VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2022, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 10, Number 3, July-September 2022 pp: 01-09

An Improved Semantic Query Expansion Approach Using Incremental User Tag Profile for Efficient Information Retrieval

Muhammad Ahsan Raza[1], Muhammad Ali[1], Maruf Pasha[1] and Mubashir Ali[2*]

[1]Department of Information Technology, Bahauddin Zakariya University, Multan 60000, Pakistan
[2]Department of Software Engineering, Lahore Garrison University, Lahore 54000, Pakistan
*Corresponding author email: mubashirali@lgu.edu.pk

ABSTRACT

The World Wide Web (WWW) comprises a wide range of information, and it is mainly operated on the principles of keyword matching which often reduces accurate information retrieval. The Keyword matching mechanism faces word mismatch problems while retrieving relevant information. Furthermore, the inherent ambiguity of short keyword queries demands enhanced methods for Web retrieval. Ontological-based query expansion is one of the primary methods for Web retrieval, and it handles the vocabulary mismatch problem by computing semantics from the ontology knowledgebase. However, the retrieval of information relevant to user interests is a major challenge. In this paper, we seek to improve retrieval performance by leveraging user preferences and ontology semantics in the process of query expansion. The expansion words are added to the user query using WordNet lexicon and domain ontology. Additionally, the search intent of the user is also added as expansion words by exploiting a tag-based user profile. When it comes to obtaining relevant documents, the proposed framework outperforms the keyword-based approach by achieving a 76% F1-score. This noticeable improvement accurately reflects the importance of including user intents in the process of semantic query expansion.

KEYWORDS

Natural language processing, Ontology, Semantic query expansion, Entity recognition, COVID-19

JOURNAL INFO

HISTORY: Received: August 15, 2022 Accepted:September 16, 2022 Published: September 21, 2022

I. INTRODUCTION

The World Wide Web (WWW) has become an integral component of modern society because it offers a worldwide place for information with vast publishing, archiving, exchanging, and locating capabilities that can be used to meet the information needs of a person or organization. Because of the exponential growth in both the quantity and complexity of online information, efficient data retrieval is becoming more challenging. An information retrieval (IR) system sorts the documents based on the entered search query, with the most relevant documents appearing at the top of the list.

IR system employs different methods in its effort to understand the underlying semantics and meaning of user search queries and their accompanying textual material. Recent IR systems relied on the Semantic Web, the Social Web, and personalization to improve information search [1]. For instance, one method examines the textual information included in a source of knowledge via statistical analysis or the analysis of natural language. In another method, with the assistance of the Semantic Web, the IR system can make use of the presentation of formal information (namely an ontology) that is machine-understandable. There exist ontologies and taxonomies that may either be generated automatically with the assistance of a computer or manually by a person. Finally, the social Web emerges as a result of user use and interaction. The IR system makes use of the social Web to retrieve information semantically.

Personalized IR systems aim to provide users with more relevant and specific results [2]. To improve the accuracy and efficacy of the personalized IR, it is necessary to have a deep understanding of the user data, securely store it, and conduct a thorough analysis. However, owing to the limited space available for user input and the richness of natural language, search queries tend to be ambiguous, thus making it tough for the IR system to meet the demands of the user [3].

Query Expansion (QE) is the most effective method for addressing the problem of insufficient words, which is common in online search queries. Numerous QE methods (including ontology-based and linguistic) have been presented in recent years to remedy the vocabulary mismatch problem [3-4]. In this paper, we combine Semantic QE with user-centric data (also known as personalized data) for resolving the complexity of short queries. The user information needs are the primary focus of this study. To this end, the proposed semantic QE approach incorporates the user data (namely, user profile) which contains user habits, the user's semantic understating of concepts as well as the search history. The user data and ontology aid the QE process in precise and comprehensive information retrieval.

We provide a personalized search method that incorporates the user's profile, which contains ontological ideas, as well as their search history and habits, the semantic understating of concepts, and the vector-based on synonyms terms gathered from the WordNet API. The second goal is to provide the most appropriate search results based on the information we've gathered about the user from the first goal, making the personalized search as seamless as possible. Finally, by taking the user's profile into account, we may reveal the synonyms and hidden meanings of terms in the user's search query, leading to improved information retrieval.

A. Motivation

It becomes tedious to retrieve user-related data as the amount of data available to researchers grows exponentially. Personalized search engines are becoming increasingly popular as such engines can customize their findings to each user by taking into account their past searches and other information about the person searching. However, these search engines are insufficient to fulfill the needs of a user and understand the semantics of a given query [5].

The COVID-19 outbreak sparks fresh discussion on how to best retrieve data and counteract disinformation. The evolution of the SARS-CoV-2 virus, with its rare combination of lethality and contagiousness, led to the COVID-19 pandemic. Since December 2019, the disease has spread rapidly owing to the present road and air transportation infrastructure and the dense population of major cities, where people are more likely to come into close contact with one another. Numerous publications in several scientific domains have been published as a consequence of research into COVID-19 and SARS-CoV-2 to understand more about the epidemic, keep tabs on its progress, and eventually contain it. Therefore, a more efficient IR method is required that understands the user behavior as well as the hidden meanings of the query.

The primary aim of the present study is to enhance the usefulness of the IR system by expanding search queries with user-relevant terms. To achieve this aim, we have proposed an improved QE approach that combines semantics and user profile in the expansion of query. Moreover, we have carried out experiments using the COVID-19 dataset over the Lucene IR system.

II. LITERATURE REVIEW

The amount of information available on the WWW has grown exponentially, which means that the information is also getting more complicated. Information retrieval is becoming more difficult, time-consuming, and resource-intensive as the volume of available data grows; users also have increased difficulty in separating useful results from the vast quantities of noise they encounter during searches.

The WWW lays the foundation of several contemporary query extension strategies that have been developed to enhance the efficiency of IR Systems. QE strategy employs a wide variety of methods to decipher the underlying semantics and meaning of a search query. Some of the more prominent methods are listed below.

Ojha & Deepak [6] proposed the development of a medical query expansion approach based on metadata. For each query term, new topics are generated using Latent Semantic Indexing, which accepts a query as input and expands it. The approach exploited Lesk similarity with Normalized Pointwise Mutual Information is used to align a collection of ontologies derived from PubMed keywords semantically. With the help of Recurrent Neural Networks, a "Knowledge Tree" is made, which is then used to organize the metadata from Google Books. Lastly, Wikidata, CASNET, and the Hepatitis Knowledge Base are used to add more information to the terms in the Knowledge Tree.

Authors in [7] describe ontology building using domain-specific knowledge. ConceptNet and Global ontology are put to use in the process of building fuzzy ontology. For the many semantic connections that are part of the Global ontology and ConceptNet, a fuzzy membership function has been defined. To better understand a query, most semantically relevant terms are found using a developed fuzzy ontology.

Authors in [8] discussed the shortage of Arabic semantic sources, and suggest a hybrid QE framework for Arabic retrieval. The semantic query expansion uses English ontologies to supplement the restricted Arabic sources (Arabic Wordnet) presently accessible. The framework used a specialized machine learning translation model to assure that the process of Arabic-English translation does not drift.

The problems caused by unclear queries (such as vocabulary mismatch) are studied in [9]. The study represented that disambiguation of computer programming questions may be achieved by utilizing an ontology-based query expansion approach. The suggested model used the Cosine similarity algorithm to enhance the expanded query and, the search results.

Wu & Pan [10] presented an approach to semantic expansion based on a pay-as-you-go fashion mechanism. In the first step, after gathering information from users' search criteria, an initial query network is built and semantic similarities between the data (knowledgebase) and query network are determined. After that, sorting the similarity value in decreasing order yields the expanded sets of attribute values and edges. Finally, from the graph, an expanded query is formed by combining all or part of these additional values.

The authors in [11] utilized pseudo-relevant online knowledge derived from the top N web pages provided in response to the original query by Google, Bing, and DuckDuckGo to extract the first set of expanded keywords. The TF-IDF score, the kNN-based cosine similarity score, and the correlation score were the three weighting models that are used. The TF-IDF score weights the individual terms retrieved from the web content. The kNN-based cosine similarity score scores the expansion terms to produce the term-term association. Finally, the correlation scores are added to obtain the relevant expanded terms.

Sun et al. [12] provided an innovative query expansion and entity weighting (QEEW) technique that makes use of the connections between the entities in the entity catalog to improve the performance of query reformulation. The catalog consists of queries from users, responses from assistants, and related entities.

In [13] authors proposed a method that chooses terms based on (i) the semantic similarity between tags that compose a query, (ii) the proximity of the query-user relationship to customize expansion, and (iii) a strategy to expand user queries on the fly. An in-depth analysis of three major public datasets collected from Delicious, Flickr, and CiteULike illustrated the efficacy of the methodology.

The semantic-based QE technique helps cope with ambiguity and vagueness, as well as the limitations of an information retrieval system that include poor recall and low precision value. On the other hand, the QE techniques based on personalized data are centered on the goal of improving such Web queries by adding expansion terms obtained from the personal information repository of each user, which indirectly personalizes the results of the search. User intent is neglected in the Semantic method, and the semantic aspect of the question is not addressed in personalized techniques. Thereby, we devised a QE framework to take into account the query's semantics and user-related intent in the expansion process to provide more useful results. Thus, our approach differs from previous works in three aspects: first, we exploit linguistic knowledge to disambiguate the term senses accurately to support vocabulary mismatch issues. Second, we generate computer science-relevant concepts from ontology. Finally, the extracted concepts are evaluated and selected using a graph-based similarity method to formulate a precise expanded query that reflects the users' information requirements.

III. PROPOSED FRAMEWORK

Users on the Internet are frustrated by the inadequacy of existing search engines. This is because all users are being serviced by the same set of resources. Since everyone has their own set of interests and preferences, a single approach cannot possibly serve everyone. This highlights the need for personalized data (i.e. user profile) that can cater to individual needs in the process of QE. The proposed framework of QE exploits semantic structure and user profile to expand a query with user-relevant terms. Our QE framework is presented in Figure 1. The key components of the proposed framework are (1) Query Preprocessing, (2) Semantic Expansion, (3) Profile Building, (4) Profile Re-Building, and (5) Query Expansion & Information Retrieval. An explanation of how each component functions is included here as well.

The procedure begins with the user initiating a query and ends with the retrieving of the results. The first stage occurs when the user submits a query. By examining the user query words, we may choose the most relevant search terms and discard the irrelevant ones. At this point, the user query is broken down into words that are stemmed afterward. For instance, if a user enters the query 'spreading of COVID', we receive two words: spread and COVID. The phrase 'query processing' is used to describe this stage. The next step is to extract the meaning behind the user query. After the query words have been simplified via stemming the ontological connections are exploited to obtain the additional terms for QE. This stage is named semantic expansion.

After the semantic expansion stage, the user information in terms of tags is collected to create its profile (i.e., profile building stage). Later, in the profile re-building stage, the user tags are regularly updated via other user profiles. The last stage expands the query with terms obtained from an ontology and tags-based user profile. We have used weight allocation algorithms (such as the Wu & Palmer similarity formula) for each additional word to decide which additional words are most significant. Once the query has been successfully reformulated, submitting the reformulated query to the IR system is the final step of this stage. Relying on the expanded query, the search engine gives back a list of results that are then sorted.

In the following subsections, we will describe the components of the proposed model in detail.

A. Query Preprocessing

Query preprocessing is the initial step in our framework in which we remove stop words, stem query words, and recognize entities.

1) Stopping

Stopping is a method used to remove frequently used terms from indexes and searches [14]. It is necessary to go through this step since the majority of query keyword inclusion is useless in terms of finding user-relevant resources. The stop words have a grammatical purpose, but they are not very useful for differentiating one text from another. For example, the word 'the' is in most documents. If a user uses the word 'the' in a search query, it will not change the range of possible results much. Stop words include articles, conjunctions, and prepositions, and a typical stop list has between 300 and 500 words.

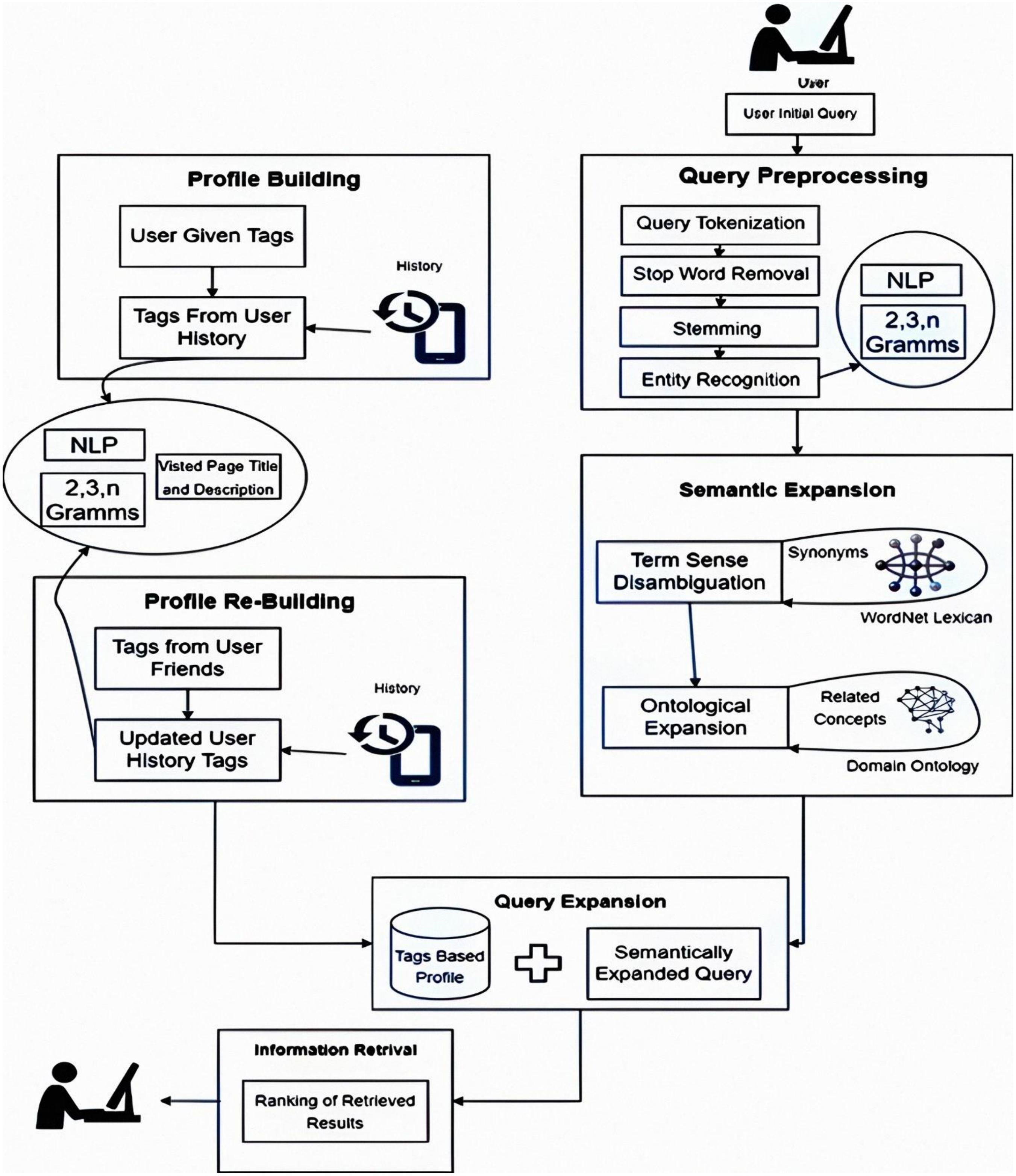

Figure 1: Proposed Approach for QE

2) Stemming

To go back to keywords' root forms, many keywords undergo a process called stemming, which involves deleting suffixes [15]. Stemming also involves the elimination of gerunds (which begin with the word ‘ing’), plurals, and past tense.

3) Entity Recognition

In search queries, keywords typically stand in for the entities being sought. If you type ‘harry potter walkthrough’ into the search bar, we assume you are searching for a game, but if you type ‘mountain view at a hotel in mountain view with pool’, we assume you are looking for a specific physical place. Entities in a query are separated by constructing various permutations of its constituent keywords and then checking each permutation independently to see whether it is an entity. We employ a grammatical method for merging the terms, that is, 2-Gram, 3-Gram, and N-Gram.

B. Semantic Expansion

The semantic expansion stage provides expansion words chosen both from Wordnet and the domain ontology. To expand a query semantically, we have adopted a semantic expansion algorithm [16]. A detailed explanation of each sub-step of this stage is provided below.

1) WordNet

In the first step, we utilize Wordnet to identify synonyms for our query keywords (identified from the previous stage). These synonyms serve as additional terms for QE.

2) Ontology

The COVID-19 pandemic continues to spread. The recent coronavirus pandemics have driven the critical need to thoroughly explore these diseases using cutting-edge technology. An ontology can provide a unique knowledge structure through which one can better examine infectious illnesses caused by coronaviruses.

The CIDO is a community-created, open-source biological ontology [17] to facilitate the study of coronavirus infectious disease. Figure 2 represents a portion of CIDO ontology. The CIDO provides standardized human and machine-readable annotations and representations of information related to infectious diseases caused by coronaviruses, including their etiologies, transmission, pathophysiology, diagnosis, prevention, and treatment. The CIDO development process complies with the OBO Foundry's Principles. In the semantic expansion stage, we have used CIDO ontology to obtain semantically related terms for a user query.

C. Profile Building

The stage creates the user profile using the following sub-steps.

1) User Information Collection

Collecting information about the user is a crucial first step in developing a profile that accurately reflects that user's preferences, habits, and needs. It is also essential to the process of user search history to understand the current demands of an individual user. Through the user's behaviors inside the system (i.e. search history), we may infer many things about user preferences.

A user's greatest source of information and explanation is the user. Every individual has superior knowledge about him or herself. Therefore, we consider this and continue to emphasize the user's explicit data. We give users a way to interact with our system so that they can add information in the manner of tags. The User Profile is then updated with these tags. Figure 3 shows the interface where a user's personal information (including self-information, job or work information, location, age, and hobbies) is entered into the system. This information represents user behavior and thus is useful in obtaining expansion terms for the query.

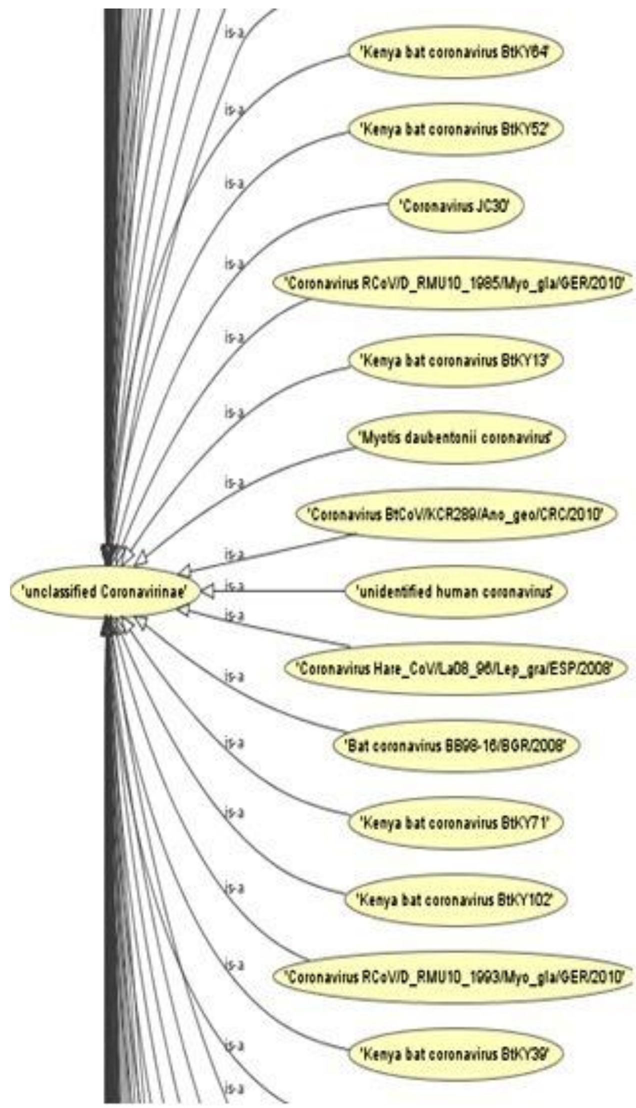

Figure 2: Snapshot of part of CIDO

Figure 3: Tags representing User Requirements

2) User Search History

A user's past search activity stores information about visited web pages. This history information helps find additional terms that relate to the user's search. We have also collected additional keywords from the user history and stored them in the user profile. To this end, we have filtered out entity keywords from a user's browsing history by first extracting the page title and metadata, and then using the entity recognition approach. We keep the users' profiles up-to-date by taking into account the last 10 websites they have visited. Table 1 shows a sample of tags collected from a user history.

Table 1: Tags obtained User History

User 1:
Disease, Allergy Disease, Infection, Cause Chest Infection, Symptoms Chest Infection, COVID-19, Coronavirus, Origin of Covid, Covid Affect Lungs, Omicron, Coronavirus Infection Disease, Long Covid, Respiratory Issues, Heart Transplant, Heart Attack Causes, Coronavirus Viral Disease

D. Profile Re-Building

The user history may evolve. The new pages that a user visits online were not get included in the initial user profile. If we do not regularly update a user profile, it will quickly become outdated. In light of the above, we have settled on expanding the query and retrieving more precise results by updating the user profile from recent history to include more and more user-relevant tags.

E. Query Expansion and Information Retrieval

The query expansion stage expands the user query using two sets of additional terms. The first set of additional terms is obtained from the outcome of the semantic expansion stage. The second additional terms set is obtained by exploiting the tags within the user profile. We have used Wu & Palmer similarity technique [18] to obtain the significant terms (relevant to the user) from these two sets of additional terms. The significant terms are then added to the original user query to obtain an expanded query which is then submitted to the IR system to retrieve the results.

Apache Lucene provides a robust and feature-rich IR system. The following are the two key responsibilities of Lucene:

- To compile an index of the articles that you want to review in order of priority.
- Search the index, process the query, and provide the results.

We have used Lucene and Vector Space Model (VSM) for the retrieval of query results.

1) Vector Space Model (VSM)

The VSM is an algebraic model used in many contexts, including data processing, retrieval tasks, document indexing, and ranking [19]. Using this model, we may describe and compare various documents and searches by their content, or keywords. Using Eq. 1, we can determine the degree of similarity between query and documents.

$$\mathrm{Sim}\,(D,P) = \frac{D.P}{||D||\,||P||} = \frac{\sum_{i=1}^{m} d_i p}{\sqrt{\sum_{i=1}^{m} d_i^2}\,\sqrt{\sum_{i=1}^{m} p_i^2}} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

where D represents the document and P represents the user query.

IV. RESULTS AND EVALUATION

This section evaluates the effectiveness of our semantic QE model and represents the results. TREC COVID-19 Open Research Dataset Challenge (CORD-19) dataset [20] and the COVID ontology were utilized in the research experiments. Overall, 50 queries were submitted to the Lucene IR via the proposed QE. The QE system provided the additional terms with the help of ontology and an updated user profile. Using the weighted Wu & palmer method, the highest-score terms that best describe the user's original query were then added to form an expanded query. Table 2 shows a sample of 10 queries and the expanded terms provided by the proposed system against each query.

Table 2: Query Samples

Original Terms	Expanded Terms
Covid Disease	Covid, Coronavirus, Covid-19, infection, illness, SARS-CoV-2
Coronavirus Symptoms	Covid, Covid 19, viral disease, SARS-CoV-2, indications, signs
Coronavirus Origin	Covid, beginning, Launch, Outbreak, start, SARS-CoV-2, Covid-19,
Viral Disease	Infection, Pandemic, outbreak, illness, disease
SARS vaccine	Viral vaccine, disease, Covid-19, Coronavirus Vaccine, Vaccinium

Evaluation measures such as Precision, Recall, and F1-score were used to assess the performance of the proposed framework.

Precision: Precision refers to the process of selecting just those results from the recovered set of results that are directly connected to the query. Eq. 2 represents the formula for determining precision.

$$\text{Precision} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{R_i}\sum_{j}^{R_j}P(D_j) \text{ ---------------------- (2)}$$

Recall: It is a categorization of the significant results that have been retrieved successfully. The formula for the computation of the recall is given in Eq. 3.

$$\text{Recall} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{r_i}{R_i} \text{ --------------------------------- (3)}$$

To retrieve the results from the chosen dataset, first, we input the user query directly to Lucene without adding any term to it (i.e., without expansion). Second, we add expansion terms obtained via our proposed approach to the original query and then submit the expanded query to Lucene (i.e., with expansion). From the obtained results via Lucene, we have calculated the precision. Table 3 presents the precision percentage with and without the expansion terms for 10 user input queries.

Table 3: Precision with and without query expansion

Query	Precision % Without QE	Precision% With QE
Q1	44	74
Q2	42	68
Q3	38	64
Q4	52	80
Q5	46	64
Q6	38	62
Q7	38	58
Q8	42	70
Q9	52	72
Q10	58	80

The following Figure 4 shows the precision results in graphical representation for with and without query expansion methods.

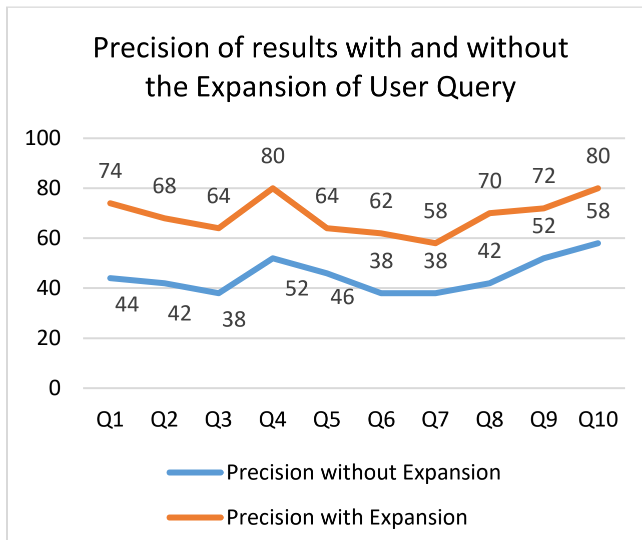

Figure 4: Precision With and Without Query Expansion

Table 4 shows the recall percentage with and without query expansion, while Figure 5 gives a graphical representation of the recall measure.

Finally, we calculated F1-Score for average precision and average recall to measure the overall performance of our proposed QE approach. The F1 performance is shown in Figure 6.

Table 4: Recall with and without query expansion

Query	Recall % Without QE	Recall % With QE
Q1	34	74
Q2	42	66
Q3	54	78
Q4	40	80
Q5	36	76
Q6	42	84
Q7	34	80
Q8	36	76
Q9	38	74
Q10	40	70

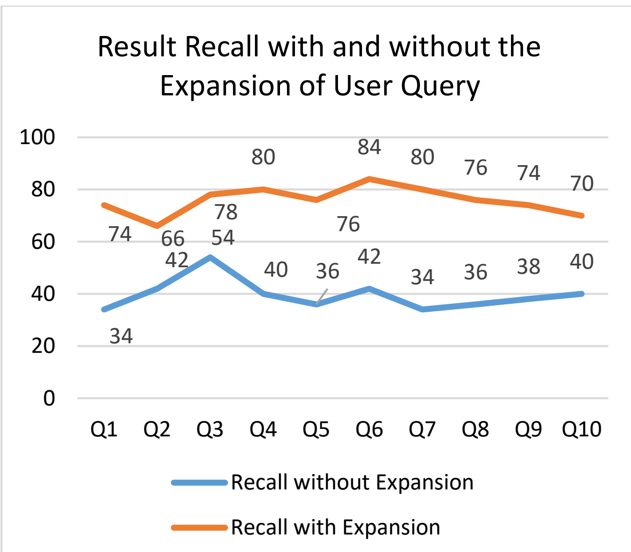

Figure 5: Recall with and without Query Expansion

V. CONCLUSION:

In this paper, we have presented an approach for QE which keep track of user intents as well as the semantics of the given query. Our proposed approach consists of several steps. First, we expanded the input query semantically with the help of Wordnet to overcome the word-mismatch problem and CIDO domain ontology to obtain the semantically related terms. For user intent, we made users' tag profiles with the help of the user-given tags and the search history. In addition, we have also kept the user profile updated using the user's friend profile and the latest search history. The profile is rebuilt to track the latest user preferences and behaviors. In the next step, we expanded the query with tags obtained from the updated user profile and with semantically related terms computed from ontology. In the end, the expanded query was submitted to the IR system (i.e., Lucene) to retrieve the results. The IR system then ranks the retrieved result so that a user can get the most related document on the top.

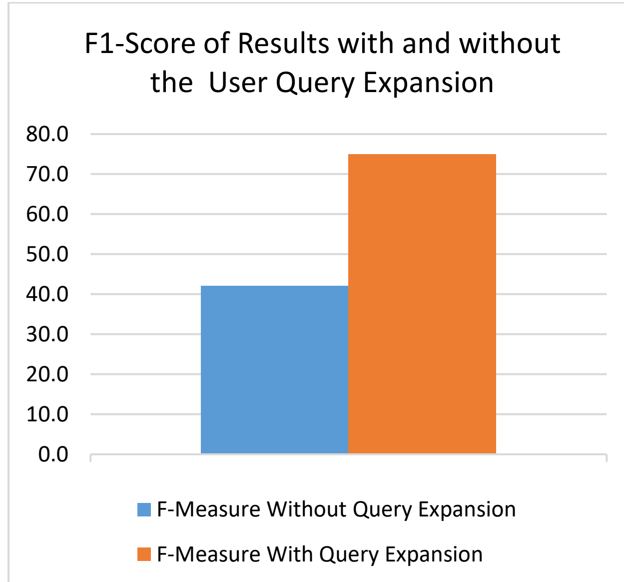

Figure 6: F1-score with and without Query Expansion

An expanded query produced results that are both exact and accurate. The highest level of precision that we attained with a user query (that is not expanded) is 58%. However, with the corresponding expanded query, 80% highest precision is obtained. Without expansion, the maximum recall for a query is 54%. When we added additional terms with the assistance of our QE approach, the recall is improved to a maximum of 84% for a given query. The F1-score for the proposed semantic QE approach is computed as 76%, whereby it was just 42.3% without expansion. We believe that this noticeable improvement in the F1-score is because the proposed semantic QE approach has not only focused on the semantics of user queries but also focused on user interests and preferences.

CREDIT AUTHOR STATEMENT

Muhammad Ahsan Raza: Conceptualization, Methodology, Visualization. **Muhammad Ali:** Data curation, Writing- Original draft preparation, Investigation. **Maruf Pasha:** Software Validation. **Mubashir Ali:** Writing-Reviewing, Visualization and Editing.

COMPLIANCE WITH ETHICAL STANDARDS

It is declared that all authors don't have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

REFERENCES

[1] S. Pani, B. Sahu, J. Mishra, S. N. Mohanty, and A. Panigrahi, "Pragmatic Analysis of Social Web Components on Semantic Web Mining," in Social Network Analysis, ed, 2022, pp. 83-108.

[2] P. Khare, "User Profile Mining and Personalization of Web Services," International Journal of Computer Applications, vol. 105, pp. 12-15, 11/14 2014.

[3] M. Raza, R. Mokhtar, N. Ahmad, M. Pasha, and U. Pasha, "A Taxonomy and Survey of Semantic Approaches for Query Expansion," IEEE Access, vol. PP, pp. 1-1, 01/24 2019.

[4] D. K. Sharma, R. Pamula, and D. Chauhan, "Semantic approaches for query expansion," Evolutionary Intelligence, vol. 14, pp. 1101-1116, 2021.

[5] D. Malhotra and O. P. Rishi, "A comprehensive review from hyperlink to intelligent technologies based personalized search systems," Journal of Management Analytics, vol. 6, pp. 365-389, 2019/10/02 2019.

[6] R. Ojha and G. Deepak, "Metadata Driven Semantically Aware Medical Query Expansion," Cham, 2021, pp. 223-233.

[7] S. Jain, K. R. Seeja, and R. Jindal, "A fuzzy ontology framework in information retrieval using semantic query expansion," International Journal of Information Management Data Insights, vol. 1, p. 100009, 2021/04/01/ 2021.

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

[9] M. Khedr, F. El-Licy, and A. Salah, "Ontology based Semantic Query Expansion for Searching Queries in Programming Domain," International Journal of Advanced Computer Science and Applications, vol. 12, 01/01 2021.

[10] N. Wu and Y. Pan, "Semantic query expansion method based on pay-as-yougo fashion for graph model," Journal of Physics: Conference Series, vol. 1971, 2021.

[11] H. K. Azad, A. Deepak, C. Chakraborty, and K. Abhishek, "Improving query expansion using pseudo-relevant web knowledge for information retrieval," Pattern Recognition Letters, vol. 158, pp. 148-156, 2022/06/01/ 2022.

[12] Z. Sun, S. Lu, C. Ma, X. Liu, and C. Guo, Query Expansion and Entity Weighting for Query Reformulation Retrieval in Voice Assistant Systems, 2022.

[13] M. R. Bouadjenek, H. Hacid, and M. Bouzeghoub, "Personalized Social Query Expansion Using Social Annotations," in Transactions on Large-Scale Data- and Knowledge-Centered Systems XL, A. Hameurlain, R. Wagner, F. Morvan, and L. Tamine, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, pp. 1-25.

[14] I. H. Witten, A. Moffat, and T. C. Bell, "Managing Gigabytes: Compressing and Indexing Documents and Images," IEEE Transactions on Information Theory, vol. 41, p. 2101, 1995.

[15] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, pp. 130-137, 1980.

[16] M. Raza, R. Mokhtar, N. Ahmad, and M. Ashraf, "Sensual Semantic Analysis for Effective Query Expansion," *International Journal of Advanced Computer Science and Applications*, vol. 9, 01/01 2018.

[17] S. Jupp et al., "A new Ontology Lookup Service at EMBL-EBI," In: Malone, J. et al. (eds.) Proceedings of SWAT4LS International Conference 2015, 2015.

[18] D. Guessoum, M. Miraoui, and C. Tadj, "A modification of wu and palmer semantic similarity measure," in The Tenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2016, pp. 42-46.

[19] M. W. Berry, Z. Drmac, and E. R. Jessup, "Matrices, Vector Spaces, and Information Retrieval," SIAM Rev., vol. 41, pp. 335–362, 1999.

[20] A. I. F. Al, A. Goldbloom, P. Lin, P. Mooney, C. Schoenick, S. Kohlmeier, *et al.* (30-May-2022). *COVID-19 Open Research Dataset Challenge (CORD-19)*. Available: https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge