Sensual Semantic Analysis for Effective Query Expansion

Muhammad Ahsan Raza[1], M. Rahmah[2], A. Noraziah[3]
Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Kuantan, Malaysia

Mahmood Ashraf[4]
Department of Information Technology
Bahauddin Zakariya University
Multan, Pakistan

***Abstract*—The information has evolved rapidly over the World Wide Web in the past few years. To satisfy information needs, users mostly submit a query via traditional search engines, which retrieve results on the basis of keyword matching principle. However, a keyword-based search cannot recognize the meanings of keywords and the semantic relationship among the terms in the user's query; thus, this technique cannot retrieve satisfactory results. The expansion of an initial query with relevant meaningful terms can solve this issue and enhance information retrieval. Generally, query expansion methods consider concepts that are semantically related to query terms within the ontology as candidates in expanding the initial query. An analysis of the correct sense of query terms, rather than only considering semantic relations, is necessary to overcome language ambiguity problems. In this work, we proposed a query expansion framework on the basis of query sense analysis and semantics mining using computer science domain ontology, followed by working prototype of the system. The experts analyzed the results of system prototype over test dataset and Web data, and found a remarkable improvement in the overall search performance. Furthermore, the proposed framework demonstrated better mean average precision and recall values than the baseline method.**

Keywords—Semantic computing; information retrieval; computational intelligence; ontology; term sense disambiguation

I. Introduction

At present, the volume of information over the World Wide Web (WWW) has been increasing continuously. Current search engines share this diverse information pool of the WWW and retrieve results by using simple keyword-based matching. These search engines cannot recognize the semantic relevance between search text and student query, thus receiving increased results that are irrelevant to computer science. In this situation, designing a system that interprets user search requirements correctly, rather than providing results by merely performing keyword-based matching, is challenging.

Query expansion is a technique that can be used for effective information searches to satisfy users' requirements. The query expansion process involves augmenting the initial user query with additional terms that are related to user requirements. Currently, among several query expansion techniques studied in [1], ontology browsing is considered a prominent query expansion technique. Ontology provides semantics to plain text [2]; thus, finding additional query-related concepts by exploring the semantic relations is useful in exploring semantic relations.

Focusing on computer science discipline, where data are unstructured and dispersed over WWW, this research work proposes an alternate sense-based semantic query expansion framework to overcome the word mismatch problem of keyword-based searches. Given a user query, the approach initially captures the set of senses for query terms. Then, the relevant concepts from ontology are extracted on the basis of term-sense data. Finally, the extracted concepts are used to expand the initial query for obtaining user-centric results.

Our approach extends the model presented in [3] via disambiguation of query term sense and semantic similarity strategy for selecting and ranking expanded terms.

The remainder of this paper is organized as follows. Existing techniques for query expansion based on ontology are reviewed in Section 2. Section 3 outlines the major steps of our approach alongwith the ontologies used in the query expansion method. The query expansion framework is discussed, and the functionality of each component of the framework is clarified in Section 4. Section 5 details the experiment results and analysis of this work. Section 6 presents the conclusion and highlights the future work.

II. Related Work

Existing keyword-based search techniques have been used for retrieving information from large unstructured data on the WWW [4]. Such techniques retrieve results on the basis of matching the keywords from the user query. However, keyword-based techniques lack semantic orientation and cannot capture the user information requirements.

Query expansion is used to improve the performance of information retrieval system and retrieve results that are user-relevant. Ontology is useful in query expansion because it provides a means for discovering unstated concepts that can be used to expand the initial user query. Early works have explored the use of ontology in query expansion techniques that have been extended eventually in different ways, such as domain-specific ontology [5-7], general ontology [8-9] and linguistic expansion [10-11].

Bhogal, Macfarlane, & Smith in [12] have reviewed the role of ontology in discovering the terms for expanding seed query, whereas [13] provided a comprehensive overview of recent query expansion techniques for supporting an effective information retrieval. Authors in [14] contended that general and similar concepts related to the original query can be identified using thematic relations of ontology. The researchers used semantic relations and qualifiers (i.e., specified in seed query) to filter possible features for reformulation of a new expanded query. The work focused on geographical test data and queries and showed improvements in the accuracy of results. In [15], authors leveraged ontology to obtain the rarely occurred opinions about a product. The authors are of the view that such opinion targets have high chance of relatedness with frequently occurred targets. The proposed hybrid architecture showed improved results over existing techniques using semantic data.

Regarding semantic expansion, Gan & Hong [23] explored Wikipedia knowledgebase and three corpuses (CACM, ADI and CISI) to extract the terms relationships. They constructed a Markov network to select the relevant candidates for query expansion. Their experimental results showed that the proposed method outperforms the baseline model. Another application of query expansion includes word sense disambiguation, in which linguistic knowledge is exploited to select the correct word sense. For example, authors in [16] tested the use of WordNet (a famous thesaurus for providing word senses, e.g., set of synonyms) in query expansion. Their method achieved a 57% disambiguation rate using the standard expansion procedure.

However, 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 the 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. Overview of Approach

Our approach offers two expansion modes, namely, term sense disambiguation and semantic expansion. The former mode aims to solve the vocabulary mismatch problem, thus facilitating users to write textual queries in their own vocabulary. The latter expansion mode relies on ontology in selecting the concepts and relations that are relevant to user query. The rationale behind using two stages of expansion is that, the query itself makes the machine understand the user's demands.

The major steps in our approach are presented as follows:

Step1: Submit the initial query to the system as a set of terms Q.

Step2: Convert the initial query into a standardize query $Q' = \{q_1, q_2, q_3, . , q_K\}$ by removing the noise and stop-words, where $|Q'| = K$. For instance, query 'Algorithm for searching' can be represented as {{algorithm}, {searching}}.

Step3: Search thesaurus to extract set of senses $Sq'_i = \{s_1, s_2, s_3, \ldots , s_N\}$ for each term $q'_i \in Q'$, where $| Sq'_i| = N$ and $1 \geq i \leq$. K. Compute semantic similarity score for each sense in Sq'_i against q'_i and arrange senses in descending order of obtained scores. Moreover, include q'_i into corresponding Sq'_i as follows:

$$Sq'i = \{ \{q'i\} \cup \{s1, s2, s3, \ldots , sN\} \} \qquad (1)$$

Step4: For each element of Sq'_i, browse ontology to find number of relevant concepts and add them in set $C_i = \{c_1, c_2, c_3, \ldots , c_M\}$, where $|C_i| = M$. Thus, we obtain vector $\overrightarrow{C_i}$, against each q'_i (see Fig. 1).

Step5: Calculate the semantic similarity score of each element of $\overrightarrow{C_i}$ against corresponding q'_i. Fig. 1 illustrates that each concept is assigned a similarity score.

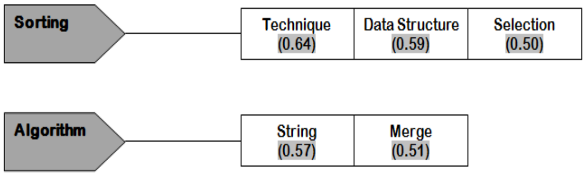

Fig. 1. Concept Vectors for Each Query Keyword.

Step6: Expand the user query Q′ with concepts that achieves high similarity score.

Step7: Submit the expanded query to the information retrieval system for results.

In the present work, we use WordNet thesaurus and computer science domain ontology to reformulate the initial query, in an attempt to understand user requirements in semantic manner.

A. WordNet Lexicon

Many researchers have focused on using the WordNet lexicon for query expansion work. The lexicon represents precise word relationships that are further categorized into 26 types, such as hyponym and synonym. Miller first introduced the WordNet lexicon in 1995 [17], while the latest available version is 3.1.

We use WordNet 3.1 and only focus on the synonymy relationships (namely, synsets) of the lexicon. These synsets provide a means for obtaining term senses to disambiguate user query.

B. Computer Science Domain Ontology

The use of computer technologies in our lives has caused the development of computer science as a distinct discipline. Computer science is an appealing discipline given the implementations that concern every aspect of life. Furthermore, this discipline has various sub-fields, such as database systems (the study of fundamental properties of relations and query processing) and programming languages (the study of approaches to describe problem-solving computations).

Ontology can be used in organizing the data in computer science discipline, thus enabling to browse relations among concepts semantically. We select the computer science domain ontology [18] developed at the University of Athens by Michael Sioutis and encoded in the web ontology language

(OWL). This ontology formally describes all branches of computer science (e.g., algorithms, artificial intelligence, programming languages, and data structures.) using relationships among these branches, such as hasPart and isPartOf.

Figure 2 depicts the portion of the graphical hierarchy of domain ontology. This portion indicates the concepts and their relations. For example, the programming methodology and languages concept is related to the computer science concept via isPartOf relationship.

We use computer science ontology to extract concepts that are semantically related to user search query. The search results in using such concepts when added to original user query are better than the original query.

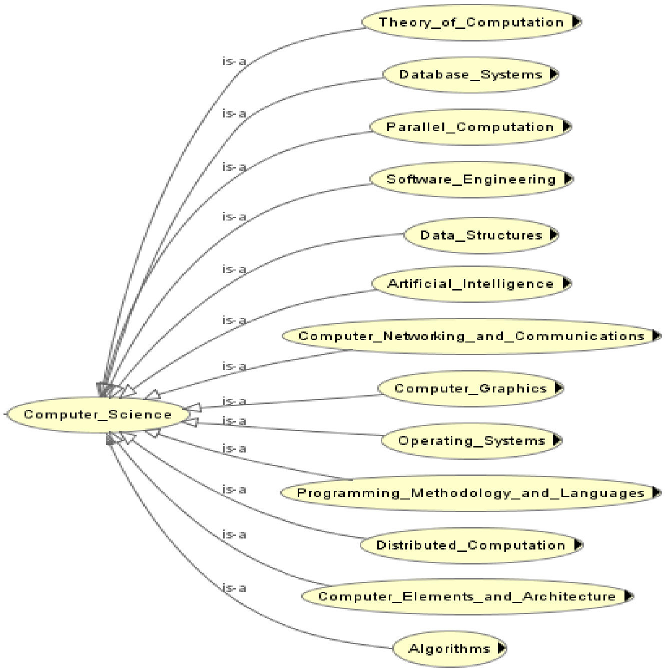

Fig. 2. OWLViz View of Computer Science Ontology.

IV. Query Expansion Framework

Figure 3 demonstrates an overview of the proposed framework for query expansion based on ontology. The framework is based on five main units, namely, user interface, query refinement, sensual semantic expansion, similarity inference, and query constructor components. These components are described in the following subsections, starting from the initial user query to generating final results. Our approach is based on expanding the initial query that focuses on sense disambiguation, computer science domain semantics, and use of semantic similarity method to filter the set of candidate expansion terms.

A. *User Interfaces*

The user poses initial search query Q and views the results returned by the information retrieval system via a user-view interface.

B. *Query Refinement*

Query refinement module adds several basic structures to the unstructured initial student query. The two basic operations of this component are tokenization and stemming. At the end of these operations, a pool of keywords (called standard query Q′) from the initial student query has become available.

1) *Tokenization*: Tokenization splits the query into words called tokens on the basis of a space character. Thus, we obtain two types of tokens, namely, word and space tokens. Furthermore, stop-words (e.g., the, a, and an) are removed from word tokens to obtain the query keywords.

2) *Stemming*: Stemming helps in identifying basic forms of query keywords by removing the affixes from each term. We use Porter stammer [19] to stem.

For example, for a given query Q = 'an algorithm for sorting', the query keywords after tokenization are {sorting, algorithm}, and the stammer provides us with the standard form of query Q′ as {sort, algorithm}.

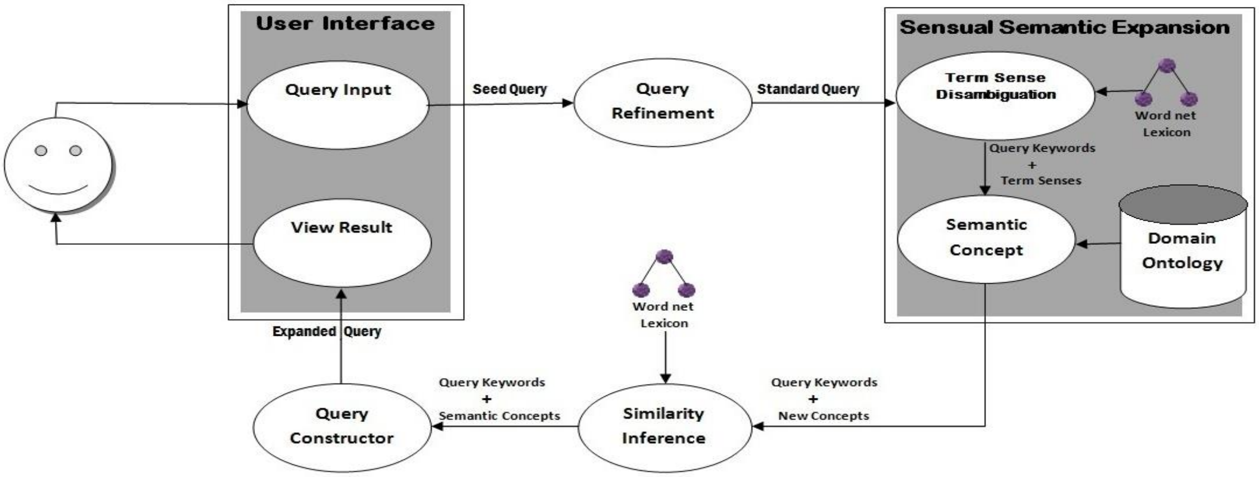

Fig. 3. Ontology based Query Expansion Model.

C. Sensual Semantic Expansion

Sensual semantic expansion (SEE) component supports the sense disambiguation of query Q′ and obtains a set of expansion terms (semantic concepts). The basic function is to match query term senses against computer science ontology concepts to find unstated concepts that imitate the user's interest. The SSE module processing is conducted via two phases.

1) Term sense detection: In this phase, multiple senses for each keyword of query Q′ are extracted via synsets of WordNet lexicon. These senses provide a means for referring to the same keyword in multiple ways. Moreover, this information helps in avoiding the retrieval of irrelevant candidate concepts for expansion.

2) Semantic concept identification: Given the query senses, this phase uses a knowledgebase (which in our case is computer science domain ontology) to discover the semantic concepts. Each sense of query keyword is matched with ontology concepts. If a match is found, then classes (concepts) related to matched ontology concept are extracted via hyponym and hypernym relationships.

Note that if match is found, then the SSE module omits ontology search for the remaining senses of a keyword; otherwise, it rejects a keyword. Algorithm1 describes the SSE sub-tasks. For each keyword of Q′, the algorithm obtains the sense vectors, and sense data are used to extract concepts that are related to initial query Q.

Algorithm 1 Algorithm for sensual semantic expansion

```
Input  :  Q′ , The set of query keywords
          DO, domain ontology
Output :  SC , Set of semantic concepts

FOR each keyword(i) in Q′
  // Term sense detection
  IF keyword is found in WordNet
    Compute synset for keyword(i) // set of synonyms
  END
  // Semantic concept identification
  FOR each sense(j) in synset
    IF sense(j) is found in DO
      // Traverse hypernym relationship
      // and hyponym relationship to one level
      Extract ontology concepts
      Add concepts in SC
      BREAK;  // omit search for remaining senses
    END
  END FOR
END FOR
```

D. Similarity Inference

The inclusion of query senses during term sense detection task may return computer science concepts (i.e., semantics) that are loosely related to user requirements. For example, in query Q of our example, the system must provide the concepts that represent different techniques of list sorting. Thus, the identified semantic concepts must be analyzed to check whether these concepts are representative of the original query Q or not.

In this phase, we adopt a semantic similarity measure for the following purposes: (1) to arrange the query senses on the basis of their scores and (2) to select among the set of candidate expansion concepts (with high scores) recognized by the SEE component. We use Zhou similarity measure [20] to evaluate the similarity score of each query Q′ keyword against corresponding senses and expansion concepts. The scores obtained using Equation (2) show the relatedness of concepts with query Q′. Moreover, we set k=0.5 to obtain a hybrid (i.e., path-based and information content-based) similarity value on the basis of the WordNet lexicon.

$$Sim\_score\ (w,c) = 1 - k \times A - (1-k) \times B \qquad (2)$$

Where

$$A = \left(\frac{\log(len(w,c) + 1)}{\log(2 \times \max(depth_w\ , depth_c) - 1)} \right)$$

And

$$B = \left(\frac{IC(w) + IC\ (c) - 2 \times IC\big(lso\ (w,c)\big)}{2} \right)$$

E. Query Constructor

The query constructor component formulates the expanded query. It receives the list of high-similarity-scored concepts and combines them with initial query Q to create an expanded query.

The query after the expansion is then automatically posted to the information retrieval system, which retrieves results for the user.

V. Experimental Results and Analysis

The above mentioned approach (called SSE) is implemented with a prototype of a system that uses tools, such as NetBeans, Jena API, and ARQ engine. We aim to retrieve the most relevant information for the users without requiring to navigate through irrelevant results. To obtain the search results, we use Atire search engine [21] as basic retrieval model. We evaluate our approach using Communications of the ACM (CACM) test collection, which consists of documents from the domain of computer science [22]. In addition, we have extracted 50 queries from the CACM topics and have selected the top 20 expansion terms in our experiments.

A. Evaluation Metrics

To measure the effectiveness of retrieval, we use two metrics, namely, mean average precision (MAP) and recall. The MAP indicates the accuracy of retrieved results, whereas recall denotes the completeness of results. We define these measures as follows.

$$MAP = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{R_i}\sum_{j}^{R_j} P(D_j) \qquad (3)$$

$$Recall = \frac{1}{Q}\sum_{i=1}^{Q}\frac{r_i}{R_i} \qquad (4)$$

Where Q is the number of queries, R_i represents total number of relevant documents for *i-th* query, P refers to precision, D_j is the *j-th* document from top retrieved set of documents and r_i reflects the number of relevant documents retrieved for *i-th* query.

B. Results Analysis

In the experiment, 80 computer science students (10 undergraduate, 30 graduate, and 40 postgraduate) were divided into 4 groups randomly. All of these groups were trained to retrieve results via the Atire search engine for queries that were expanded using our approach (SSE) and for the unexpanded queries (called baseline). In addition, the groups were required to record the score of relevant results (i.e., documents relevant to a given query) achieved using the SSE and baseline method. For the sensitivity analysis, the groups were requested to measure the relevancy score separately for the top 50 and top 100 retrieved results of each submitted query.

For the performance analysis, three measurement variants, namely, MAP@50, MAP@100 and Recall@100 were calculated. The difference among these measures is based on the analysis of the top N retrieved documents (where N can be 50 or 100). The results for baseline and SSE model in terms of MAP and Recall are summarized in Table 1.

TABLE I. COMPARISON IN TERMS OF MAP@50, MAP@100 AND RECALL@100 MEASURES VIA ATIRE SEARCH ENGINE

Evaluation Measures	Baseline	SSE Approach
MAP@50	0.15	0.25
MAP@100	0.11	0.23
Recall@100	0.48	0.74

The SSE approach achieved a considerable improvement over the baseline method in the MAP and recall values. When top 50 documents are retrieved, the SSE approach shows improvement about +10% in MAP as those of baseline method. The results trend is found very similar for top 100 retrieved documents. Moreover, we realize that SSE method can improve the recall measure about +26%, when 1000 documents are retrieved. This observation further confirms the effectiveness of proposed SSE system.

The SSE approach is better than the baseline method given the following reasons: (1) SSE leverages user query senses at the initial stage. Therefore, the new approach helps in identifying the correct sense for the original query terms. (2)The SSE method avoids including unnecessary expansion terms by considering the computer science domain ontology and semantic similarity method.

The efficiency of the SSE procedure was further evaluated using Google, which is a search engine that is widely used by users. Table 2 reports the results measured with MAP and Recall for both Atire based and Google based SSE. An interesting observation is that the performance improvement in the Google-based SSE method is similar to the Atire-based SSE method for top 50 (MAP@50) and top 100 (MAP@100) retrieved documents. Therefore, the SSE approach can stably improve the retrieval accuracy for the Web-based search. By contrast, the Recall@100 result for the Google-based SSE is less substantial than the method implemented in the Atire toolkit but still much better than the baseline method (i.e., +15%).

TABLE II. Comparison in Terms of Map@50, Map@100 and Recall@100 Measures Via Atire and Google Search ENGINES

Evaluation Measures	Atire Based SSE	Google Based SSE
MAP@50	0.25	0.24
MAP@100	0.23	0.21
Recall@100	0.74	0.63

Finally, the results were plotted in a 2Dchart for the MAP and recall values. Fig. 4 displays that the SSE approach and Google-based SSE search outperform the baseline method. The results trend indicates that SSE method achieved better Recall value when Atire retrieval system is adopted, in contrast to Google-based retrieval. We believe the main reason for this is that the pool of expansion terms is kept small in size (20 terms). In Atire-based SSE, the fewer expansion terms provides an effective guidance in selection of relevant results.

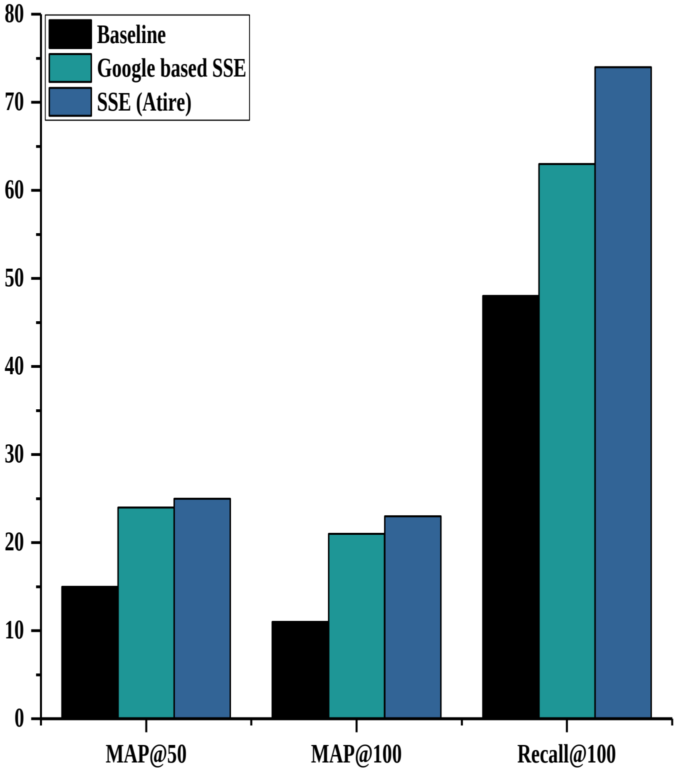

Fig. 4. Column Plots of MAP@50, MAP@100 and Recall@100 Measures.

VI. Conclusion and Future Work

In this work, we have addressed the problem of accurate search by focusing on reformulating the user query to become self-explanatory. We have proposed a sensual semantic framework for query expansion and have used semantic structures i.e., ontologies. In particular, the SSE method helps in extracting the correct sense of a query term from WordNet ontology. The semantic expansion process takes place by browsing computer science ontology for additional terms related to query terms and query senses. The generated expansion terms are then analyzed using similarity inference to select terms closely related to query senses. Experts have evaluated our prototype on the CACM collection and the Atire search engine. Our system has obtained the optimal results for MAP@50, MAP@100, and Recall@100 using test dataset. Moreover, we have tested the capability of SSE system on WWW using Google search engine. The difference between the Atire-based SSE and Google-based SSE methods for MAP@50 and MAP@100 is insignificant. This comparative analysis indicates that our approach is also useful in retrieving precise information from a diverse information pool of WWW. Our model has outperformed the baseline method, thereby indicating that the concept of query sense analysis along with semantic expansion can provide a breakthrough in retrieving relevant information for users.

Our future work includes enhancing our prototype for large standard ontologies (in contrast to domain-specific ontology) and making the prototype available to researchers to test its authenticity and detailed analysis in various domains.

References

[1] Carpineto, C., & Romano, G., “A survey of automatic query expansion in information retrieval,” *ACM Computing Surveys (CSUR), 44*(1), 1, 2012.

[2] Zhu, X.-H., Wu, T.-J., & Chen, H.-C., “An Interoperable Model for the Intelligent Content Object Based on a Knowledge Ontology and the SCORM Specification,” Journal of Educational Computing Research, 56(5), 723-749, 2018. doi: 10.1177/ 0735633117725764

[3] Raza, M.A., et al., “Query expansion using conceptual knowledge in Computer Science,” in 5th International Conference on Software Engineering & Computer Systems (ICSECS17), 2017. Langkawi Islang, Malaysia.

[4] Yonggang Qiu, and Hans-Peter Frei, “Concept based query expansion,” SIGIR conference on Research and development in information retrieval, pp. 160–169, New York, NY, USA, 1993.

[5] Waseem Alromima, Ibrahim F. Moawad , Rania Elgohary and Mostafa Aref , “Ontology-based Query Expansion for Arabic Text Retrieval,” International Journal of Advanced Computer Science and Applications(IJACSA), 7(8), 2016. http://dx.doi.org/10.14569/IJACSA.2016.070830.

[6] Xinhua, L., Xutang, Z., and zhongkai, L., “A domain ontology-based information retrieval approach for technique preparation,” Physics Procedia. 25:1582-1588, 2012.

[7] Nacim Yanes, Sihem Ben Sassi, and Henda Hajjami Ben Ghezala, “Ontology-based recommender system for COTS components,” Journal of Systems and Software, Volume 132, 2017, Pages 283-297, ISSN 0164-1212.

[8] Adeel Ahmed and Syed Saif ur Rahman, “DBpedia based Ontological Concepts Driven Information Extraction from Unstructured Text,” International Journal of Advanced Computer Science and Applications(ijacsa), 8(9), 2017. http://dx.doi.org/10.14569/IJACSA.2017.080954.

[9] Yuanfeng He, Yuanxi Li, Jiajia Lei, C.H.C Leung, “A framework of query expansion for image retrieval based on knowledge base and concept similarity,” Neurocomputing, Volume 204, 2016, Pages 26-32, ISSN 0925-2312.

[10] C. H. C Leung, and Alfredo Milani, “Collective evolutionary concept distance based query expansion for effective web document retrieval,” 2017.

[11] Bhawani Selvaretnam, and Mohammed Belkhatir, “A linguistically driven framework for query expansion via grammatical constituent highlighting and role-based concept weighting,” Information Processing & Management, Volume 52, Issue 2, 2016, Pages 174-192, ISSN 0306-4573.

[12] Bhogal, J., A. Macfarlane and P. Smith, “A review of ontology based query expansion,” Information Processing and Management 43(4): 866-886, 2007.

[13] Azad, H.K. and A. Deepak, “Query expansion techniques for information retrieval: a survey,” CoRR, 2017. abs/1708.00247.

[14] Mauro, N., et al., “Concept-aware geographic information retrieval,” in Proceedings of the International Conference on Web Intelligence 2017, ACM: Leipzig, Germany. p. 34-41.

[15] khan, K., Ullah, A., & Baharudin, B. “Pattern and semantic analysis to improve unsupervised techniques for opinion target identification,” Kuwait Journal of Science, 43(1), 129-149, 2016.

[16] Crimp, R. and A. Trotman, “Automatic term reweighting for query expansion,” in Proceedings of the 22nd Australasian Document Computing Symposium 2017, ACM: Brisbane, QLD, Australia. p. 1-4.

[17] G. A. Miller. “WordNet: A lexical database for English,” CACM 38, 11, 39–41, 1995.

[18] Sioutis M. “Computer Science Ontology,” Department of Informatics and Telecommunications, University of Athens, 22 June 2009. http://cgi.di.uoa.gr/~std04153

[19] Porter, M. F., “An algorithm for suffix stripping. In Readings in Information Retrieval,” K. S. Jones and P. Willett Eds., Morgan Kaufmann, 313–316, 1997.

[20] Zhou, Z., Wang, Y. and Gu, J., “New model of semantic similarity measuring in Wordnet,” in Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, 2008, November 17-19, Xiamen, China.

[21] A. Trotman, C. L. A. Clarke, I. Ounis, S. Culpepper, M.-A. Cartright, and S. Geva, “Open Source Information Retrieval,” A Report on the SIGIR 2012Workshop. SIGIR Forum 46, 2, 95–101, 2012.

[22] Salton, G., Fox, E. A., & Wu, H., “Extended Boolean information retrieval,” Commun. ACM, 26(11), 1022-1036, 1983. doi: 10.1145/182.358466

[23] Gan, L., & Hong, H., “Improving Query Expansion for Information Retrieval Using Wikipedia,”. *8*(3), 27-40, 2015. doi: 10.14257/ijdta.2015.8.3.03.