International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 14s,2025
https://theaspd.com/index.php

Formulation Of E-Content Search Approach Using Thesaurus Based Ontology For Engineering Education Environment

Kamaljeet Kaur Mangat[1], Dr. Amandeep Verma[2]
[1]Punjabi University Centre for Emerging and Innovative Technology, Mohali, Punjab, India, Kamalmangat@pbi.ac.in
[2]Punjabi University Centre for Emerging and Innovative Technology,Mohali, Punjab, India, Vaman71@gmail.co

Abstract
With the advancement of technology and digitization of educational resources there is a plethora of information available on the world wide web. Many a times the search process does not yield the desired relevant results against the query posed by the user. In academic and technical context, the user is unable to retrieve domain specific relevant content because of vocabulary mismatch, lack of semantic understanding, and insufficient contextual information in search systems. In order to address these limitations, the present paper has leveraged the capabilities of a technical thesaurus using ontological engineering to expand the query terms along with the capabilities of NLP techniques to refine and reformulated the user queries in engineering education environment. The active role of users in the evaluation process has helped to achieve the relevant, diverse and user-satisfactory results.

Keywords: *Query Expansion, IEEE Thesaurus, e-content, Ontological Engineering, NLP, Searching*

INTRODUCTION

There is a significant rise in the volume of online educational material with the rapid expansion of digitization. The learners from all of the different faculties, irrespective of their subject background, access educational material through various digital platforms such as search engines, digital libraries, databases etc. Still the retrieval of relevant content remains a challenge because of unstructured and heterogeneous nature of e-content resources. The Traditional e-content search depends upon the keyword-based retrieval models, where user queries are matched against the indexed terms in documents, The Boolean model and vector space models[1] provide simplicity and computational efficiency but they suffer from vocabulary mismatch and semantic relatedness problems. The use of ontologies in semantic search models has transformed e-content retrieval as they understand the context and structure of the terms posed in the user query for searching. In addition to this, the incorporation of NLP based steps such as preprocessing pipeline, keyword extraction and n-gram creation further help to refine and reformulate the query. The use of query expansion techniques such as synonym substitution and ontology-based enhancement using thesauri bridge the gap between user vocabulary and document language [2]. The focus of the proposed approach is on automatic query expansion. The statistically driven approaches are considered to be better as compared to those based on generic ontology [27]. This is because of the ambiguity of query terms based on synonymy. Since the present problem is related to a specific application, that is, engineering education, ambiguity handling is less difficult due to specific terms as in[3] for geographical terms. Thesaurus is an organized type of controlled vocabulary that can be adopted as domain specific ontology[4]. The use of thesaurus as an ontology for retrieving content for a particular domain has been demonstrated in various fields such as healthcare [5], agriculture [6] etc.

The section II presents the background study related to ontological engineering .Section III presents the design methodology and the and formulation of the query expansion of the proposed work.

BACKGROUND

Ontological Engineering

Ontological engineering is an efficient method for the knowledge representation and management in various domains[7]. The purpose of ontological engineering[8] is to “provide a basis of building models of all things in which computer science is interested". Ontologies range from lexicons, to dictionaries and thesauri and further

to first order predicate logic theories in computer science[9]. In any of these forms, ontologies encourage standardization of the terms used to represent knowledge about a domain. Qin and Paling [10] demonstrated how the GEM controlled vocabulary was transformed into a more expressive ontology, offering a structured approach to handle multidimensional metadata in educational materials. Hilera et al.[11] introduced an evolutive method that begins with glossaries and systematically transforms them into taxonomies, thesauri, and eventually full ontologies. Their methodology underscores the increasing semantic expressiveness gained at each transformation stage. Similarly, Castilho et al. [12] combined corpus analysis with WordNet to generate a domain-specific thesaurus for data privacy and used it to enrich existing ontologies. Their approach bridges natural language processing and ontology modeling through automatic semantic expansion. In practical applications, Sandhu et al.[13] applied a domain thesaurus in a case-based reasoning system for detecting dengue cases. The use of a keyword-aware thesaurus facilitated precise case identification and proved to be extensible to other healthcare applications. Thesauri, when transformed into ontologies, provide a semantically enriched foundation for query expansion. Unlike traditional thesauri that offer lexical-level expansions through predefined relationships like synonyms (USE/UF), broader terms (BT), and related terms (RT), ontologized thesauri encode these relations with formal semantics. From controlled vocabularies to ontologies, each transformation enhances the semantic clarity and interoperability of information systems. When integrated into IR frameworks, thesauri not only support richer indexing and retrieval capabilities but also serve as intermediaries in ontology engineering, knowledge sharing, and domain-specific reasoning. Such integration has been shown to significantly improve recall and precision in domain-specific IR tasks, particularly when using resources like MeSH [5], AGROVOC [6], or WordNet. In the context of e-content searching, ontologies enable semantic search, thereby allowing the systems to understand the meaning behind user queries rather than relying solely on keyword matching.

PROPOSED APPROACH

The integration of Natural language processing techniques in IR help to understand the structure and meaning of content and thereby advance the searching process[14]. The present study use thesaurus as ontology for reformulation of query. The user query is refined using NLP techniques and reformulated by undergoing ontology based query expansion.

Design Methodology

The various steps involved in the formulation of e-content search approach are shown in figure 1 and the corresponding NLP tasks in preprocessing pipeline are shown in figure 2.

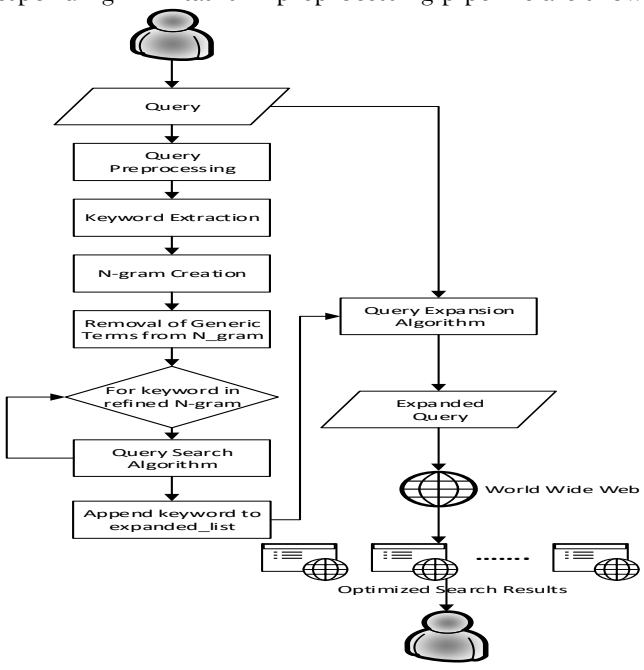


Figure 1: e-content search approach

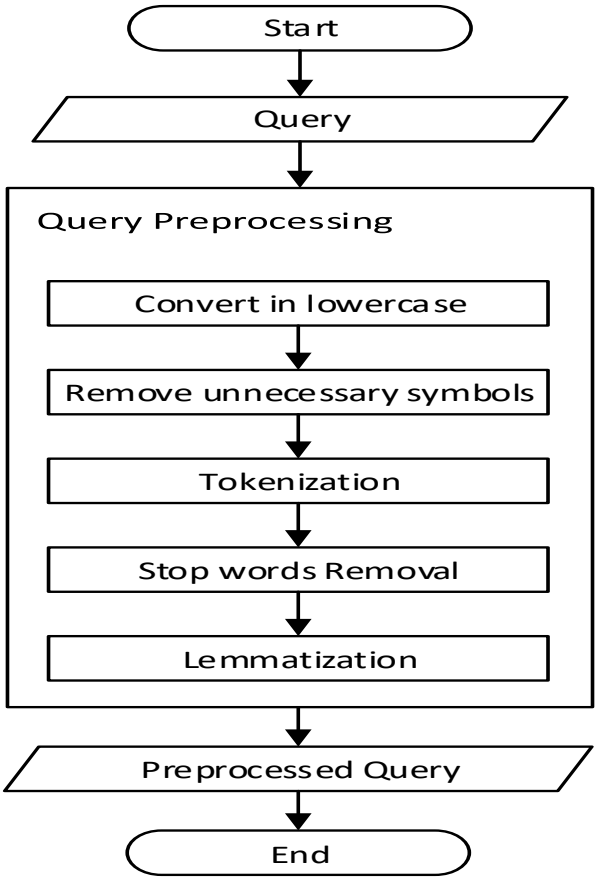


Figure 2: NLP Preprocessing pipeline

Formulation of Query Expansion

The condition was set as per the subject domain expert for expansion of input term. The same has been demonstrated in [15].

Condition:

If Query term matches with TOP TERM in thesaurus, then

Query expanded with Narrow terms

Else if Query matches with Narrow Term then expansion term is top term

The initial input query with a few keywords is reformulated with similar significant terms referred to as query expansion[16]

If a user query consists of single term, such that, $Q = \{ t_i\}$

Then $Q_{exp} = QUT'$

where $T' = \{t'_1, t'_2, \ldots . t'_m\}$ is the set of new terms that are added from Data source DS.

The key aspect in the query expansion is the set T', that append the original query term with additional significant terms to enhance the relevance of search results. The choice of the data source is another significant factor for query expansion. The data source for the present work is the domain specific thesaurus of engineering domain, where the original query term is enhanced by the semantically corresponding terms.

The new terms against a given query are extracted from the data source here, IEEE thesaurus based on the condition such that if query term matches with top term in thesaurus, then the resultant query expand the given query term with the corresponding narrow terms; otherwise, if query matches with narrow term, then it is appended with top term. If there is no narrow term corresponding to a given query term then it is expanded with related terms.

$Q = \{q\}$

DS= {BT, TT, RT, NT}

if
$q \leftarrow TT : q \leftarrow q + NT$
$else$
$q \leftarrow NT : q \leftarrow q + TT$
if
$NULL \leftarrow NT : q \leftarrow q + RT$
$returnQ$
If a user query consists of n terms, such that

$$Q = \{ t_1, t_2, \ldots . , t_n \}$$

then the expanded query is given by, $Q_{exp} = (Q - T'')\ UT'$
Where the stop words $T'' = \{t_{i+1}, t_{i+2}, \ldots \ldots t_n\}$ are removed.
The English lexical database, WordNet has also been used for semantic relatedness among the generic words based on their similarity measure to refine the query. Now the expanded query is given by

$$Q'_{exp} = (Q - T'')UT_w UT'$$

Where $T_w = \{t_{w1}, t_{w2}, \ldots \ldots . . t_{wn}\}$ is the set of relevant WordNet synsets,
$T_w = T_W - T_x$, where T_W is set of all synsets in WordNet corresponding to a word,
$T_x = \{t_{x1}, t_{x2}, \ldots \ldots . t_{xn}\}$
whose similarity measure is below threshold value and when the query posed contain the domain specific term along with generic terms then the posed query is checked with domain ontology first and then with the generic ontology viz wordnet for finding more terms. The pseudocode for query expansion is given in *algorithm 1* that use the descriptive terms such as Top Terms (TT), Narrow Terms (NT) and Related terms (RT) from the thesaurus.

Algorithm 1: Query Expansion algorithm

$Q = \{q\}$
if $q_i \leftarrow TT : q_i \leftarrow q_i + NT$
$else$
$q_i \leftarrow NT : q_i \leftarrow q_i + TT$
if NULL $\leftarrow NT : q_i \leftarrow q_i + RT$
$if\, q_j \in T_w : q_j = q_j + t_w$
if
$t_{wi} = sim(t_{wj}) : q_j \leftarrow t_{wi} + t_{wj}$
$q_i + q_j \leftarrow q$
$return\ Q$

IMPLEMENTATION

Data Simplification Tool

The first step is toward the development of novel data simplification tool for the XML file creation of the given thesaurus[17].

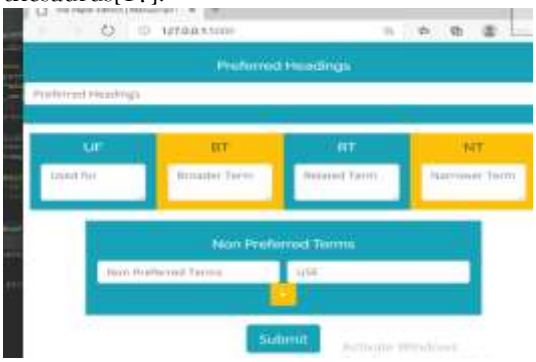


Figure 3: Snippet of Simplification tool for XML creation

A snippet of the interface for the same is given in *figure 3*. The XML file is accessible by the search tool to expand the given user query to generate more meaningful terms.

Search Tool Development

The proposed approach is implemented using following tools and technologies as shown in *Table 1*.

Table 1: Implementation Tools& Technologies

TASK	Tools Utilized
IDE	PyCharm
Programming Language	Python
NLP Library	NLTK
Text Preprocessing	RegEx
Web framework	Flask
Front End Development	HTML, Javascript
CSS Framework	Bootstrap

The steps undertaken for the development of technique for the e-content search are (i) Query Admission where the user poses its query in normal English language for searching the results. The query can be a term-based query or sentence-based query posed by user. (ii) Query Preprocessing phase includes the conversion of text to lowercase, removal of unnecessary symbols, tokenization, stop word removal, and lemmatization. It is the process of preparing a user's search query before it is used by a search system. It transforms the query into a cleaner, standardized, and often more effective form to improve retrieval accuracy and relevance.(iii) Keyword Extraction is done using Rapid Automatic Keyword Extraction algorithm(RAKE)[18], which is a domain independent keyword extraction algorithm that determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurrence with other words in the text.(iv) n-gram creation involves taking a sequence of text and breaking it down into all possible sub-sequences of a specified length (n) [19]. For instance, unigrams (n=1) consider individual words, while bigrams (n=2) and trigrams (n=3) capture short-range dependencies between words. (v) Generic terms are removed from the query by using lexical database WordNet[20] to refine the terms for expansion. For query expansion , the XML file built from ontology is used and the algorithm from the formulation of query expansion is implemented. A snippet of user interface is shown in figure 4.

Figure 4: User interface for Query Input

EXPERIMENTAL SETUP AND RESULTS

[21] proposed an entropy measure that is used to assess that how uniformly the query terms are distributed across documents. Top ranked results have higher entropy and are more related to the query. Therefore, it has been decided for the present study that only top 10 documents returned by google search to be considered as results of the input query. There are two experiments in this study, one is for the simple query posed by the user in comparison with the reformulated query after query expansion. The second experiment aim to evaluate and compare the relevance of the resultant links in the interactive web search that are retrieved from the original query and expanded query.

Selection of Queries

The queries pertaining to identified domain are selected. There are total of 30 queries, 10 each for a particular category as shown in *Table 2*.

Table 2:Selection of Queries

Category	Description	Number of Queries
D1	CSE	10
D2	ECE	10
D3	ME	10
Total		**30**

For evaluation purpose total 30 queries were selected from three branches of engineering education domain. Each category contains 10 sentence-based queries.

The results are formulated as under in equation (i)

$$xQn = \{Q_s, Q_o\}$$

$$x = \{C, E, M\} \in D, Domain \ \ and \ n = 1 \ to \ k$$

$$\forall \ xQn \ \exists \ R \ where \ R = \{r_1, r_2 \ldots\ldots, r_k)$$

.....equation (i)

Top *k* results are retrieved for each *xQn*. Where k=*10*

In order to reduce the subjective bias to the expected results and make the results more relevant, the binary relevance method and graded relevance methods are used.

Users Selection for Assessment of Search Results

User evaluation[22] has been conducted to confirm that whether the results are accepted by real users. Questionnaires were used to obtain the feedback from the user. The questionnaire contains the Title and URL of results of simple query and expanded query along with the columns for binary and graded relevance assessment to be filled by user.

[23] take into account varying degrees of relevance of retrieved documents, rather than treating them as simply "relevant" or "not relevant" (binary relevance). These metrics aim to reflect the user's satisfaction by assigning weights or scores to different levels of relevance (e.g., marginal, fair, high). The results are ranked by the user experts on graded relevance score. The results corresponding to simple as well as optimized query are given graded scores as per the user level expertise for each of the retrieved link. The participants included the faculty at different levels, students of PG and UG courses and engineers as shown in *table 3*. They were told to rank the documents on the scale of 0 to 3 (0: not relevant,1: somewhat relevant, 2: relevant, 3: highly relevant)

An instance of $R \in xQn$ is given to the users for computing relevance level.

Table 3: Users for Evaluation

User	Background	Number
Industry	Engineering, Management	6
Faculty	Engineering	4
	Non-Engineering	6
Student	PG-Sciences	3
	UG-Engineering	5
Total		**24**

The one set questionnaire contains three queries each from different domain and grouped together for evaluation by each user. All of the 24 evaluators provided the feedback for 72 queries and rest of the 18 queries were assessed by 6 users from industry, thereby evaluating 90 queries. This way each query is evaluated by three users and accordingly relevance is evaluated and subjected to further treatment.

Performance Metrics

The performance criteria to evaluate the queries is based on precision, Sub-topic recall and ERR thereby checking the relevance, diversity and user satisfaction of the results.

Precision

Precision is a metric that measures the accuracy of the retrieved documents. It is defined as the proportion of retrieved documents that are relevant to the query.

Precision is given by

$$Precision = \frac{Number\ of\ Relevant\ \ Retrieved}{Total\ number\ of\ Retrieved}$$

Sub-Topic Recall

S-Recall measures the diversity of the results as resultant link cover a range of topics after the expansion of the query. It measures the proportion of unique subtopics covered by top k links by the total number of relevant subtopics known for the query.

Sub-topic recall is given by

$$s - recall = \frac{Number\ of\ unique\ subtopics\ covered}{Total\ number\ of\ known\ subtopics}$$

It is a graded relevance measure and relevance score is given as (0: *not relevant, 1: Somewhat relevant, 2: Relevant, 3: Highly relevant*)

Expected Reciprocal Rank

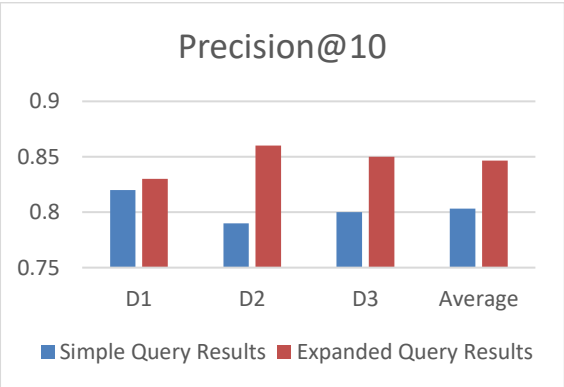
Since the evaluation of results is user centric, another parameter Expected Reciprocal Rank (ERR) models a user examining results sequentially and stopping at a satisfactory document, accounting for graded relevance.

Expected Reciprocal Rank (ERR) is given by

$$ERR@k = \sum_{i=1}^{k} \frac{1}{i}.P(i)$$

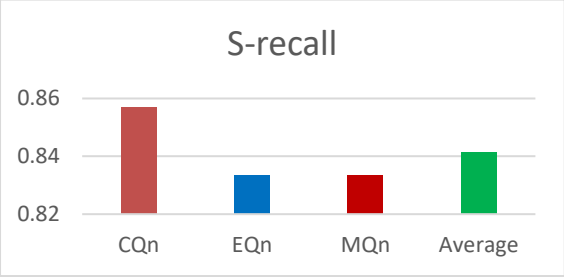
RESULTS AND DISCUSSION

Precision@k: The results of evaluation of precision@k are shown in chart 1for 10 queries each for respective domain. It returns the relevant links retrieved against the total results returned by the system. The results show that there is no significant influence by the type of domain of input query for searching. The results of simple query yield an average precision of 0.8033 where in for the expanded queries gave 0.8467 thereby increasing the performance of system by 4% as shown in the graph 1.



Graph 1: Precision@10 for all domains and average

S-Recall measures the diversity of the results as resultant link cover a range of topics after the expansion of the query. It measures the proportion of unique subtopics covered by top k links by the total number of relevant subtopics known for the query. The results contain a total of 101 unique subtopics out of which 85 are covered in the resultant links thereby giving a coverage of more than 84% of the topics as shown in *graph 2*.



Graph 2: Subtopic recall of all domains and their average

Since the evaluation of results is user centric, another parameter Expected Reciprocal Rank (ERR)[24] models a user examining results sequentially and stopping at a satisfactory document, accounting for graded relevance. The *graph 3* shows the average ERR for simple queries is 74% and for the expanded ones is around 80%.

Graph 3: Average Expected Reciprocal Rank

The results signify that there is an improvement in the evaluation parameters for the reformulated query. The XML file built from a standard thesaurus enhanced the initial search query with more meaningful terms that yield better results.

CONCLUSION AND FUTURE WORK

The present work aimed to expand the initial user query with the help of ontological engineering approach in the area of engineering education domain. There is a significant contribution of different users in this study. They were actively involved in the selection and building of dataset, selection of queries and evaluation of results. The role of ontological engineering for building domain knowledge is the major contribution of the study. Since the dataset belongs to a specialized domain then the evaluation parameter technicality can be explored for future work.

REFERENCES

[1] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Inf Process Manag*, vol. 56, no. 5, pp. 1698–1735, 2019, doi: 10.1016/j.ipm.2019.05.009.
[2] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, vol. 27, no. 3, pp. 129–146, May 1976, doi: 10.1002/asi.4630270302.
[3] D. Buscaldi, P. Rosso, and E. S. Arnal, "A WordNet-based Query Expansion method for Geographical Information Retrieval," in *Workshop on GeoCLEF*, Vienna, Austria, 2005.
[4] D. Kless, L. Jansen, and S. Milton, "A content-focused method for re-engineering thesauri into semantically adequate ontologies using OWL," *Semant Web*, vol. 7, no. 5, pp. 543–576, 2016, doi: 10.3233/SW-150194.
[5] M. Van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga, "A Method for Converting Thesauri to RDF / OWL," *Springer Verlag*, pp. 17–31, 2004.

[6] F. Amarger, J. P. Chanet, O. Haemmerlé, N. Hernandez, and C. Roussey, "SKOS Sources Transformations for Ontology Engineering: Agronomical Taxonomy Use Case," *Communications in Computer and Information Science*, vol. 478, no. November, pp. 314–328, 2014.
[7] S. Cakula and A. M. Salem, "E-Learning Developing Using Ontological Engineering," *WSEAS Transactions on Information Science and Applications*, vol. 10, no. 1, pp. 14–25, 2013.
[8] R. Mizoguchi and M. Ikeda, "Towards Ontology Engineering," *Japanese Society of Artificial Intelligence*, 1998, [Online]. Available: https://www.researchgate.net/publication/266883056
[9] I. Jurisica, J. Mylopoulos, and E. Yu, "Ontologies for Knowledge Management : An Information Systems Perspective," *Springer-Verlag Knowledge and Information Systems*, no. August 2003, pp. 380–401, 2004, doi: 10.1007/s10115-003-0135-4.
[10] J. Qin and S. Paling, "Converting a controlled vocabulary into an ontology: The case of GEM," *Information Research*, vol. 6, no. 2, 2001, doi: 10.1108/07419051111145118.
[11] J. R. Hilera, C. Pagés, J. Javier Martínez, J. A. Gutiérrez, and L. De-Marcos, "An Evolutive Process to Convert Glossaries into Ontologies," *Information Technology and Libraries*, pp. 195–204, Dec. 2010, [Online]. Available: http://www.genomicglossaries.com/content/ontolo
[12] F. M. B. M. Castilho, R. L. Granada, B. Meneghetti, L. Carvalho, and R. Vieira, "Corpus+WordNet Thesaurus Generation for Ontology Enriching," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, 2012, pp. 257–262. [Online]. Available: http://www.cpca.pucrs.br/VisualizationTool/Resource/Corpus.
[13] R. Sandhu, J. Kaur, and V. Thapar, "An effective framework for finding similar cases of dengue from audio and text data using domain thesaurus and case base reasoning," *Enterp Inf Syst*, vol. 12, no. 2, pp. 155–172, Feb. 2018, doi: 10.1080/17517575.2017.1287429.
[14] R. John and S. S. Govilkar, "Survey of Information Retrieval Techniques for Web using NLP," 2016.
[15] A. Shiri and C. Revie, "Query Expansion Behavior Within a Thesaurus-Enhanced Search Environment: A User-Centered Evaluation," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 4, pp. 462–478, 2006, doi: 10.1002/asi.
[16] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Inf Process Manag*, vol. 56, no. 5, pp. 1698–1735, 2019, doi: 10.1016/j.ipm.2019.05.009.
[17] The Institute of Electrical and Electronics Engineers (IEEE), *2019 IEEE Thesaurus*. 2019.
[18] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in *Text Mining: Applications and Theory* , M. W. Berry and J. Kogan, Eds., Chichester, U.K.: Wiley, 2010, pp. 1–20.
[19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511809071.007.
[20] G. A. Miller, "WordNet: A Lexical Database for English," *Commun ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[21] B. He and I. Ounis, "Studying Query Expansion Effectiveness," in *European Conference on Information Retrieval 2009 LNCS-5478*, B. Mohand, C. Berrut, J. Mothe, C. Soule, and Dupuy, Eds., Springer-Verlag Berlin Heidelberg, 2009, pp. 611–619. doi: https://doi.org/10.1007/978-3-642-00958-7_57.
[22] S. Plansangket and J. Q. Gan, "A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search," *Artif Intell Res*, vol. 4, no. 2, pp. 119–125, 2015, doi: 10.5430/air.v4n2p119.
[23] J. Kekäläinen and K. Järvelin, "Using Graded Relevance Assessments in IR Evaluation," *Journal of the American Society for Information Science and Technology*, 2002.
[24] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM conference on Information and knowledge management*, New York, NY, USA: ACM, Nov. 2009, pp. 621–630. doi: 10.1145/1645953.1646033.