Int. J. Advance Soft Compu. Appl, Vol. 13, No. 2, July 2021
Print ISSN: 2710-1274, Online ISSN: 2074-8523

User Interest Driven Semantic Query Expansion for Effective Web Search

Muhammad Ahsan Raza[1], M. Rahmah[2], Fahad Qaswar[3], Anam Shafqat[4] Muhammad Rauf[5] and Sehrish Raza[6]

[1]Department of Information Technology, Bahauddin Zakariya University, Multan, Pakistan
e-mail: ahsan_0136@yahoo.com
[2, 3]Faculty of Computing, College of Computing & Applied Science, Universiti Malaysia Pahang, Pekan, Pahang, West Malaysia
e-mail: drrahmah@ump.edu.my, fahadqaswar200@gmail.com
[4]Department of Computer Science, Government College University, Faisalabad, Pakistan
e-mail: aanie.me@gmail.com
[5]Department of Computer Science, Institute of Southern Punjab, Multan, Pakistan
e-mail: muhammadrauf78645@gmail.com
[6]Institute of Computer Science and Information Technology, The Women University, Multan, Pakistan
e-mail: sehrish.6025@wum.edu.pk

Abstract

Retrieving user-relevant content from a large volume of data available on the Web via an input query is a difficult task. A user query may not be able to specify user information needs due to the ambiguous and limited number of query terms. The semantic query expansion (QE) strategy offers a solution to this problem by expanding the query with additional terms, which are semantically similar to the original query. However, this strategy does not consider individual user interest in the generation of expansion terms. In this article, semantic QE is improved by combining the notion of ontology knowledge and user interest. The proposed semantic QE technique involves a computing domain of the input query via ontology, generates expansion terms from the user browsing history, and finally selects expansion terms that represent user preferences on the basis of the semantic similarity between expansion terms and query and user feedback. The experimental evaluation indicates that expanded queries produced by the proposed technique retrieve more personalized contents over Web search than

initial user queries. The obtained results achieve 86.4% average precision, which proves a positive impact of incorporating user preferences in semantic QE.

Keywords: *Browsing history, Information retrieval, Ontology, Personalize search, Semantic computing.*

1 Introduction

In recent years, with the exponential growth of the World Wide Web (WWW), search mechanisms have changed positively, that is, from simple keyword-based matching to semantic search [1]. However, the retrieval results from the current searching techniques may contain irrelevant data; thus, users must struggle to achieve precise and correct outcomes. For the retrieval of user-interested documents, improving queries during WWW search is necessary. When dealing with search queries, experienced and inexperienced users are improperly trained and have no knowledge of a particular domain and thus cannot achieve their relevant outcomes. One way to improve user-written queries is through the automatic expansion of queries with additional relevant terms.

The query expansion (QE) mechanism provides an opportunity for finding the additional terms relevant to an initial user query, thereby expanding it to satisfy user needs. A variety of QE techniques are applied in literature to enhance the original user queries for the retrieval of improved results via searching systems [2]. Semantic QE is an approach of computing and inserting meaningful information to search a query. According to [4], the semantic methods of QE rely on knowledge structures, such as thesaurus or ontology, for the extraction of meaningful terms to expand the user query. However, the deficiency of semantic QE systems lies in distinguishing different users and predicting their individual needs. Thus, the inclusion of user interests in the semantic QE method is indispensable to meet the demands of individual users.

This research aims to semantically expand a user query while realizing user interests to attain a personalized Web search. The semantic element of the proposed QE paradigm is achieved via ontology knowledgebase, and user preferences are collected using the browsing history. It lets new users handle short and vague queries. The proposed system attempts to obtain user-related content over the WWW on semantic basis. The main achievements of this study are two-fold, as described below.

1. Design of semantic QE model based on domain ontology to extract the possible domains of the query at the conceptual level.
2. The integration of user browsing history and domain-ontology semantics to distinguish individual user preferences over the WWW search.

The rest of this article is organized as follows. QE techniques based on knowledge structures and personalize data are reviewed in Section 2. Section 3 describes the

functionality of key components of our approach alongwith the knowledge sources used in the QE strategy. The evaluation metrics and analysis of results are discussed in Section 4. Section 5 outlines the conclusion and presents the future work.

2 Related Work

The QE mechanism determines and appends additional related terms to a user query to improve the searching of information retrieval (IR) systems. Existing QE techniques can be categorized as statistical techniques [3] and semantic techniques [5]. The semantic approach for QE gains the attention of the research community, as this approach exploits knowledge structures, such as ontology, thesaurus or topic maps [5]. These knowledge structures are useful in computing meaningful terms via implicit relationships, which otherwise are difficult to obtain from textual corpus.

Recent semantic QE techniques have explored different variants of ontology (including domain-specific and domain-independent ontologies) in the generation of expansion terms. For example, authors in [6] exploited Arabic WordNet knowledgebase to obtain concepts similar to user query. The proposed hybrid QE technique expanded queries with high accuracy in comparison to existing QE techniques. Cui et al. [7] created a medical ontology graph and used it to generate candidate expansion terms in the medical field. The authors perceive that medical ontology is a better source for deriving query-related terms than the thesaurus. The experimental results showed improved results for the proposed method over the thesaurus-based expansion. In [8], authors leveraged two knowledge sources, namely, domain-independent WordNet ontology and Wikipedia, to extract semantic data (i.e., candidate terms) on the basis of an input query. The expanded queries exhibited improved results over Web than non-expanded queries.

Authors in [3] pointed out another vein of QE, namely, personalized QE that includes user information (e.g., user profile) as data source in the process of the user query expansion. For example, Chen et al. [9] proposed a food recommendation model, whereby the QE module within the model extracts user interest from dietary and health data sources. The proposed personalized recommendation model was compared with non-personalized models in terms of expert judgment and mean average precision (MAP) metrics. The proposed model achieved a higher score in human evaluation and a better MAP value than the counterpart models. In [10], authors developed a user profile from browsing data and knowledge graph. The user profile was then exploited by the proposed algorithm to rank the retrieval results for achieving personalized search. Overall, the proposed method showed +35% improvement in precision compared with the base system. Similarly, [11] generated personalized expansion terms from user profiles. These user profiles were created from the resources of folksonomy knowledgebase, which is tagged and annotated by users to represent their topics

of interest. The experimental evaluation over three datasets showed better MAP values for the proposed QE method using user profiles than the baseline model.

However, our proposed QE method differs from existing semantic and personalized QE approaches in the following three aspects: first, we use ontology knowledgebase to identify the domain of the user search query. Second, we generate expansion terms from a browsing history document set on the basis of an identified query domain. Last, we take the benefit of the semantic similarity method and user feedback in the selection of expansion terms, which closely reflect user interests.

3 Semantic QE System Based on User Preferences

The proposed system is composed of four main steps, as shown in Fig. 1. In the first step, keywords are extracted from the given query. In the second step, after the noise is removed from the query, a set of expansion terms is generated by exploiting domain ontology and user history. In the third step, the expansion terms, which are close to the keywords (generated in the first step), are calculated to rank them. In the final step, an expanded query is formulated on the basis of highly ranked terms. This expanded query then facilitates the generation of user-related content via the IR system.

3.1 User Interface

The user interface of this model is the screen with which users interact with the system. It is used to give a query and obtain results from the IR system. After the processing of query via our proposed QE system, the results are retrieved by the IR system on the basis of the expanded query.

3.2 Keyword Extraction

Keyword extraction is a task of automatic identification of terms, which best describe the subject of a query. It consists of two sub-steps: stop word removal and stemming.

3.2.1 Stop Word Removal

A user query may contain noise words (e.g., of, the, an, to, who), which affect the overall system performance. These words are said to be stop words. In this phase, the stop words are removed from the user query to extract the keywords. To remove noise words, we use Java tokenizer and a list of English stop words given by [12].

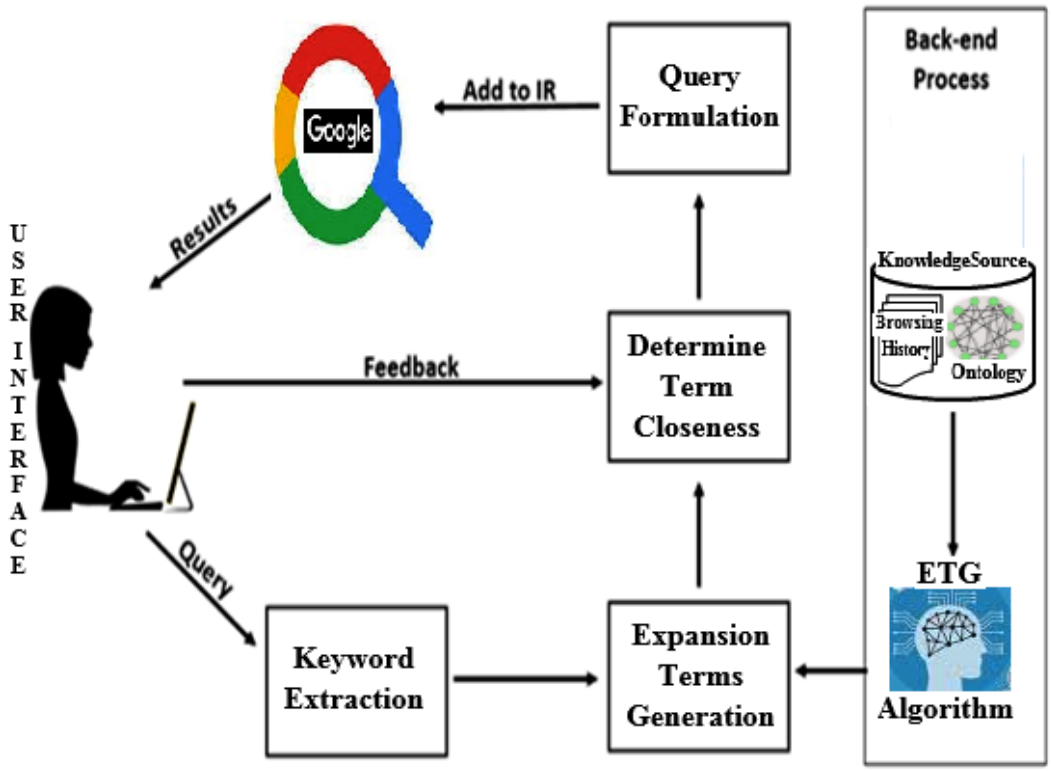

Fig. 1: Major components of semantic query expansion system

3.2.2 Stemming

The query keywords obtained after the removal of stop words may be in different forms; for instance, keyword "expand" has forms such as "expanded" or "expansion." Stemming is the process that reduces a keyword to its stem word [13]. We use the Porter algorithm, which removes suffixes from the keywords. Stemming provides keywords that can be helpful in the generation of expansion terms by using the QE algorithm.

3.3 Expansion Term Generation (ETG)

After the removal of noise words, a set of expansion terms is generated via the proposed ETG algorithm. It processes two knowledge sources, namely, domain ontology and user browsing history on the basis of original query keywords, and provides a set of expansion terms as output. In this section, we first describe the knowledge sources, and then highlight the performance of the ETG algorithm.

3.3.1 Knowledge Sources

The ontology and browsing history knowledge sources are utilized to depict the user query semantics and user preferences, respectively.

a) Ontology

Ontology represents the model (concepts and their relationships) of a specific domain [18] and therefore is a key source for obtaining additional meaningful concepts about the domain. We utilize computer science (CS) ontology, which is constructed by [14] with the help of the CS curriculum. The protégé view of CS ontology is illustrated in Fig. 2. The left pane of the figure shows the concept hierarchy of the ontology, whereas the right pane shows the detail of each selected concept from the CS hierarchy, such as equivalent and disjoint classes.

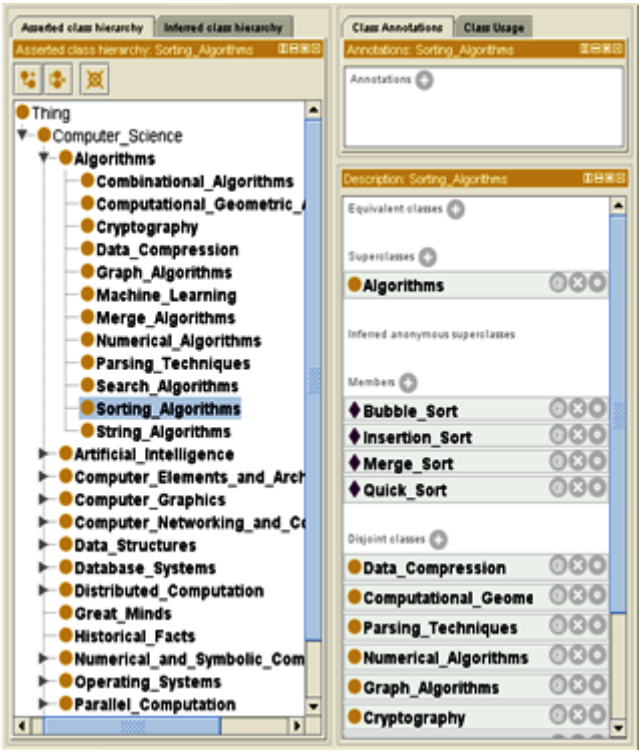

Fig. 2: Protégé view of ontology representing classes hierarchy

b) Browsing History

User browsing history contains documents clicked by a user. These documents indicate the kind of information in which the user is interested. Authors in [15] observed that a browsing document set may contain noise, such as unwanted document or shift in user interest. However, they argued that using a complete set of user-browsed documents is reliable in obtaining most user preferences. In this research, we only focus on documents available in the browsing history. The idea is to exploit these documents for extracting user interests in the form of expansion terms.

3.3.2 Expansion Algorithm using Ontology and Browsing History

The main objective of the ETG algorithm is to utilize user preference along with ontology knowledge for the expansion of user query to retrieve user-relevant results. The generation of expansion terms by using the ETG algorithm involves two key aspects: the first is to judge the domain of a query input by a user. Ontology relationships are exploited for this purpose. The second is to find user interests (appropriate expansion terms) from the browsing history on the basis of an identified domain of the query.

a) Identifying Query Domain

The ETG algorithm relies on the ontology structure to recognize the domain of the input query. To this end, we follow the steps in the sensual semantic expansion (SSE) module [16]. However, we omit the sense identification steps of SSE and focus on parent–child relationships to obtain the domain of the given query. The process of identifying the query domain proceeds as follows:

1. Find a match in the ontology for each keyword in query Q.
2. For each match, extract parent and child concepts via IS-A relationships in the ontology. Append extracted concepts in Q´ to formulate a new query.
3. Use Q´ to retrieve further expansion terms from the browsing history.

b) Finding Expansion Terms

This sub-step of the ETG algorithm determines the expansion terms from the set of documents listed in the browsing history of the user. For each document, the cohesion relations (i.e., correlation score between terms) between document terms are computed with the query keywords and domain (as identified in the previous section). Document terms with high correlation scores are then selected as expansion terms. The cohesion score can be measured using equation 1.

$$CoScore_{D(t)} = \prod_{K(i) \in Q} \left(P(D_{(t)} | K_{(i)}) + 1 \right) \qquad (1)$$

where

$$P(D_{(t)} | K_{(i)}) = \frac{P(D_{(t)}, K_{(i)})}{P(K_{(i)})}$$

D(t) denotes a document term, K(i) depicts a keyword in query Q, and P(D(t) | K(i)) represents the probability of the co-occurrence of the document term and query term.

In sum, appropriate expansion terms can be determined using the following steps:

1. For each term in document D, calculate its cohesion score with all terms in Q′ by using equation 1.
2. Select terms with high cohesion scores (i.e., above threshold values) as expansion terms E. Append E to Q′.
3. Submit Q′ to the IR system to retrieve user-related content.

The step-by-step computation of the ETG algorithm for the generation of terms from ontology and browsing history is given in Algorithm 1.

Algorithm 1 ETG Algorithm

Input : Q , The set of query keywords
DO, Domain ontology
BH, Browsing history

Output : Q′ , The set of expansion terms

// Identifying Query domain

```
1   FOR each keyword(Ki) in Q
2   |  IF keyword is found in DO
3   |  |  Extract Ki domain(s) via ISA relationship in DO
4   |  |  Add domain(s) in Q′
5   |  END-IF
6   END-FOR
      // Finding expansion terms
7     FOR each domain(Dj) in Q′
8     |  IF Dj is found in BH document
9     |  |  Compute cohesion score with document terms
10    |  |  Select terms with high cohesion
11    |  |  Append terms in Q′
12    |  END-IF
13    END FOR
```

3.4 Determine Term Closeness

From the set of expansion terms (i.e., output of the ETG algorithm), this step further determines which terms have high relevancy to the initially input query. Researchers have used various measuring schemes for determining term closeness; for example, [17], [19], and [20] used the BM-25 scheme, the KLD measure, and the TF-IDF similarity score, respectively. Meanwhile, we use two similarity stages to calculate the weightage of terms: (1) Wu and Palmer (WAP) similarity scheme and (2) user feedback.

3.4.1 WAP Similarity

WAP similarity is used to measure the weightage of terms by using the characteristics of WordNet thesaurus [21]. Highly relevant expansion terms are

selected on the basis of a threshold value. The formula of WAP is described in equation 2.

$$WAP_{weight} = \frac{2 \times D\ (LCS)}{\left(D(T_1) + D(T_2)\right)} \quad (2)$$

where Ti is the term whose WAP weight is needed to be calculated, D represents the depth of Ti in a graph of the WordNet thesaurus, and LCS depicts the least common node among the terms Ti.

3.4.2 User Feedback

In this step, the user gives feedback about the terms to be finally used for the expansion of the initial user query. The chosen expansion terms (that are selected on the basis of WAP weight) are presented to the user for the further pruning (i.e., closest to information needs) of expansion terms. Finally, the user-selected terms are added to the original query.

3.5 Query Formulation

Highly relevant terms (i.e., outputs of two-staged similarity schemes) are added to the original query to formulate a new precise expanded query. The two sets of terms, that is, original terms and the selected expansion terms are added using the Boolean operator AND and OR. The newly formulated query is given to the IR system for the retrieval of user-related content. In our system, we use Google for two reasons: (1) it is a popular IR engine and (2) the retrieval effectiveness of the proposed system can be well validated over huge WWW corpus.

4 Performance Evaluation

To measure the performance of the proposed QE technique, the expanded queries are submitted to the Google engine to obtain user-relevant documents. The next sub-sections present the performance measures, which are utilized to show the overall performance of our system, and the result discussion.

4.1 Performance Metrics

Precision is the common metric used by researchers for exhibiting the performance of QE techniques [9]. This metric shows the ability of the proposed technique to withhold the retrieval of irrelevant results (i.e., documents unrelated to user needs). Equation 3 shows the formula for the calculation precision.

$$P = \frac{RLS}{RD} \quad (3)$$

where RLD refers to the relevant number of documents in the result, and RD means the total number of documents retrieved for a query.

This research utilizes two variants of precision metrics to show the performance of the proposed semantic QE, which is based on user interest. The first variant is Precision@50 (P@50), where precision is computed at the top 50 retrieved Web documents. The second variant is Precision@100 (P@100), which shows precision at the top 100 Web results.

4.2 Result Discussion

We compare our proposed semantic QE technique, which incorporates user preferences, with a base technique (i.e., initial query search without expansion). We input two queries: the initial user query and the corresponding expanded query in the Google IR system. The top 50 and top 100 results of the IR system are given attention because most users usually traverse five pages of Google search results at most [22].

Fig. 3 displays the precision values at the top 50 results (P@50) for the initial and expanded queries. From the result of 10 sample queries, the maximum precision is obtained for expanded query number 9 (i.e., 92%). However, the precision percentage for the same initial query is 72%. That is, our proposed model is +20% better in achieving precise results than Google. Meanwhile, a low precision percentage (which is 70%) is obtained for expanded query number 8. For the same query, the proposed model exhibits +30% improvement compared with the initial query precision. Moreover, the maximum precision percentage for the initial query technique is 72%, which is the minimum percentage achieved in our proposed technique. Considering the P@50 result of all 10 queries, our proposed semantic QE technique, which is based on user interest, outperforms the base search technique.

To further verify the accuracy of the proposed system over the Google engine, we also test the precision results for the top 100 retrieved results (i.e., P@100). This verification is important to show that the proposed technique performs even better if more retrieved Google results are considered for precision calculation. Fig. 4 illustrates the P@100 results for the initial and corresponding expanded queries. The lowest precision (which is 70%) is obtained for expanded query number 8 by using the proposed technique. This query still has +30% improved precision compared with the initial query. The same percentage difference can be observed in the P@50 result for query number 8. Therefore, the proposed technique performance is not degraded even at the top 100 retrieved Google search results. In addition, the maximum precision percentage for the initial query technique is 66%, which is even less than the minimum percentage achieved in our proposed technique (i.e., 70%). Considering P@50 and P@100 results, our proposed semantic QE, which incorporates user interests, outperforms the initial query search technique.

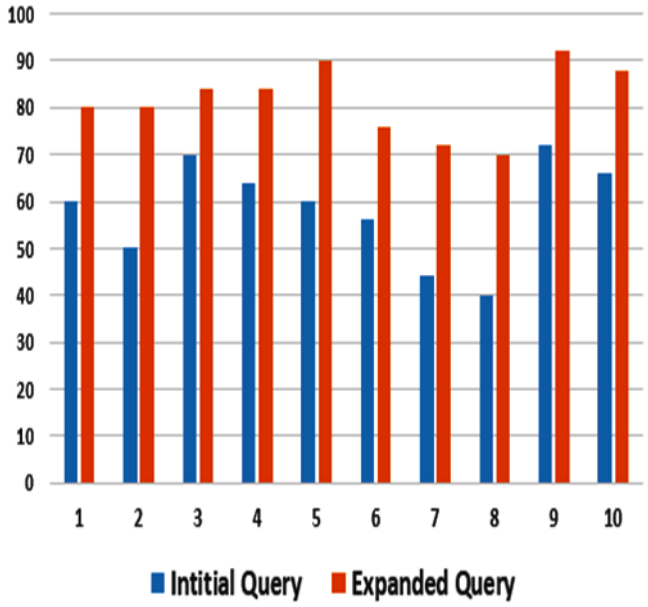

Fig. 3: Precision measures for queries at top 50 results

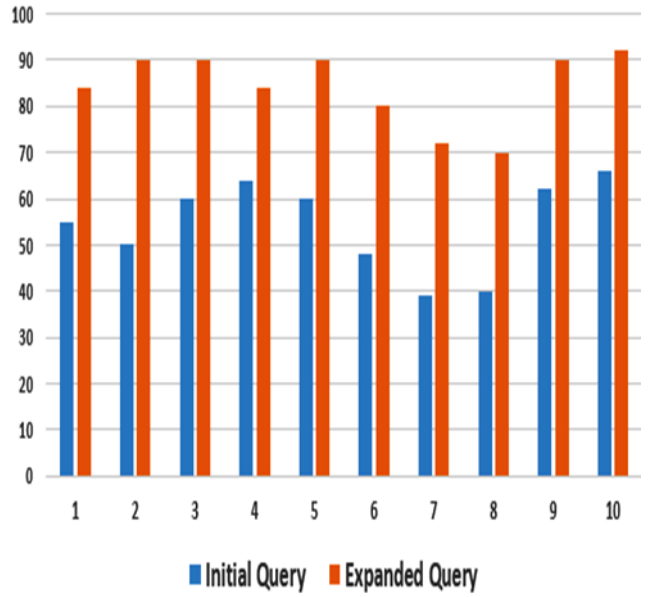

Fig. 4: Precision measures for queries at top 100 results

We also calculate the average precision to show the overall performance of the proposed and base systems. Fig. 5 displays the average precision values at the top 50 results (i.e., average P@50) and top 100 results (i.e., average P@100). The average P@50 percentage for the initial query system is 58.2%, whereas that for the proposed system is 81.6%. That is, the proposed semantic QE system, which focuses on user interest, achieves far better precision percentage than the base system (i.e., initial query search system). In terms of average P@100, our

proposed semantic QE technique shows a +29.8% average improvement compared with the base system. The average P@100 for the initial query system decreases compared with the average P@50 results. Meanwhile, the proposed QE technique achieves better average P@100 than average P@50. Thus, as the number of results (from top 50 to top 100 results) increase, the precision of the proposed method is improved (which is +2.6%). This improvement suggests that the proposed QE method, which incorporates user preferences, expands queries with semantically and user-related concepts and better satisfies the user search requirements than non-expanded queries (i.e., initial queries input by users).

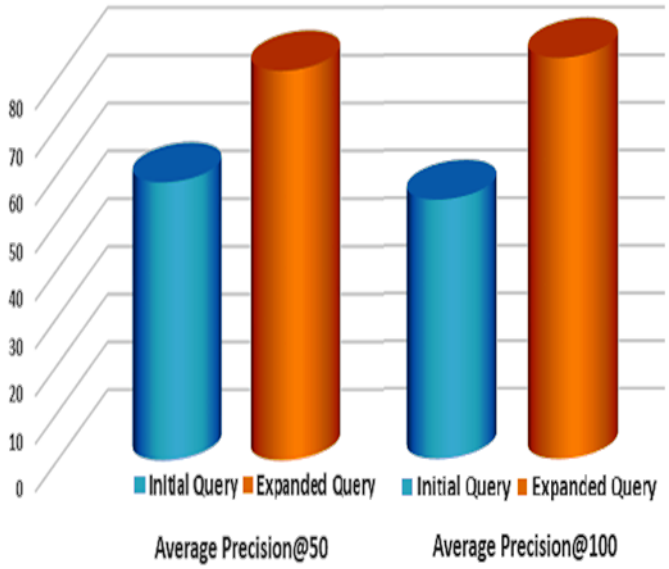

Fig. 5: Average Precision for queries at top 50 and 100 results

5 Conclusion

The mismatch between search query terms and documents affects the retrieval results of the existing IR system. Many semantic QE methods have been proposed to solve the term mismatch issue. These methods take advantage of the ontology knowledge source to expand search queries with terms semantically relevant to the original search query. However, semantic QE does not consider individual user interests, which can be extracted from the Web browser history.

The use of ontology can help in obtaining domain semantics, whereas user preferences can be collected from the browsing history. Based on these ideas, we incorporate user interests into the process of semantic QE. Compared with existing semantic QE techniques, our system identifies the user query domain via ontology semantics (by exploiting ontology relationships) and captures user intents from history logs (via correlations). The retrieval results over Google for queries expanded by our system achieves better precision than the initial query

results. The proposed technique achieves 81.2% and 86.4% average precision for top 50 and 100 Google results, respectively. Therefore, combining semantics and user interests can achieve substantial improvement in precision results.

In this research, we focus on domain-specific ontologies for the identification of an initial query domain and utilize history logs for the generation of an expansion term set. In the future, we want to explore the effect of a large ontology (domain independent), as a domain-specific ontology contains limited terms. We further plan to exploit browsing history features, such as query session and document click time, in the process of expansion term extraction.

ACKNOWLEDGEMENTS
The authors would like to thank the Ministry of Higher Education, Malaysia for providing financial support under Fundamental Research Grant Scheme (FRGS) No. FRGS/1/2018/ICT04/UMP/02/1 (University reference RDU190112) and Universiti Malaysia Pahang for research facilities.

References

[1] Chen, L., Shang, S., Yang, C., & Li, J. (2020). Spatial keyword search: a survey. GeoInformatica, 24(1), 85-106. doi: 10.1007/s10707-019-00373-y
[2] Pasha, M., & Raza, M. A. (2011). Analyzing Query Expansion Techniques. Paper presented at the 2011 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS 2011), Guangzhou, Guangdong, China.
[3] Raza, M. A., Mokhtar, R., & Ahmad, N. (2019). A survey of statistical approaches for query expansion. Knowledge and Information Systems, 61(1), 1-25. doi: 10.1007/s10115-018-1269-8
[4] Alromima1, W., Moawad, I. F., Elgohary, R., & Aref, M. (2016). Ontology-based Query Expansion for Arabic Text Retrieval. International Journal of Advanced Computer Science and Applications (IJACSA), 7(8).
[5] Raza, M. A., Mokhtar, R., Ahmad, N., Pasha, M., & Pasha, U. (2019). A Taxonomy and Survey of Semantic Approaches for Query Expansion. IEEE Access, 7, 17823-17833. doi: 10.1109/ACCESS.2019.2894679
[6] Almarwi, H., Ghurab, M., & Al-Baltah, I. (2020). A hybrid semantic query expansion approach for Arabic information retrieval. Journal of Big Data, 7(1), 39. doi: 10.1186/s40537-020-00310-z
[7] Cui, X., Zhai, P., & Fang, Y. (2020). Semantic Query Expansion based on Entity Association in Medical Question Answering. Journal of Physics: Conference Series, 1642, 012022. doi: 10.1088/1742-6596/1642/1/012022
[8] Azad, H. K., Deepak, A., & Abhishek, K. (2020). Query Expansion for Improving Web Search. Journal of Computational and Theoretical Nanoscience, 17(1), 101-108. doi: 10.1166/jctn.2020.8635
[9] Chen, Y., Subburathinam, A., Chen, C., & Zaki, M. J. (2021). Personalized Food Recommendation as Constrained Question Answering over a Large-scale Food Knowledge Graph. ArXiv, abs/2101.01775.

[10]Wiem, C., Mohammad, O. W., Haiyan, L., & Omar Ghaleb, E. (2020). Context-Aware Personalized Web Search Using Navigation History. International Journal on Semantic Web and Information Systems (IJSWIS), 16(2), 91-107. doi: 10.4018/IJSWIS.2020040105

[11]Bouadjenek, M. R., Hacid, H., & Bouzeghoub, M. (2019). Personalized Social Query Expansion Using Social Annotations Transactions on Large-Scale Data- and Knowledge-Centered Systems XL (pp. 1-25). Berlin, Heidelberg: Springer Berlin Heidelberg.

[12]NL, R. (2011). Long stopword list. Retrieved February 29, 2021

[13]Porter, M. F. (1997). An algorithm for suffix stripping Readings in information retrieval (pp. 313–316): Morgan Kaufmann Publishers Inc.

[14]Raza, M. A., Rahmah, M., Raza, S., Noraziah, A., & Hamid, R. A. (2019). A Methodology for Engineering Domain Ontology using Entity Relationship Model. International Journal of Advanced Computer Science and Applications (IJACSA), 10(8), 326-332.

[15]Rastegari, H., & Shamsuddin, S. M. (2010). Web Search Personalization Based on Browsing History by Artificial Immune System. International Journal of Advances in Soft Computing and Its Applications, 2(3).

[16]Raza, M. A., Rahmah, M., Noraziah, A., & Ashraf, M. (2018). Sensual Semantic Analysis for Effective Query Expansion. International Journal of Advanced Computer Science and Applications(IJACSA), 9(12). doi: http://dx.doi.org/10.14569/IJACSA.2018.091208

[17]Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., & Daigremont, J. (2011). Personalized social query expansion using social bookmarking systems. Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

[18]Zakaria, N., Hassan, R., Othman, R., & Asmuni, H. (2015). Maturity-based analysis of lightweight ontology from the aspect of extensibility, reusability and evolutionary. International Journal of Advances in Soft Computing and Its Applications, 7(2), 54-74.

[19]Singh, J., & Sharan, A. (2015). Relevance feedback based query expansion model using Borda count and semantic similarity approach. Computational intelligence and neuroscience, 2015.

[20]Azad, H. K., & Deepak, A. (2019). A new approach for query expansion using Wikipedia and WordNet. Information sciences, 492, 147-163.

[21]Hamza, M. A., Ab Aziz, M. J., & Omar, N. (2020). Sentence Similarity Measurement for Smart Education Based on Thematic Role and Semantic Network Techniques. International Journal of Software Engineering and Computer Systems, 5(2), 37–65.

[22]McMahon, C., Johnson, I., & Hecht, B. (2017). The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.