Their approach relies on combining multiple techniques in ranking and filtering the results. It follows five stages: building the term pool, ranking, semantic filtering, choosing optimal terms by a genetic algorithm and reweighing the terms. First of all, the term pool is constructed by the top retrieved documents based on Okapi-BM25 as a matching function. After that, four different approaches are used for further expansion of the terms from the pool: Kullback-Leibler divergence, co-occurrence, the binary independence model and Robertson selection value. Then, all resulted terms will be ranked using four different techniques: Borda, Condorcet, Reciprocal and SumScore. They are used for rank combining of the terms. Consequently, the ranked terms will be filtered semantically using Word2Vec similarity measure with the TRECCDS corpus. With a genetic algorithm, only the optimal combination of the terms will be selected, where each gene represents a candidate term and each chromosome represents a combination of candidate terms. Recall is used as a parameter of the fitness function. Also, crossover and mutation operators are used to generate the offspring. In the final stage, the Rocchio algorithm is used to reweight the final expanded terms. The approach was tested stage by stage with other state-of-the-art benchmarks and showed good improvements in terms of recall and precision.

Fang et al. proposed an expansion method that uses the advantages of the semantic and sequential information of the words to build a retrieval system for biomedical documents [33]. Their method is based on the semantic sequence dependency model (SSDM). Initially, they trained a domain-specific corpus by using a subset of the MEDLINE database. The Skip-gram model was the language model used to generate the word embedding. When the user enters their query, all the synonyms will be extracted from the corpus. The SSDM will combine the query keywords with the extracted terms for the query expansion. This is done by replacing the query keywords with their related terms with the maximum replacement of three words; thus, the expanded queries will be the result of all possible combinations of the query and its related terms. During the combination phase, a score is given to each combination by using Cosine similarity measurement. Also, during the training phase, each document from the database is given a score as well. Consequently, for the final results, a summed weight of the query and the document scores is issued and used to rank the retrieved results. To test the results, the authors compared their work with a benchmark model and with the conventional sequence dependency model (SDM) by using a thousand questions proposed by experts. They got a mean average precision with a higher value of 0.024. The authors presented a thoughtful approach by studying the impact of different combinations of the expanded queries. Moreover, since it uses a neural network for word embedding, it has the ability to detect common phrases that could be very useful in domain-specific searching. However, the time complexity for long queries should be considered.

Almarwi et al. [34] presented a genetic approach that focuses on the weighting phase. They used Wordnet as their main source for expansion. All the collected synonyms from Wordnet go through three weighting approaches. Then, they considered choosing the optimal weight by applying particle swarm optimization (PSO). The weights were used as positions of particles of the algorithm. The main purpose of applying optimization is to improve the term selection. To test their proposed work, they collected an Arabic corpus and applied multiple experiments to analyze the effectiveness of including and excluding each weights approach. Eventually, they had good recall and precision for their entered queries.

Xiaoyan et al. presented a framework that improves the search for locations based on a semantic approach [20]. It contains two main phases: the location-semantic relationship measuring and the top-k typical and relevant object selection. The first phase depends on three major measurements. The first measurement is used to find the location similarity between two spatial objects. Euclidean distance is used here by considering the latitude and longitude of the objects. The second measurement considers the semantic relevancy between keywords and between the documents of the object’s text. Alchemy API is used to extract distinct keywords; then, the intra and inter correlation between the keywords is constructed by using the Jaccard coefficient and IDF, respectively. Both correlations are combined to define the keyword coupling relationships for each pair of keywords. After that, cosine similarity measure is used to calculate the semantic similarity between documents based on the computed keyword coupling relationships. The third measurement is done by combining the Word2Vec technique and the convolutional neural network. Then, the final semantic relationship between two spatial objects is determined by combining all previous measurements. The second phase defines the top-k relevant objects based on the Gaussian kernel function. To test the approach, many
The question “what is this tweet about?” is viewed as a summery of information. This summery is constructed from documents to avoid any mis-contextualization. The purpose of this work is to find an answer to the tweet as a single query. The main desire to employ semanticity is to retrieve as many relevant documents as possible to avoid any mis-contextualization. Mainly, the purpose of this work is to find an answer of a given tweet by retrieving related documents. To do so, they yield on the semantic query expansion techniques, selection since only the first candidate term is selected, which may cause a loss of other relevant terms.

Ensaf et al. [12] developed a framework that focuses on the Holy Quran. Mainly, it returns all related verses based on the searched concept. A word2vec model is trained on an Arabic corpus based on CBOW. They started by working manually on a Quran dataset and provided each verse with a suitable topic. To produce more useful word vectors, they trained the Word2Vec model using an Arabic corpus that was built using different resources. After that, the entered query will have an average vector along with each topic of the Quran. Then, using the cosine similarity, the first related topic is returned. Consequently, the related verses are collected based on the selected topic. To test the framework, they used three popular datasets. As a baseline, they applied unigram and Okapi methods. They measured the framework performance with precision and recall. The experiment with Omiotis measurement provided the best results for one of the datasets, while the Wikipedia measurement outperformed the other measurements in different datasets. Also, they compared the framework performance with other existing works, and it outperformed them on some datasets. This framework introduced a multilevel semantic measurement; thus, more related documents will be retrieved. In addition, it considers the relatedness between the query terms and the documents before the expansion. However, the threshold of semantic degree for selecting most related terms is not specified.

Masoumeh et al. [36] proposed a hybrid recommender system to generate recommendations in discussion groups by taking advantage of the information about both the content and the user. The method has three main stages: content-based filtering, collaborative filtering and hybrid filtering. Basically, based on the content of the posts and tags, similar posts will be collected. The tags will be extracted, and a semantic hierarchical structure of them will be built by using WordNet as an expansion source. In the collaborative-based filtering phase, the implicit ratings of all users is obtained to find the users most similar to the active users. The users rating will be based on the behavior of the users such as their comments or their favorite posts. Consequently, the similarity between the active user and the users will be determined by using Cosine similarity measure or Pearson correlation coefficient. As a result, the users similar to the active user will be collected. In the content-based filtering phase, the system will then generate a recommendation list of the user’s questions from the hierarchical structure of posted question tags. If the question tags do not exist in the hierarchical structure, a search for their synonyms is conducted using WordNet. As a result, posts similar to the active user’s question will be collected. Finally, in the hybrid stage, the most similar posts with the most interactions from similar users will be returned. The method was compared with other recommendation systems, and it achieved an improvement in precision by 35%. This approach is useful for enhancing the quality of the discussion groups website. Also, by suggesting similar existing posts, it saves on system resources by minimizing the need for saving posts that already exist. However, using similarity measurement between the new tag and its WordNet synonyms can detect more related tags and thus more related posts.

Jamal et al. proposed a framework to improve a method for automatically enhancing the relevant terms by using external information sources [35]. The semantic measurement is calculated for the query and the related documents and for the query and the candidate terms. Firstly, the related documents are extracted based on the probability of the query terms being on them. Then, an automatic relevance feedback method is used to select the most semantically related documents. For the relatedness between the query and the documents, a combination between three measurements is used; these are omiotis, Wikipedia link-based and pointwise mutual information measurements. The first one is based on calculating both the semantic compactness and semantic path elaboration of two words using WordNet. The second measurement is based on the relatedness between terms based on Wikipedia articles. Two terms are related when they are pointed by the same Wikipedia articles. The last measurement measures the co-occurrence of two terms in a large document collection by how often two terms occur together. After that, the top-M documents are selected, and their words and the query terms will be measured with the original query by the above measurement techniques to establish the candidate terms. To test the framework, they used three popular datasets. As a baseline, they applied unigram and Okapi methods. They measured the framework performance with precision and recall. The experiment with Omiotis measurement provided the best results for one of the datasets, while the Wikipedia measurement outperformed the other measurements in different datasets. Also, they compared the framework performance with other existing works, and it outperformed them on some datasets. This framework introduced a multilevel semantic measurement; thus, more related documents will be retrieved. In addition, it considers the relatedness between the query terms and the documents before the expansion. However, the threshold of semantic degree for selecting most related terms is not specified.

In the work of [37], the authors presented a framework for tweet contextualization. It aims to determine the context of a given tweet by retrieving related documents. To do so, they yield on the semantic query expansion techniques, where they treated the tweet as a single query. The main desire to employ semanticity is to retrieve as many relevant documents as possible to avoid any mis-contextualization. Mainly, the purpose of this work is to find an answer of the question “what is this tweet about?” by viewing a summery of information. This summery is constructed from
Table 4: Comparison of the linguistic structure approach frameworks.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Year</th>
<th>Feature Extraction approach</th>
<th>Terms similarity measurement</th>
<th>Expansion Source</th>
<th>Search dataset</th>
<th>Language</th>
<th>Evaluation metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>[28]</td>
<td>2015</td>
<td>POS</td>
<td>The similarity score is the total number of common words between the expanded query and the identifier divided by the total number of words.</td>
<td>WordNet</td>
<td>Codes</td>
<td>English</td>
<td>83% 91%</td>
</tr>
<tr>
<td>[32]</td>
<td>2016</td>
<td>K-Mean Clustering</td>
<td>The ratio of the number of semantically related images in a cluster and the total number of related images of n semantic clusters combined with distance from the image to the query.</td>
<td>Image Database</td>
<td>Image</td>
<td>English</td>
<td>60%</td>
</tr>
<tr>
<td>[9]</td>
<td>2017</td>
<td>LM</td>
<td>SVD, Cosine similarity measure</td>
<td>Query logs, top clicked documents</td>
<td>Text</td>
<td>English</td>
<td>24% (short queries 40.54% (long queries))</td>
</tr>
<tr>
<td>[30]</td>
<td>2017</td>
<td>TF-IDF, Word2vec</td>
<td>Based on TF-IDF values.</td>
<td>Wikipedia</td>
<td>Twitter</td>
<td>English</td>
<td>nCG 33.94%</td>
</tr>
<tr>
<td>[31]</td>
<td>2017</td>
<td>Term Frequency</td>
<td>Based on the final score, which is the intersection of the Wikipedia links that have the query with the links that have both the query and the candidate term.</td>
<td>Wikipedia</td>
<td>Text</td>
<td>English</td>
<td>80%</td>
</tr>
<tr>
<td>[22]</td>
<td>2017</td>
<td>Word2vec</td>
<td>Word2vec</td>
<td>Dataset</td>
<td>Text</td>
<td>English</td>
<td>23.9% 32.5%</td>
</tr>
<tr>
<td>[33]</td>
<td>2018</td>
<td>Words embeddings</td>
<td>Cosine similarity.</td>
<td>Biomedical corpus: MEDLINE database</td>
<td>Text</td>
<td>English</td>
<td>33.62%</td>
</tr>
<tr>
<td>[35]</td>
<td>2019</td>
<td>Not mentioned</td>
<td>Omioitis measurement, Wikipedia link-based measurement, Pointwise mutual information measurement</td>
<td>WordNet, Wikipedia</td>
<td>Text</td>
<td>English</td>
<td>87.54%</td>
</tr>
<tr>
<td>[20]</td>
<td>2019</td>
<td>Word2vec</td>
<td>Euclidean distance, Jaccard, Cosine similarity</td>
<td>Dataset</td>
<td>Text</td>
<td>English</td>
<td>55%</td>
</tr>
<tr>
<td>[12]</td>
<td>2020</td>
<td>Word2vec: CBOW</td>
<td>Cosine similarity.</td>
<td>Classic Arabic Corpus</td>
<td>Text</td>
<td>Arabic</td>
<td>91.95%</td>
</tr>
<tr>
<td>[36]</td>
<td>2020</td>
<td>Not mentioned</td>
<td>Cosine similarity, Pearson correlation coefficient</td>
<td>WordNet</td>
<td>Text</td>
<td>English</td>
<td>72.5%</td>
</tr>
<tr>
<td>[37]</td>
<td>2021</td>
<td>POS</td>
<td>Occurrence frequency</td>
<td>WordNet</td>
<td>Tweets</td>
<td>English</td>
<td>94.59%</td>
</tr>
</tbody>
</table>

related documents that represent the tweet concept. To collect such documents, a semantic query expansion approach (SemQEx) is represented. It relies on WordNet as an expansion source. The query (tweet) is annotated via TreeTagger, which is NLP tool that applies POS on the query terms. Then, candidate terms are selected from WordNet based on their similarities with the query terms. This is accomplish by considering the synsets and their definitions in WordNet. Indeed, the similarity score is calculated based on the occurrence frequency of the query term in the synsets and its definition. Thus, it is considered as candidate term if the original term appears as it synset or in its definition. To test their work, they applied a dissimilarity score to test the improvement of their work and succeeded in decreasing the score with their approach.

Table 4 presents a summary of the frameworks based on their techniques, sources, datasets, language used and evaluation metrics.

6 Ontology-Based Approach

Ontology is a very powerful structure of the terms. It is built based on the semantic relationship between the terms. Since the semantic relevance is already defined in the ontology, some frameworks have taken advantage of it and used
it to expand the queries. In this section, we introduce, in historically ascending order, some recent frameworks that are based on using the ontology approach for query expansion.

In Li et al. [38], the authors focused on searching in the ontology itself using RDF queries. They tried to find solutions to queries with too few results or empty results. They calculate the similarity between the query and its candidate by considering the similarity of two triples in the ontology and the IDF score of the candidate triple. They developed a relaxation algorithm to produce relaxed queries and ranked their importance based on their semantic similarity with the original query. Their methodology was based on ontology relaxation of RDF triples. This means dealing with the entailment in RDF. They studied the semantic similarity with two factors: semantic overlap-ratio and semantic depth. The former represents the number of mutual elements in two RDF triples. The second depends on the ontologies’ hierarchical depths of the concepts. It is measured by the distance between the hierarchical depths. Also, every RDF is weighted by using TF-IDF. The final similarity degree will be computed by using semantic overlap-ratio, semantic depth and the RDF triple weight. The relaxation method, then, will choose the top-k relaxed queries to be executed based on the total similarity degree and a defined threshold. They tested the method by using a dataset with 100 K distinct triples. Queries with empty answers were used. They compared their work with other published methods and they got the minimal running times. Their algorithm outperforms the other algorithms with higher recall. However, based on their precision results, their ranking methodology needs more development.

Gerard et al. proposed a feedback-based approach using appropriate ontologies for homonyms to retrieve images semantically [39]. The first phase of the system focuses on building the ontology. The ontology is constructed for the homonyms of the search keywords. Multiple ontologies can be added as required. The system will be saving the homonymous list to help the homonym LookUp Directory, which is a HashMap that has a key for multiple values. These keys are considered as indices to the ontology tree. When the user entered the query, after it gets cleaned, the remaining words will be passed to the semantic algorithm. Based on homonym LookUp content, this algorithm is used to choose the related ontologies. After that, they will be classified using the SVM classifier. After the classification, the semantic similarity is measured between the terms and the class labels of the ontologies. An OntoEntity is then formulated by adding the matching class label, the matching homonym LookUp content and the similarity score. Further, the similarity is measured between the OntoEntity and the metadata, i.e. the description of the image. When the user sees the results and clicks on one of them, this will be considered later in the algorithm and used as content information. Thus, the algorithm will work based on the lookup directory along with the content information and construct the semantic similarity by using semantic equivalence matching. The content-based analysis is done by keyword matching between the query original terms and the selected image URL words. Based on that, the classes of ontologies will be prioritized and the most semantically related ones will be displayed. All the semantic measurement is done using pointwise mutual information (PMI) measurement. The proposed framework was tested and compared with four other similar works. It achieved 95.33%, 96.41% and 95.87% on precision, accuracy and F-measure, respectively. The framework constructed a separate phase for building an ontology repository, which makes updating it much easier. It incorporates the user intervention in building the ontology, which may be misleading if the user enters inconsistent data into the ontology. It also incorporates the user intervention in the semantic measurement but without relying on it to determine the semantic score.

Faiza et al. [8] proposed a Quran search engine. They built a Quranic ontology. The Holy Quran has special characteristics; this means that the words used inside the Quran may have different meanings compared with when they are used outside of it. The authors constructed and built the Quranic ontology with respect to its linguistic and semantic concepts. Relationships have been established between the key concepts in the Quran e.g. relation of synonym, relation of antonym, relation of hyponymy and relation of similarity. Their search engine works through several steps. After applying lemmatization for each word, all candidate terms are collected from the Quranic ontology, i.e. all child nodes of the term in the ontology will be collected. Then, the system will collect the terms that are semantically related to the original query terms. Two Quranic search engines were used to evaluate their approach: the Alfanous Quran search engine and Corpus Quran. The approach system gives 70% average precision, which is the highest precision among other search engines. Indeed, using recall as an evaluation metric is required since the system uses a limited dataset to search, which is the Quran. The approach’s strength is the terms expansion phase. However, it does not provide a ranking mechanism since it expands the query with various terms, which can result in a massive number of verses.

Ahmed et al. [40] worked on taking advantage of existing semantic web repositories as well as building a bridge between non-English speakers and English resources. This was done by implementing a natural language interface (NLI) to reach out English ontologies. This interface will match the entered query with the most relevant semantic web query written in SPARQL, which could be used to get RDF results. The SPARQL query is created based on the expanded queries of the entered query. Indeed, the authors’ architecture is built in four main stages. First, the
entered query will be preprocessed to eliminate any useful words and then get translated to English. After that, in the query mapping phase, the query is classified to create the correct SPARQL query. After that, if any measurement is listed in the query, it will be identified that it should use the SPARQL query. Next, semantic mapping is performed between the query terms and ontology name entities to identify the most relevant ones to the ontology topics. Finally, based on the relationship between the terms in the ontology, the relationship between the terms will be identified. As a final step, the SPARQL query will be generated based on all the expanded queries terms and be able to be executed to gain the related RDF results. To test their work, they built an integrated ontology from food and health ontologies. They compare the performance of the approach with a manual approach. 389 questions were used. First, their concepts were identified manually. After that, the same experiment was conducted using the proposed approach. Their approach’s precision, recall and F-Measure are 89%, 79% and 84%, respectively. The proposed framework could capture more semanticy if the original terms were expanded before matching them with the entity names components. In other words, if the query terms do not match any topic from the entity names, the SPARQL query could be empty.

Furthermore, alsamadi et al. [41] tried to build a bridge between Arabic questions and DBpedia. They proposed a framework to create SPARQL queries from the entered question. For preprocessing, they performed tokanization, stemming and named entity recognition. The latter is the most important since it could identify the names out of the question. To perform that, they used a java library called FARASA. After that, they collected the resource identification and labels, DBpedia in their case. Then, they acquired the properties from the question, which connect the subject with the objects. Since, DBpedia does not have Arabic labels for most of its represented properties, they performed the properties collection with different approaches: the baseline approach, Wikidata-based approach, and dependency parsing-based approach. The dependency parsing-based approach generated the best results for their framework. After that, they classified the triples to subject-based triple and object-based triple and formulated the SPARQL query based on that. After evaluating their work, they got 84% precision.

In [21], the authors tried to take advantage of DBpedia attributes values by choosing the most appropriate ones. To achieve that, they used an integration of topic modeling to find how topics and attributes are semantically related. Their methodology has two main phases. In the first phase, the relationship between the query and the document is determined by a language model. Then, Bo1 is used to assign the probability distribution in order to select the top Bo1 expansion terms from the returned documents. In the second phase, for each Bo1 expansion terms, DBpedia attributes are selected, whether it is single or multi valued. After that, Latent Dirichet Allocation (LDA) is applied to Bo1 expansion terms from DBpedia by using two attributes to collect the final candidate expansion terms. LDA, indeed, is used to find the distribution of the topic in documents and the distribution of words of the document to topics. Their experiments showed that LDA can solve the problem of multi-valued attributes of DBpedia, which means increasing the efficiency of the selected terms.

The work of [27] considered working with fuzzy ontology to show its impact on the query expansion process. To perform this, their framework has three main phases, fuzzy ontology construction, query expansion using the fuzzy ontology and retrieve the needed information based on the expanded query. First, in order to build the fuzzy ontology, a dictionary is created by applying text mining on a certain corpus with the help of external ontologies. The main purpose of the text mining is to extract the main concepts on the selected corpus. Then, ConceptNet ontology is used to define the relationships of the concepts whether it is synonymous, hierarchical or functional relationships. Then, based on the type of the relationship, a semantic weight is assigned for each pair of the concepts to complete the creation of the fuzzy ontology. After that, the fuzzy ontology is used to expand the user’s query. This is performed by selecting the top three related concepts based on their weights. Then, the information retrieval process is performed using the expanded query. To test their work, they constructed a fuzzy ontology using Solar domain using a database from the UCI Repository. Then, the expanded query is tested throughout the web search engines. Their framework reached 80% precision among Yahoo, Google and Bing.

Table 5 presents a summary of the frameworks based on their techniques, sources, datasets, language used and evaluation metrics.

7 Hybrid-Based Approach

In this approach, the framework produces strongly related terms. It explores the linguistic structure of the term, and it takes the advantage of the strong semantic relationship between the terms that are already built in the ontology. In this section, we introduce, in historically ascending order, some recent frameworks that are based on using the hybrid approach to expand the query.
<table>
<thead>
<tr>
<th>Ref.</th>
<th>Year</th>
<th>Feature Extraction approach</th>
<th>Terms similarity measurement</th>
<th>Expansion Source</th>
<th>Search dataset</th>
<th>Language</th>
<th>Evaluation metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[40]</td>
<td>2016</td>
<td>Not mentioned</td>
<td>Based on matching the query terms and name entities in the ontology</td>
<td>Ontology</td>
<td>Text</td>
<td>English</td>
<td>89% 79%</td>
</tr>
<tr>
<td>[38]</td>
<td>2017</td>
<td>TF-IDF</td>
<td>Based on semantic overlap ratio and semantic depth of two triples</td>
<td>Ontology</td>
<td>RDF triples</td>
<td>English</td>
<td>90%</td>
</tr>
<tr>
<td>[39]</td>
<td>2018</td>
<td>Lemmatization</td>
<td>PMI: Pointwise Mutual Information</td>
<td>Ontology</td>
<td>Images</td>
<td>English</td>
<td>95.33%</td>
</tr>
<tr>
<td>[8]</td>
<td>2019</td>
<td>Lemmatization</td>
<td>Collect all child nodes of the term in the ontology.</td>
<td>Quranic ontology</td>
<td>Quran verses</td>
<td>Arabic</td>
<td>70%</td>
</tr>
<tr>
<td>[21]</td>
<td>2021</td>
<td>LDA</td>
<td>A combination of LDA and Bo1 is applied</td>
<td>DBpedia</td>
<td>Text</td>
<td>English</td>
<td>45% 68%</td>
</tr>
<tr>
<td>[27]</td>
<td>2021</td>
<td>Text mining</td>
<td>A fuzzy membership weight using ConceptNet</td>
<td>ConceptNet and external ontologies</td>
<td>Text</td>
<td>English</td>
<td>80%</td>
</tr>
</tbody>
</table>

Table 5: Comparison of ontology approach frameworks.

Yuanfeng et al. [7] worked on an image retrieval system that is based mainly on the CYC knowledge base. Their method is constructed through three phases. First, an interface is provided to the user to pass the initial query. Second, in the query expansion candidates generator, using the CYC knowledge base, all related terms are extracted. This is done by digging further into CYC ontologies to find related concepts and passing them to the query expansion candidates ranking system as the expansion candidates array. Third, in the query expansion candidates ranking system, all the array elements’ semantic similarities will be measured. The semantic similarity is calculated based on the WordNet [25] similarity measurements. After that, the most relevant candidates will be returned to the user to expand their options. Then, based on their deeper selection, the final results will pop out. To prove their approach efficiency, the authors used two comparison strategies. First, they tested the system with and without their solution. The results showed that the expansion approach had 27% higher precision than the other one. For the second comparison, they compared the results of their approach with the results of Bing, a well-known search engine. The proposed approach produced more effective candidates. The effectiveness of their approach is 6.4% higher than Bing. The proposed model states a good interaction between the user and the system. It depends on their dynamic feedback rather than their search history. Also, it uses multiple expansion resources to assess the semantic relationship between the query and the candidates. However, it does not consider the image features in the semantic measurement. It focuses only on analyzing the query text and its related concepts.

Meriem et al. [1] tried to overcome the short queries issues on microblog platforms. Mainly, it is based on extracting terms using text mining. Their methodology is based on two main phases. In the first stage, all the candidate terms are generated. In this stage, three steps are used. First, unstructured texts of articles form Wikipedia that are related to the query will be selected. The selection is done by using TF-IDF, and then association rules are applied between query terms and expansion terms to choose the candidate terms. The second step focuses only on the definition of the Wikipedia-related articles. The terms are extracted from the first sentence and the first paragraph of the article. In the third step, SPARQL is used to match related terms of the query with the DBpedia using the description of the concept. In the second stage, the candidate terms will be filtered based on a semantic relatedness measure that includes semantic analysis of Wikipedia and the resulting confidence of the association rules. All the terms that have semantic relatedness measures that are greater than a defined threshold will be selected. To evaluate their work, they performed several experiments on Twitter. The experiments showed good performance when the extraction techniques were combined and both filtering techniques were used.

Giuseppe et al. [17] illustrate a framework to semi-automate semantic graph generation. They structured the system by using a neural network and used SPARQL as a training set. SPARQL is extracted by retrieving all the properties that exist in the queries and stored as URIs. Also, the data of these properties will be retrieved. After that, triples from the queries data are extracted. Then, for each concept among the queries, the label of the highest levels classes of the concept is extracted. Before the clustering phase, the neural language engine takes the SPARQL triples and transforms them to sentence for training. After that, vector representation is generated for SPARQL variables, which will be clustered based on the close proximity using cosine similarity. At this point, by using the GUI, the user can edit...
the results of the clustering. Finally, in the semantic graph mapping phase, the median value of Levenshtein distances is used to assign an attribute in the data source with its close cluster. To test the system, the authors compared the precision of their semantic mapping with the semantic mapping generated by domain experts. The ontology used was DBpedia. The result showed a complete overlap between them. The approach introduced the use of AI in the semantic expansion field. However, to make the framework more intelligent and effective, the user intervention should be less involved and linguistic sources such as WordNet could be useful. In addition, Levenshtein distances focuses on the spelling of the variables; therefore, it does not consider the semantic relationships. Semantic analysis could be performed by stemming the variables in the cluster and then using semantic measurement to accept the new ones.

Later, Shengli et al. [42] presented a framework that holds a similar contribution of [40]. They proposed a semantic query graph to construct the information semantically with the help of a query expansion approach. The main idea behind this is to translate the NL question query (NLQ) to SPARQL. The framework works in four phases. First, in the query dependency parsing, the question structure will be analyzed by using Stanford Parser, an NLP parsing tool. This analysis includes determining the dependency relations of the question words. Then, the dependency parsing tree is generated based on the relationship between the words. After that, the tree structure will go through an optimization process that includes merging words, removing less semantic dependencies and joining dependencies; then, a dependency parsing graph is constructed. Secondly, based on the graph from the previous step, the semantic query graph is constructed with respect to the main entity on the question, the question type and the verb type if the question contains one. After that, using string similarity score, the semantic similarity is calculated by using WordNet to match the semantic query graph with the knowledge base: BDpedia Spotlight. Then, the relation edges in the graph are mapped to predicates and the entity nodes to the attribute value. Finally, the corresponding SPARQL is generated. This is done by traversing the semantic query graph where the edge represents the SPARQL conditional statement. The SPARQL variable will be the variable node, and the resource will be the entity node. Four different datasets were used to test the approach. The results were obtained by using their approach and three other ones. When their approach was used with two datasets, it showed a good performance. However, the approach did not introduce an optimal solution, as it failed in capturing and translating some questions. Using a different knowledge base may produce better performance of the approach since BDpedia stores the resources in a format that may differ from the ones in the question.

In the work of [4], a hybrid query expansion framework was proposed for biomedical field queries. It aimed to overcome the vocabulary mismatch problem which can occur due to the diverse lexical variants in the biomedical field. They used three main sources to enrich the query: CDI (clinical diagnosis information), Wikipedia and Mayo Clinic. Their framework consists of five stages. In the first stage, the main goal is to extract biomedical concepts. This is performed by using MetaMap annotation which extract the desired concepts from the unstructured clinical notes and each one is assigned by an ID called CUI (Concept Unique Identifier). In the second stage, queries are created from the extracted concepts and sent to Google Custom Search Engine (GCSE). Then, GCSE fetches both Wikipedia and Mayo Clinic to collected related results. After that, the original query is expanded further by using pre-trained word embedding model by collecting candidate terms that are semantically related to the original query via cosine similarity measurement. Indeed, they used three types: domain-specific, domain-agnostic and hybrid. Furthermore, the resulted candidate terms and CDI are merged and sent to PubMed corpus, a pre-processed indexed biomedical corpus, to obtain the best combination for the query. Finally, based on the final query, the related biomedical literature is retrieved. Their experiments showed in precision when they used the hybrid word embedding model.

The work of [43] represented a query-by-document framework. It aims to take the seed document as a seed and construct a query string. This query, then, will be sent to an API to fetch for related scientific documents. Their methodology presents a query expansion mechanism with different order and purpose than the previous works we had viewed. First, they used the seed corpus in order to formulate the targeted search domain. To do so, a list of keywords are extracted from that corpus to determine the characteristic of the domain knowledge. These keywords are chosen based on their TF-IDF weights. Then, to fetch for the related documents, a query is constructed from the list of the extracted keywords. To accomplish this, they used Monte Carlo (MC) sampling principle, where the query is generated by choosing candidate keywords from the list that have a probability distribution based on their TF-IDF values. After that, the document frequency is calculated based on its appearance in every MC iteration. Indeed, the document frequency will be used to rank the relevance of the resulted documents. Besides, the semantic rank of the results is calculated based on cosine similarity measurement. To test their work, they used two case studies. In the first one, they targeted two different search field and attached a high recall value, 83.9%. The second case study, they compared the methodology with different sampling techniques and concluded that MC reached the highest linguistic relevance results.
Table 6: Comparison of hybrid approach frameworks.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Year</th>
<th>Feature Extraction approach</th>
<th>Terms similarity measurement</th>
<th>Expansion Source</th>
<th>Search dataset</th>
<th>Language</th>
<th>Evaluation metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>[17]</td>
<td>2018</td>
<td>Word2Vec</td>
<td>Levenshtein distances, Cosine similarity</td>
<td>DBpedia, RDF triples</td>
<td>Text</td>
<td>English</td>
<td>P@5: 60%</td>
</tr>
<tr>
<td>[42]</td>
<td>2019</td>
<td>Not mentioned</td>
<td>Based on the length of the greatest common subsequence over the length of the word.</td>
<td>DBpedia, WordNet</td>
<td>Text</td>
<td>English</td>
<td>Precision: 94%</td>
</tr>
<tr>
<td>[4]</td>
<td>2022</td>
<td>Not mentioned</td>
<td>Cosine similarity</td>
<td>clinical diagnosis information, Wikipedia, Mayo Clinic</td>
<td>Text</td>
<td>English</td>
<td>P@5: 48%</td>
</tr>
<tr>
<td>[43]</td>
<td>2022</td>
<td>TF-IDF</td>
<td>Cosine similarity</td>
<td>domain specific corpus</td>
<td>Text</td>
<td>English</td>
<td>83.9%</td>
</tr>
</tbody>
</table>

Table 6 presents a summary of the frameworks based on their techniques, sources, datasets, language used and evaluation metrics.

8 Discussion

This section offers a succinct overview of researchers’ interest in semantic query expansion (QE). It delves into the commonly used approach, similarity measurement, language, sources, and optimization, while also addressing challenges and potential future research directions. To provide a comprehensive understanding of the QE field, we present our observations based on a recent literature review in Tables 4, 5, and 6.

The linguistic structure approach has emerged as a popular choice among researchers, as evidenced by Figure 8. This approach provides several advantages, including enhanced efficiency in capturing semantic similarity. Researchers can apply multiple linguistic techniques, which offers greater flexibility. Furthermore, there are several corpuses available that can be used as a source for semantic expansion. Creating a new corpus using web data for a linguistic model is also a viable option. While some researchers focus solely on ontology, the mixed approach is more commonly used, as it provides a broader scope and greater accuracy in semantic query expansion.

Moreover, after conducting our research, we found that three techniques stood out for their effectiveness: cosine similarity, Word2vec, and TF-IDF. Cosine similarity is a commonly used measure for accurately determining the relatedness between vectors in a vector-space. Instead of measuring the actual distance between vectors, it evaluates how close they are within a specific topic [12]. Word2vec is another effective method that employs a neural network model to handle large sets of text. On the other hand, TF-IDF proves to be useful when assessing the significance of terms based on their frequencies in specific documents.

In order to identify potential future directions in the QE field, we conducted a thorough literature review and compared our findings across various aspects, which we believe represent open issues in the field. These aspects include the use of machine learning, as it can greatly assist in NLP tasks; the language used, which can provide insight into QE’s popularity among different nations; the approach taken, which can reveal which methods are currently dominant in the field; and the use of optimization and SPARQL translation, which can help achieve optimal results. Lastly, we considered the domain and whether the framework can be used for general purpose search. Our findings are summarized in Table 10.

10 Natural Language
<table>
<thead>
<tr>
<th>Ref.</th>
<th>Search Domain</th>
<th>Key concept</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| [28] | Codes                  | • Matching exact POSs pairs of query terms and candidates.  
• Extracting methods identifiers in the code to match them with the candidates. | Taking advantage of the methods identifiers.  
The approach performance based on the code’s quality. Ontology will perform better.                                                                 |                                                                                                  |
| [10] | Patents                | Using the patent abstract as query and International Patent Classification (IPC) as metadata to reduce the search time. | Integrated sources, the use of IPC to get more accurate and fast results.                                                                                                                                   | They did not clarify a threshold to accept the term based on its semantic similarity.           |
| [30] | Tweets                 | • Consideration of the user’s semantic and social features.  
• Defining a boolean logic keyword relevance filter.  
• Defining tweet quality model | Tweet’s quality consideration.                                                                                                                                  | The threshold adjustment method is making random assumption in the tweets’ distribution time.     |
• Applying the LSA method using the query context. | Using the user’s query log to define the semanticity based on the user’s preferences.                                                                                                                      | The log size could affect the time efficiency.                                                    |
| [35] | General                | Applying different levels of semantic relatedness; a document against a query and each query term with terms in the document. | Providing multiple measurement methods.                                                                                                                               | The threshold for the semantic degree is not specified.                                         |
| [32] | General                | Using the user’s selected images to be the centroids of the clusters.     | The consideration of the variance of the feature to determine its importance.                                                                                                                                   | • The phase of building the clusters around the centroids lacks defining the semanticity between them.  
• The main methods are centered around the user intervention, which could be misleading.         |
| [12] | Islamic holy book: Quran | Building and training a word2vec model using CBOW on the Arabic corpus and using it for Quran searching. | • The construction of a vector for the query as a whole by averaging the vectors of its words.  
• Applying the machine learning technique to construct the vectors.                                                                 | Only the first candidate term is selected, which may cause the loss of other relevant terms.         |
| [31] | General                | Providing a sub-state framework where the surviving related candidate terms for the current state will be moved to the next state to extract more related ones. | The persistence of the original query across states to minimize query drift for generated enhancement terms  
The algorithm is based on revisiting Wikipedia, which could be computationally expensive. |                                                                                                  |
| [33] | Biomedical Articles    | Integrating both sequential and semantic information to expand the query. Investigating which word-embedding algorithm can best serve for biomedical articles. | Ability to detect common phrases, which could be very useful in domain-specific searching.                                                                                                              | The time complexity for long queries should be considered.                                      |
| [36] | Discussion groups      | Applying semantic measurement in a recommender system for discussion groups using both content and user information. | By suggesting similar existing posts, it saves the system resources by minimizing the need for saving posts that already exist.  
Similarity measurement between the new tag and its WordNet synonyms can detect more related tags and thus more related posts. |                                                                                                  |
| [20] | Locations              | Using semantic query expansion in spatial keyword query.                  | The consideration of relatedness of among locations and among their textual documents.                                                                                                                          | The system precision is low.                                                                    |
| [22] | General                | Using different similarity measurements to expand the query semantically and using a genetic algorithm to select the optimal candidates. | Multiple ranking and filtering  
The initial term pool should be generated out of semantically related documents.                                                                                                                                  |                                                                                                  |
| [34] | General                | Applying several similarity measurements and applying optimization to select the optimal weight for each candidate | Multiple ranking and filtering  
The initial term pool is selected from Arabic Wordnet, which is a limited source                                                                                                                                   |                                                                                                  |
| [37] | Tweets                 | Reling on semantic query expansion for tweets classification               | Multiple ranking and filtering  
The initial term pool is selected from Arabic Wordnet, which is a limited source                                                                                                                                   |                                                                                                  |

Table 7: Linguistic structure frameworks summary.
<table>
<thead>
<tr>
<th>Ref.</th>
<th>Search Domain</th>
<th>Key concept</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>[8]</td>
<td>General</td>
<td>Building a Quranic ontology based on the semantic relationship between the concept and its child nodes.</td>
<td>Using the topic’s nodes as candidate terms.</td>
<td>It lacks in providing a useful ranking method.</td>
</tr>
<tr>
<td>[39]</td>
<td>General</td>
<td>Constructing a feedback-based approach using appropriate ontologies for homonyms to retrieve images semantically.</td>
<td>• The ability of building an up-to-date ontology.</td>
<td>The ontology construction involves user intervention, which may cause inconsistent ontology data.</td>
</tr>
<tr>
<td>[38]</td>
<td>General</td>
<td>Calculate the similarity between the query and its candidate by considering the similarity of two triples in the ontology and the IDF score of the candidate triple.</td>
<td>• The consideration of triples similarity.</td>
<td>Based on their precision results, their ranking methodology needs more development to capture more relevant results.</td>
</tr>
<tr>
<td>[40]</td>
<td>Health/Food</td>
<td>Addressing the translation of the natural language query to the SPARQL semantic query using multilingual and cross-domain ontologies.</td>
<td>It supports multilingual search.</td>
<td>It lacks in expanding the query terms before matching them with the ontology topics; if the query terms do not match any topic from entity names, the SPARQL query could be empty.</td>
</tr>
<tr>
<td>[21]</td>
<td>General</td>
<td>Combining Bo1 and LDA to determine the best DBpedia attributes to fetch the suitable expansion terms</td>
<td>It showed the effect of different DBpedia features.</td>
<td>Some suitable candidate terms may not appear due to the fixed number of the Bo1 candidate terms</td>
</tr>
<tr>
<td>[27]</td>
<td>Specific</td>
<td>Constructing a fuzzy ontology using external ontologies for a specific domain and use ConceptNet to determine the weights.</td>
<td>It presents an integration of different ontologies which could serve on the source availability challenge.</td>
<td>Its weights measurement depends on other sources weights.</td>
</tr>
</tbody>
</table>

Table 8: Ontology frameworks summary.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Search Domain</th>
<th>Key concept</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Microblogs</td>
<td>Applying a text mining-based approach along with external knowledge sources to expand short queries</td>
<td>The use of association rules results in more accurate related terms</td>
<td>It consumes time since it includes visiting two external sources.</td>
</tr>
<tr>
<td>[7]</td>
<td>General</td>
<td>The usage of CYC as a semantic expansion source with the help of WordNet to find more related images.</td>
<td>The use of multiple expansion resources to assess the semantic relationship between the query and the candidates.</td>
<td>It does not consider the image features in the semantic measurement. It focuses only on analyzing the query text and its related concepts.</td>
</tr>
<tr>
<td>[17]</td>
<td>General</td>
<td>Training a neural language model using SPARQL to reconstruct semantic mapping of a data source.</td>
<td>Providing a semi-automated approach that can be used to build an ontology.</td>
<td>Levenshtein distances focus on the spelling of the variables rather than semantic distance.</td>
</tr>
<tr>
<td>[42]</td>
<td>General</td>
<td>Applying query expansion in the natural question to translate them to SPARQL queries.</td>
<td>It can be useful for end users with less experience in constructing SPARQL queries.</td>
<td>The closeness measurement should be considered more since it does not always produce the optimal answer.</td>
</tr>
<tr>
<td>[4]</td>
<td>Biomedical</td>
<td>Combining external expansion sources with word embedding model</td>
<td>The usage of multiple sources decreases the vocabulary mismatch</td>
<td>The relationship between each term and the whole query should considered since the query represents a particular biomedical case</td>
</tr>
<tr>
<td>[43]</td>
<td>General</td>
<td>The usage of Monte Carlo sampling principle to reformulate the query</td>
<td>It can perform well with less seed papers as corpus</td>
<td>The speed of the sampling process is a challenge. Also, it relies on the targeted API limits of presented results.</td>
</tr>
</tbody>
</table>

Table 9: Hybrid frameworks summary.
Table 10: General comparison of the literature

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Machine Learning</th>
<th>Support Arabic</th>
<th>Linguistic Approaches</th>
<th>Ontology Approach</th>
<th>Hybrid Approach</th>
<th>Optimization</th>
<th>NL to SPARQL</th>
<th>Domain specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>[28]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[10]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[30]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[9]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[35]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[34]</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[12]</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[31]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[33]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[36]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[20]</td>
<td>✓ ✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[22]</td>
<td>✓ ✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[21]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[41]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[39]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[38]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[40]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[27]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[34]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[7]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[42]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[43]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10 reveals that there are relatively few efforts to support multiple languages, particularly Arabic, using a hybrid approach. Moreover, existing studies typically concentrate on specific domains such as the Quran (e.g., [12]) or food (e.g., [40]), which diminishes their relevance for general applications. As a result, more comprehensive frameworks that span multiple domains are necessary.

To effectively manipulate available English ontologies, it is necessary to use English queries. However, this presents a challenge for non-English speakers. One solution is to develop a translation model that can convert Arabic queries to SPARQL queries, as suggested in previous studies like [40, 41]. Machine learning techniques can be used to enhance the accuracy of such models, particularly in word representation and similarity measurements, as shown in [31, 20, 22, 17, 34]. Additionally, using language-specific word representations, such as the Arabic word2vec model presented in [44], can improve semantic detection. By narrowing the gap between languages, these techniques can make ontologies more accessible and useful for a wider range of users.

In the realm of query expansion optimization, there have been some notable efforts presented in literature. For instance, [34] and [22] have proposed methods for optimization. In [34], particle swarm optimization was applied to choose the optimal weight for candidate terms, while in [22], genetic optimization was used to select optimal candidate combinations. While optimization can help reduce query drifting, only a handful of studies have explored this avenue. As such, exploring AI optimization with various algorithms could be a promising research path for improving semantic query expansion. This could be achieved by using semantic measurement techniques, such as cosine measurement or tf-idf, as a fitness score.

8.1 Challenges

Based on an overview of recent works on semantic query expansion, summarized in Tables 7, 8, and 9, various challenges and open issues can be identified and explored. These challenges and issues must be addressed with care as they may significantly impact the system’s performance. It is imperative to carefully analyze and address these challenges to ensure the optimal performance of the system.

Source Availability. The use of existing knowledge structures, such as WordNet, can be limited when it comes to incorporating newly adapted slang phrases. To address this issue, a utility that can match slang phrases with their exact meaning can be helpful. However, the limited number of existing ontologies poses another challenge. Building a new ontology requires time and comprehensive knowledge of the ontology domain, which can be a difficult task. Despite this, researchers can collaborate with domain experts to develop a comprehensive and efficient ontology.
Additionally, most researchers use domain-specific ontologies, but integrating multiple ontologies can increase usage and make it more generic.

Developing a semantic framework for languages other than English is a linguistic challenge. To overcome this, a hybrid framework that uses different weighting techniques can be used to determine the similarity scores of candidate terms, as demonstrated in the Arabic language framework presented in [34]. A translation phase is also necessary to take advantage of the available English sources. However, the translation must be precise to generate accurate terms and prevent query drifting. To address this challenge, [45] proposed a framework that can recognize named entities for Arabic texts using the MADAMIRA tool to capture Arabic features and avoid any loss of important words in the translation process.

**User intervention.** The success of an approach can be influenced by user intervention. Depending on the application, the user’s intervention can explicitly shape the semantic vision of the framework, which may lead to misleading results. For instance, in [32], the algorithm’s desired clusters are determined by the user’s selection of related images. However, there is a possibility that the user’s perception of the image characteristics may differ from how they are presented in the image dataset’s vector-space. This could lead to query drift and affect the approach’s performance. Therefore, it is crucial to consider the necessary level of user intervention in the framework without impeding the expansion process[5].

**User context.** Understanding the user’s tendencies is crucial in providing them with relevant information. Social and personal information can be a factor that helps achieve this goal. For instance, if a user’s personal information indicates an interest in computers, a search for "Python" should yield results related to the programming language rather than the reptile. However, it’s important to carefully consider the need for user context. While social applications like Twitter can benefit from user context [30], it may be useless and time-consuming on other platforms. Therefore, it’s essential to analyze user context and gather it only if it leads to more desirable results.

**Time complexity.** In the field of semantic query expansion, the more complex the computations, the more time they consume. In their paper,[31] proposed a framework that goes through different states, each with its own set of complex computations, making it computationally expensive. Additionally, in[15], multiple types of similarity crosschecking were used, including constructing a concept tree and using average mutual information (AMI) to calculate the similarity between words, which can be time-consuming and impact overall performance.

Therefore, it is crucial to consider time complexity when developing a framework. A reasonable balance must be found between providing accurate results and the time it takes to do so. Users generally expect their results in a timely manner, making this a significant factor in assessing a search application. To improve time complexity over a wider range, solutions such as parallel computing and HPC can be considered. However, questions may arise, such as how to use these solutions effectively, which computations to split in a parallel manner, and whether these solutions will deliver accurate results with less time consumed. For example, in [46], researchers suggested data and query transformations to execute SPARQL queries in parallel. Their main idea is to reduce the number of join operators and increase the number of union operators for the fragmentary steps of the query, allowing for inter-operator parallelism of ontological queries. This enables parallel execution across all partitions.

**Threshold selection.** In frameworks that rely on a specific threshold to select the final expanded terms, the accuracy of the threshold selection can significantly affect the framework’s overall performance. Choosing a large threshold may result in losing important and relevant results, while selecting a small threshold may generate numerous irrelevant results. An adaptive threshold, as defined and used by [30], can enhance the flexibility and feasibility of the searching framework. By conducting a historical study, data analysis, or considering user social behavior, an adaptive threshold can be effectively structured and implemented.

**Scalability.** To cater to the diverse needs of users in a search application, the system’s design should be versatile enough to work with various types of databases. While TREC is a reliable dataset commonly used in this field, relying solely on it may not ensure optimal system performance. Therefore, it’s crucial to test the scalability of the system, particularly if it’s intended for general-domain searching. By doing so, the system’s adaptability and feasibility can be enhanced.

**Query drift.** Query expansion is a valuable technique that can improve the relevance of search results. However, it is not without its challenges, including the risk of query drifting. Query drifting occurs when the expansion process is too broad, causing the search to produce unrelated results. This can be due to issues with term extraction, biased

11TREC, Text Retrieval Conference: https://trec.nist.gov/
term weighting, or biased retrieval models [47]. Low precision and recall are indicators of potential drifting, making it a significant challenge in IR model design. One solution is to set a suitable threshold, which can generate more relevant groups of expanded terms. Additionally, tracking the original query, as demonstrated in [31], can help narrow the focus of the expanded terms to meet user needs. In more advanced frameworks, optimization techniques such as genetic algorithms [22] and particle swarm optimization [34] have been used to select the most relevant terms, thus reducing the risk of drifting. It is critical to consider these challenges carefully when designing an effective IR model.

Security.

Frameworks that leverage user feedback for better results are often based on analysis of user search logs [6]. This approach can effectively detect users’ interests and improve the relevance of results. However, it’s important to consider the potential privacy risks associated with such logs, as they can reveal sensitive user data [48]. To ensure the trustworthiness of retrieval systems, user privacy must be a key consideration. For example, in [49], the authors proposed a framework that uses secure computations to prevent data leaks from the provider’s side.

In addition, protecting the privacy of queried data is equally crucial. User privileges should be clearly defined to prevent unauthorized data access or modifications. To mitigate such risks, the authors of [50] proposed a secure framework based on a blockchain structure. Their approach involves using blocks to store user transaction information and access roles, similar to the blockchain. The hash mechanism is used to securely store data and prevent unauthorized access.

8.2 Open issues and future research directions

The revolution of semantic query expansion has enriched the field of semantic web and IRS in various ways. However, there are still open issues that can pave the way for future research directions and improvements.

- Constructing an ontology can be a time-consuming and demanding task. Therefore, it is important to develop an algorithm that can automate this process while integrating the expertise of ontology developers and experts. For instance, in [39], an approach is proposed that accomplishes this in a more controlled and careful manner.

The availability of ontologies and lexicons is limited in languages other than English, such as Arabic, which can hinder the expansion of candidate queries. To overcome this, building a multilingual ontology or lexicon can improve the query expansion. In [41], an automated question/answer system is introduced that bridges the gap between Arabic questions and linked data by translating natural language questions to SPARQL queries to retrieve answers from linked data, such as Dbpedia. To achieve this, the named entity of the original query is discovered with the help of Wikipedia labels available in Arabic.

In contrast, [51] utilizes the statistical parser of the Arabic Toolkit Service (ATKS) to identify the subject, object, and predicate of the Arabic text, constructing the corresponding RDF triple. To improve the reliability of linguistic resources, slang phrases must also be considered. N-gram and word/sentence embedding can play a significant role in query processing, catching such phrases with the help of reliable sources.

- Applying AI to the expansion field is a challenging move; yet, it could add intelligence to the approaches and generate accurate and useful expanded queries. Deep learning, for instance, could help in recognizing hidden patterns among texts or images, which can provide more semanticity. Furthermore, it can provide great contextual analysis of the text and match it with an accurate topic; thus, the framework can proceed to fetch the closest topic. For instance, [12] represents a framework that applies a machine learning approach, which aims to train a word2vec model using CBOW on an Arabic corpus and use it for searching the Quran. It scored a high precision by increasing the intelligence of the vector structure. In addition, [33] used a trained word2vec model to capture both sequential and semantic information of biomedical texts. Furthermore, in applications that focus on searching in domain-specific areas, machine learning will be useful since the training dataset is precise and determined. For instance, in [52] the authors used a deep learning methodology to construct an answer/question system for building regulations to help engineers in retrieving needed information.

In the realm of semantic query expansion, incorporating real-time data is an area that is yet to be fully explored. However, a real-time semantic search has the potential to provide valuable insights on trending topics, especially during emergencies and major events when people tend to scour social media platforms for information. By leveraging the user’s semantic attributes, a framework described in [30] was able to effectively search Twitter stream data, yielding more accurate and interesting results.
The semantic search for multimedia content, such as images and videos, is still in its nascent stages. Existing frameworks rely on user intervention to determine semanticity, as seen in [32], which can be misleading and fails to provide an actual semantic expansion approach. Incorporating deep semanticity with minimal user judgment can lead to better semantic frameworks. One way to achieve this is by constructing a vector space of the images based on multiple features to create a digital description of the multimedia content. This approach can be used in [53], where the authors aim to extract semantic information from different data forms for image searching. They use a GAN framework to generate a synthetic-related image based on a multimodal query containing both an input image and textual description, which is then used to search for more related images.

9 Conclusion

With the vast amount of information available on the web, finding specific information can prove to be a daunting task. As a result, query expansion has gained popularity in recent years as a means of refining search results. Semantic query expansion, in particular, is an emerging technique that reformulates queries based on related terms that are semantically similar to the original query terms. By improving search engine efficiency, semantic query expansion can bridge the gap between retrieval systems and user needs. Linguistic structure, ontology, and hybrid approaches are the three approaches used for semantic query expansion. Linguistic structure relies on both linguistic characteristics and statistical aspects of the query, ontology uses external ontologies to find related terms, while the hybrid approach combines both methods.

In this paper, a thorough overview of query expansion approaches, particularly semantic ones, is presented. Recent frameworks have been introduced and compared, with most of them utilizing the linguistic structure approach for its efficiency and flexibility. These findings highlight the significance of semantic query expansion in information retrieval and emphasize the importance of constantly exploring new approaches to further improve its effectiveness.

Semantic query expansion is an essential aspect of information retrieval, and building an effective system requires addressing several challenges. These include source availability, time complexity, and setting an appropriate threshold. Involving user intervention and considering the user context is also crucial for enhancing the system’s performance. Additionally, researchers have identified several open issues that need addressing to improve semantic query expansion. These include building new ontologies, leveraging AI methods, considering real-time frameworks, and searching semantically for multimedia content. By exploring these approaches, we can bridge the gap between retrieval systems and user needs, leading to better search results.

References


