Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Ontology Framework in Information Retrieval Using Semantic Query Expansion Based on Feature Optimization and Machine Learning

Rupali R. Deshmukh, Dr. Anjali B. Raut

Computer Science and Engineering, H.V.P.M's COET Amravati, Amravati, India
Email: drupali1604@gmail.com

Information Retrieval (IR) systems aim to fetch relevant information based on user queries. Data and information stored online is steadily growing in volume. Indeed, when it comes to readily available, high-quality information, the web is unrivalled. By analyzing the user's query semantically, semantic information retrieval can offer results that are more relevant to their original intent. When this occurs, ontology-based knowledge representation has the potential to facilitate semantic retrieval more efficiently than alternative representation approaches. In the past twenty years, machine learning techniques have found widespread application across several fields of art, science, and technology. Information retrieval activities that make extensive use of machine learning techniques include document analysis and query expansion. Traditional IR methods often suffer from limited context understanding, leading to irrelevant results. Semantic IR systems, enhanced by ontology frameworks and feature optimization, address these challenges by incorporating domain knowledge and contextual relationships. By combining semantic query expansion, feature optimization, and machine learning, these systems can significantly improve the relevance and accuracy of retrieved results. This paper presented an ontology framework in semantic information retrieval system using query expansion based on feature optimization and machine learning. The proposed system offers a robust solution to improve the accuracy, relevance, and efficiency of search results. Our research demonstrates that the method outperforms the state-of-the-art on certain recall, precision, and f-measure metrics used to evaluate the effectiveness of semantic information retrieval. Experiments have proven that the method helps the system convey knowledge better and retrieves semantic information faster.

Keywords: Feature Optimization, Ontology, Machine Learning, Information

retrieval, Semantic Information.

1. Introduction

Millions of individuals are checking out online sites to fulfil their information demands, making them the biggest source of information. Information retrieval (IR) systems aim to give consumers highly relevant results based on their information demands. For better query mapping to a collection of documents during retrieval, keywords describing the contents of the documents are assigned during the document indexing stage. Numerous proposals in the literature have proposed statistical, neural, and semantic methods to enhance the efficiency of data retrieval tasks [1-2]. Concepts, roles, and connections may be more accurately represented with the help of ontology in semantic web, a method of knowledge representation. The current state of knowledge representation allows for the easy encoding of textual information with inheritance using methods like semantic networks and frames. Semantic network or frame searches may therefore follow class and sub-class hierarchies all the way from the root node to the target node. In semantic networks, however, there are limitations on both the amount and kind of linkages. Both the information hierarchy and the inheritance hierarchy may be preserved using the ontology-based approach to knowledge representation. Information may be retrieved utilizing both single-level and multi-level inheritance using this ontology-based representation of data. Semantic analysis and appropriate information retrieval are therefore improved by semantic information retrieval based on ontologies [3-4].

Traditional information retrieval methods index documents according to the phrases they contain, not the notion that describes them. The phrase "vocabulary gap" explains the circumstance where users and subject matter experts use disparate terminology to refer to the same subject; IR systems will not provide results that contain semantically significant information unless there is a lexical match between the two. Document representation in an IR system has been shown to effectively capture term meanings and their connections utilizing domain knowledge [5][7]. Ontologies have long been used in all fields to formally express and comprehend domain knowledge. In contrast, language gaps are no longer an issue with ontology-based document indexing techniques. However, these approaches have the drawback of being very dependent on the depth and breadth of the initial input ontology. [10][12]. Recent advances in machine learning (ML) have allowed the field of information retrieval (IR) to provide a workable solution to document indexing challenges. When co-occurring events occur, machine learning algorithms leverage word-to-word correlations; nevertheless, they ignore significant semantic relationship structures. Manually created knowledgebases that include these relational structures for semantics include ontologies and semantic lexicons. The benefits of ML modelling have prompted academics to concentrate on developing ML ranking models that can learn features and the model in tandem [15][17][20-22].

Ontology-based retrieval systems retrieve data when there is a semantic overlap between the user's query and the indexed data. Creating a semantic representation of the data and associated domain knowledge is one of the key advantages of using a domain ontology. Furthermore, links between different areas of semantic knowledge can be specified using

ontologies. Ontologies can therefore be included in the creation of diverse data-searching algorithms. The goal of this work is to give more detailed semantic indexing of scientific publications than only the Computer Science Ontology descriptions. It aims to reach results comparable to those produced by domain experts by studying the semantic level of matching ideas. This research proposes a semantic information indexing strategy for information retrieval that is built on ontologies and machine learning to address this issue. Word embeddings and ontology from computer science are utilized by the suggested method to extract the notion from a text. The suggested method supplies the ML model with potential keywords extracted from document text using natural language processing techniques. The proposed algorithms improve upon previous efforts at retrieval speed and accuracy and leverage ontology matching to provide users with answers that are more pertinent to their queries. The most significant insights from the research are listed below:

- Developed a semantic approach to machine learning and ontology-based information retrieval system for document indexing.
- Designed an approach to semantic information retrieval by combining ontological knowledge with feature optimization and machine learning which offers a robust solution for improving query accuracy and user satisfaction.
- Improved the text-based information system's effectiveness with feature optimization and machine learning methods.
- Implement a ranking based approach to retrieve relevant documents for a user query.
- Conducted the trials with the TREC-NIST dataset, taking into account several performance parameters including F-measure, recall (R), mean average precision (MAP), and precision (P).

This paper is organized as follows: A summary of earlier ontology-based semantic information retrieval and machine learning research is provided in section 2; The proposed materials and procedures are explained in section 3. The result discussion and experimental findings are presented in section 4. This paper's conclusion is defined in the last section.

2. Related Work

Here, the key concerns expressed regarding the IRS highlighted. The IR techniques that have been proposed are mainly divided into two groups: supervised and unsupervised. To make sure the system can adapt to new domains as they emerge, machine learning is used to fine-tune the ontological structures dynamically [1]. A new approach to semantic document indexing was presented [2] that combines machine learning with domain ontology. An examination of the framework for categorizing data, and a thorough categorization according to the deep learning technique presented [3]. An enhanced variant of the original COOT algorithm is the IAOCOOT technique for query expansion, which finds the semantic characteristics that correspond to the search term [4]. Using a multi-terminology, Bayesian networks (BNs), and multinomial naive Bayes classifier (MNBC) is a novel way to enhance information retrieval systems [5]. Ontology-based information extraction techniques for graphical media, textual and other data languages are examined [6]. With relevant feedback,

the novel model expands the question by utilizing a deep neural network methodology and a semantic way to find the semantic similarity between the phrases [7]. Created an AI system for document management based on ontologies [8]. The development of a new method for semantic document indexing using a Shallow Neural Network and ontologies [9]. Clustering, rough set, and Bayesian network (BN) are used as an ensemble strategy to create the MLK-rBO model, which serves four different purposes: model assessment, probabilistic network development, knowledge discovery and clustering [10].

The augmented fish swarm method is recommended to carry out the retrieval process in order to improve the effectiveness of semantic IR [11]. An automatically sorted, comprehended, searched, and summarized web material search engine was developed in order to increase relevance scores in AR domains [12]. A retrieval system based on ontologies has been developed, which uses domain ontologies to change user queries at the outset and applies semantic association during indexing [14]. To rank objects, similar to other retrieval tasks, document retrieval uses a collection of components from neural networks to extract characteristics [15]. A novel fuzzy ontology-based framework is proposed to retrieve information utilizing domain-specific knowledge utilized in ontology construction [16]. The machine learning models were trained and tested on the manually collected corpus of scientific papers utilized in the Exposome-Explorer [17]. The several text representations offered, are accountable for retrieving pertinent search results, methodologies, and evaluations performed in conceptual information retrieval [18]. The ontologies are included in machine learning algorithms and utilized to compute similarity [19]. Semantic ontology information fusion was used to build the distribution fusion model of cross-language information between English and Chinese from the perspective of systemic functional linguistics [20].

An enhanced technique for semantic information retrieval was introduced [21]. The information retrieval system i-DATAQUEST [22] may be used to retrieve the processes and experiments indicated above as a set of runnable notebooks. A thorough and multidimensional benchmarking standard for IR is offered in the form of Benchmarking-IR (BEIR) [23]. A semantic multi-stage search engine called CO-Search was created to manage intricate searches across the COVID-19 literature [24]. The majority of IR techniques have been used on Malay literature by employing the categorization system to assess their benefits and limitations [25]. A survey on different models performed for Information Retrieval and Question Answering [26]. A standardized framework for modifying the Multi-Perspective IR system was offered, along with a novel method that combines several deep learning and traditional IR models to more accurately predict the relevance of a query-sentence combination [27]. A neural framework especially created for IR is the Semantic-Aware Neural Framework for IR (SAFIR) [28], which is unsupervised and knowledge-enhanced. Agnostic prediction models are explained by a new interpretation framework that develops an interpretable model using an ontology-based sampling method [29]. The effectiveness of automatic query expansion for queries in closed collections of text documents may be enhanced [30] by a posterior filter created using the collection's restricted vocabulary.

Context-based semantic matching (SCSM) is a popular method of establishing a connection between queries and documents [31]. A hybrid strategy for query expansion was proposed [32], based on a semantic and statistical approach based on particle swarm optimization

(PSO). The goal of this semantic video retrieval system is to find appropriate films based on user-supplied queries made in natural language [33]. Multi-Perspective Sentence Relevance is a new approach to contextual information retrieval (IR) that makes use of BERT-based models. It has several applications in the field of biomedical semantics [34]. There is a new system for automatically generating ontologies that do not depend on any particular domain; it may transform a collection of unstructured texts into an ontology that is compatible with that domain [35]. An innovative fuzzy rough set-based intelligent selection of features and categorization method is proposed for semantic information retrieval to enhance the relevance score [36]. Domain ontology, a constraints-based mapper, and point-of-sale (POS), and language processing tools are all integrated into the innovative model [37]. Integrating lexical and semantic data to handle CLIR (cross-language information retrieval) [38]. The principles and methods for ontology-based information retrieval described [40] are based on considerations for ontology modeling, processing, and the conversion of ontological knowledge into database search queries. The paper explained how semantic similarity measures and ontology embeddings might exploit this background knowledge, and how ontologies may provide restrictions that improve machine learning models [41]. Document clustering using a deep ontology-based method was suggested [42].

The limitations in existing research work include manual ontology construction and maintenance, contextual understanding, feature optimization challenges, intelligent learning overfitting and model bias, dependency on quality of training data and high computational complexity in training. Addressing these limitations requires more sophisticated models, hybrid approaches combining multiple techniques, automated ontology management, and more effective machine learning algorithms tailored to the specific needs of semantic information retrieval systems.

3. Material and Methods

A. Document Corpus Dataset

The proposed information retrieval system used the TREC Dataset [13] as its document corpus. The package contains the top 'N' papers rated for each query, together with documents and queries in an extra file. This dataset contains two kinds of information: query IDs and the questions themselves. The TREC MRT dataset helps mitigate the assessment gap issues in IR to some extent. Due to the sensitive nature of the information contained inside, data limits serve as the driving force behind medical records. The Text Review Conference (TREC) was co-sponsored by the US Department of Defense and the National Institute of Standards and Technology (NIST) in 1992 as part of the TIPSTER text initiative. The main objective was to promote the development of the text retrieval method needed for large-scale estimation while supporting IR research. Spanish and Chinese were the languages for which TREC sponsored the first thorough assessments of document retrieval efficiency. TREC's computations are based on content-based digital video retrieval and answers to open-domain queries. The realistic design of the operational circumstances is made possible by the requisite test collections in TREC. There are around 3.2 million documents and 367k inquiries on it. The source texts included in the document rating dataset included passages that were part of the passage job. Our training set has 367,013 questions, and the corpus

consists of 3.2 million documents, but with an incomplete collection that was collected after the passing of data. Mapping from a positive passage ID to the matching document ID in a set of 3.2 million documents is done for every training query.

B. Methodology

The proposed system that combines machine learning with feature optimization and a distinct ontology framework for the semantic information retrieval system, as seen in Figure 1.

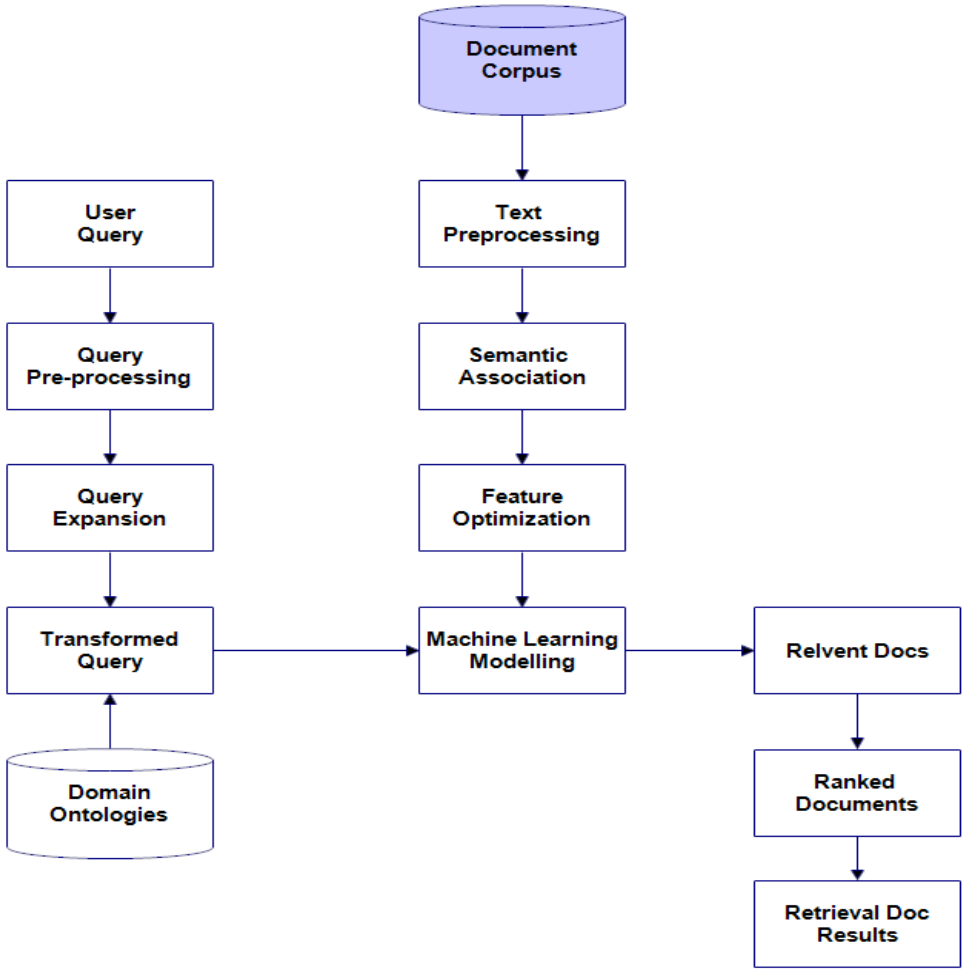

Figure 1: Machine Learning and Ontology based Semantic Information Retrieval System

Data Pre-processing

Pre-processing the query to make it more accurate or machine-readable is necessary before using it to locate the documents that satisfy the user's information requirements. A processed query is sent to the appropriate module in response to this question. There are frequently many terms in a search engine query. An alternate representation of such a

question may be a set of weighted words. Following the matching process, a typical IR system will return several results that are pertinent to the user's request. The query submitter's perception of the page's usefulness in addressing their question is the determining factor in relevance.

Query Expansion Using Domain Ontology

Domain-Based Query Expansion Adding further, semantically relevant terms to the initial user query from an organized knowledge base (ontology) unique to a certain topic is known as ontology. Concepts, their connections, and hierarchical structures inside a domain are represented by ontologies. By adding terms that may be pertinent but aren't specifically stated in the user's query, information retrieval can be improved by using domain ontologies for query expansion.

An ontology can be represented as a directed graph G = (V, E), where:

- V represents the set of concepts or entities.
- E represents the relationships (edges) between the concepts, such as hierarchical (is-a), associative (related-to), or equivalence (synonym) relationships.

Formally:

G = (V, E) where V = {C1, C2, …, Cn} and E = {R1, R2, …, Rm}

Where Ci are the concepts and Rj are the relationships between the concepts.

Given an input query Q with terms T = {t1, t2, …, tk}, the goal is to expand Q by identifying semantically related concepts from the ontology.

For each term ti in Q, the expansion is performed using the ontology graph G:

$T' = \bigcup(ti \in T)$ RelatedTerms (ti)

Where T′ contains the original terms and the expanded set of terms. The related terms can be identified based on the relationships in G, such as:

RelatedTerms (ti) = { Cj | (ti, Cj) ∈ E}

Transformed Query

Unfortunately, the user's original query did not adequately express the specific details he needed. Our algorithm makes use of three main query transformation techniques. Changing the user's original query is typical in manual searches. The retrieval system will automatically broaden the user's query as they type it in. Using the user's past behavior to generate predictions about their current purpose is one approach. This approach augments the user's original inquiry with fresh terms by assessing both local and global information resources. As a whole, this inquiry is trying to answer the first question by revealing related ideas. documents saved on the computer, the entire set of documents, or even only the first set of documents might be targeted. Adding matching terms to searches improves recall. Input from the user is essential for this procedure. After running the refining method, user receive a new query.

Indexing

The database retrieval system organizes millions of web pages that have the same amount of distinct terms. Accurate and speedy retrieval of particular information is the primary objective of indexing. As part of its indexing operation, a search engine will gather keywords and phrases from the sites it has downloaded. The inverted index keeps track of words using a backward file structure. The position of each phrase in the text is tracked by an inverted index. During the indexing of web pages, activities like lexical analysis are carried out, which are similar to the query processing phase and assist improve search engine performance. Almost all words on the page are index terms for full-text indexing. Indexing completes any search engine and enhances query performance by accelerating response times. Apart from just indexing websites, search engines also assign a site a ranking based only on the web's link structure, an attempt to approximate an "importance" assessment. To create the index, the indexing component gets text documents as input. Documents are deconstructed into tokens as part of the automated process of constructing an index. Tokens go through several text operations before they become indexing words.

Feature Optimization

Feature optimization involves selecting the most relevant features (concepts or terms) for the expanded query T′. This can be done using techniques like feature weighting, selection, and dimensionality reduction.

Let W(Cj) represent the weight or importance of concept Cj in the context of the query:

$$W(Cj) = \alpha 1 \cdot \text{Relevance}(Cj) + \alpha 2 \cdot \text{Context}(Cj) + \alpha 3 \cdot \text{Frequency}(Cj)$$

Where:

- Relevance(Cj) measures how relevant Cj is to the original query.
- Context(Cj) measures the contextual alignment of Cj with the query intent.
- Frequency(Cj) measures how frequently Cj appears in relevant documents.

The weights α1, α2, α3 can be tuned using machine learning algorithms based on training data.

Feature selection can then be performed by retaining only the top-n features with the highest weights:

$$T_{optimized} = \{Cj \in T' \mid W(Cj) > \theta\}$$

Where θ is a threshold for feature selection.

Machine Learning Modelling

The comparison of query terms with index items is the responsibility of this part. It looks up all the documents in the index that contain the keywords you entered. This is a typical search that uses processed query terms to access a document index. One indicator of how closely two pages match is the number of words they share. One may determine the relevancy of the returned documents by assessing the degree of resemblance between the index and query phrases. One technique involves locating the URLs of sites that include the query keywords

and contrasting them with the index terms. This document-query or query-keyword syntactic matching can be improved with semantic matching. The proposed approach primarily focuses on an ensemble-based information extraction module. Three machine learning algorithms—KNN, DT, and SVM—were utilized by the ensemble information extractor. The ensemble algorithms use these methods to generate a set of phrases that have already been specified.

With query '*i*' and doc '*i*' representing the query words and documents, respectively, the similarity measure is used in the information retrieval matching process in the proposed framework. During query expansion, we look for the most similar word to the question in terms of semantics to broaden it. By picking the three most important terms from the relevant ontology, we have broadened the scope of the query. We can now use these four terms as a query in any web search engine to find relevant publications based on our domain knowledge of the context. The similarity measure is adjusted according to the equations provided, and the query and document pairs are refined as (query', doc') in the expanded query matching process. In this case, "doc" refers to the newly discovered documents that the search engine has found, and "query" indicates the improved query words.

$$Sim(query, doc) = \frac{\sum_i query_i doc_i}{\sqrt{\sum_i query_i^2}\sqrt{\sum_i doc_i^2}}$$

$$New\_Sim_{(query', doc')} = \frac{\sum_i query_i' doc_i'}{\sqrt{\sum_i query'^2_i}\sqrt{\sum_i doc'^2_i}}$$

Compared to a basic Boolean matching model, the computational effort required by a probabilistic weighted model for NLP queries is higher.

Machine learning models can be used to predict and optimize the feature set based on historical query and retrieval data. For example, a classifier fff can be trained to predict the relevance of features:

$f : T' \rightarrow [0,1]$

Where f(Cj) outputs a probability score indicating the likelihood of Cj being relevant to the query.

The final expanded and optimized query can be represented as:

$Qfinal = \{Cj \in T' \mid f(Cj) > \delta\}$

Where δ is a confidence threshold.

Semantic Information Retrieval

The expanded and optimized query Qfinal is used to retrieve documents D = {d1, d2,…, dm} from the document corpus. The retrieval score S(dj) for each document dj is computed using a similarity measure:

S(dj) = cos(vector(Qfinal), vector(dj))

The ranking uses a score to establish the relative importance of each document. Search engines are based on IR, a field of science and engineering., and one of its major challenges is ranking query results. With a query q and a set of documents D that match it, the primary goal is to use some metric to sort or rank D, with the goal of showing the user the "best" results at the front of the list. Traditionally, the ranking factor has been the relevancy of documents to the given information demand of a query. Through an interface, users may access their graded papers. Based on the study, the researchers use the ranking technique. The process flow is illustrated in Figure 2.

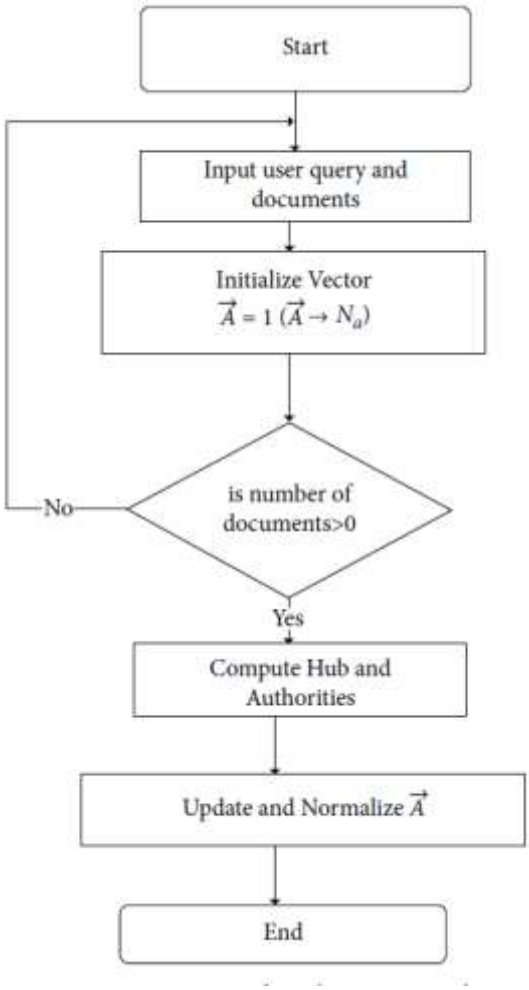

Figure 2: Document Ranking Approach

Algorithm 1: IR system
Input: Query Input Document Output: Relevant Retrieved Information

Method:

Step1: The user's information needs are defined by query input from the user interface.

Step 2: Assessing the initial inquiry makes use of domain ontology in conjunction with query processing procedures. Document collection is the first step in indexing, which entails carrying out the identical assignments given to them by the module for semantic association.

Step 3: Indexes are generated from the original documents in the context of the indexing module.

Step4: Ontology-Based Query Expansion - Expand query using ontology concepts and relations, Identify synonyms, related terms, hierarchical relationships.

Step5: The user's information needs are modeled semantically by the transformed query.

Step6: Feature Optimization - Feature weighting using machine learning Dimensionality reduction (select most relevant expanded terms based on learned weights.

Step7: The search for literature that answer this query is carried out using a semantics module. The indexing system finds all of the document keywords that match the query phrase quickly.

Step8: Semantic Information Retrieval - Document matching based on optimized query, Compute similarity scores.

Step9: As a function of the word matching score, the recovered documents are arranged in order of expected relevance.

Step10: Ranked Document Results - Rank documents based on similarity scores and Return top-N relevant results to the user.

4. Experimental Result and Discussion

A. Experimental Setup

The proposed information retrieval system is designed in Python and evaluated with testing using Pycharm IDE. The software used is Windows 11, with 16 GB of RAM, an Intel i7 processor, and 8 GB of graphics memory. The scikit-learn ML and Word2Vec toolbox are part of the with Anaconda Distribution. To forecast the optimal performance rate, performance parameters such as F-measure, Mean Average Precision (MAP), precision, and recall are utilized.

B. Evaluation Parameters

Three commonly used metrics and the proposed model were used to evaluate our results. What follows is a more detailed explanation of these metrics.

1. Recall (R): It is the percentage of relevant documents that are discovered during a search. To find this value, use the following formulas. The numbers Dr and Dt represent the number of relevant data that were and were not retrieved, respectively.

$$R_r = \frac{D_r}{D_r + D_t}$$

2. Precision (P): Relevance calculates the proportion of retrieved documents that were truly helpful. Du is the number of useless data that was obtained.

$$R_p = \frac{D_r}{D_r + D_u}$$

3. Mean Average Precision (MAP): It is computed as the average precision score for the documents that the query groups were used to get. The Average Precision (AP) of every search query is the basis for calculating the MAP. After that, the MAP is calculated and recorded as;

$$MAP = \frac{1}{n} \sum_{n} AP_n$$

4. F-measure: F-measure is as simple as summing up the precision and recall scores harmonically.

F-measure = 2 * (Rp * Rr) / (Rp + Rr)

C. Experimental Results analysis

Experiments are carried out and evaluated using TREC datasets to show how effective the proposed and baseline methods are. The information retrieval function was created to accept a query as input and return the first 'N' documents that are pertinent. This process will come after the pipeline for information retrieval (IR). Initially, the query will be pre-processed. The semantic feature vector for it will then be generated by it. It will then order the papers according to their similarity ratings. These metrics, recall, precision, and F-measure are used to assess performance. The number of most-read documents was used to determine outcomes by selecting the 10, 20, 50, and 100 as examples. Table 1 below provides an in-depth analysis of each of these components and Table 2 defines the with feature optimization effect on retrieval system.

Table 1: Evaluation of Results based on Top-N Documents for Retrieval without Optimization

Performance Parameter	Recall	Precision	F-measure
N@10	0.72	0.69	0.71
N@20	0.70	0.68	0.69
N@50	0.71	0.65	0.67
N@100	0.68	0.64	0.66

Table 2: Evaluation of Results based on Top-N Documents for Retrieval with Optimization

Performance Parameter	Recall	Precision	F-measure
N@10	0.72	0.70	0.71
N@20	0.73	0.71	0.70
N@50	0.75	0.69	0.68
N@100	0.72	0.68	0.69

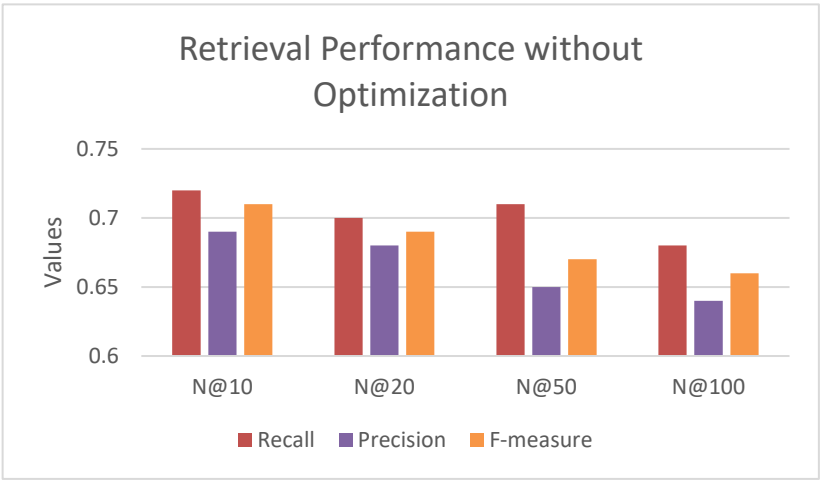

Figure 3: IR Result evaluation performance for Top-N retrieval

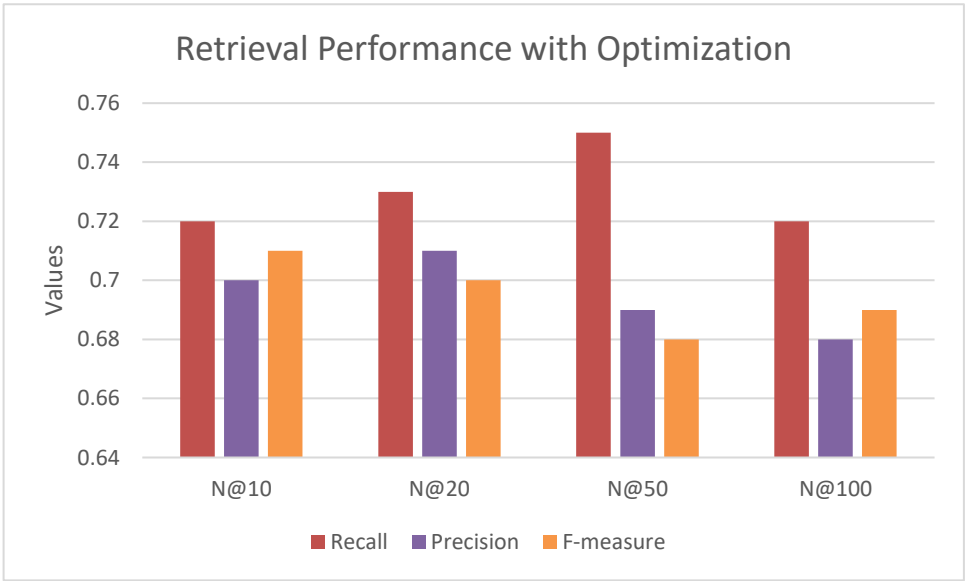

Figure 4: IR Result evaluation performance for Top-N retrieval

The TREC MRT dataset's recall, precision, and f-measure results are displayed in Figure 3 an in Figure 4 without and with feature optimization. The proposed system outperformed the others in terms of precision @10, 20, 50, and 100 respectively with optimization technique. According to the proposed method, recall @10 and 50 is 0.72 and 0.75 respectively. The developed model performed better in comparison to other retrieval documents on the f-

measure @100.

Table 3: Performance Evaluation Parameters for Test Query without optimization

Evaluation Parameters/Query	Recall	Precision	F-score	MAP
Query1	0.55	0.60	0.52	0.51
Query2	0.48	0.54	0.46	0.44
Query3	0.49	0.54	0.47	0.43
Query4	0.53	0.57	0.48	0.46
Query5	0.57	0.60	0.53	0.55

Table 4: Performance Evaluation Parameters for Test Query with optimization

Evaluation Parameters/Query	Recall	Precision	F-score	MAP
Query1	0.57	0.62	0.53	0.54
Query2	0.49	0.55	0.47	0.45
Query3	0.52	0.57	0.49	0.45
Query4	0.56	0.58	0.50	0.48
Query5	0.58	0.61	0.55	0.56

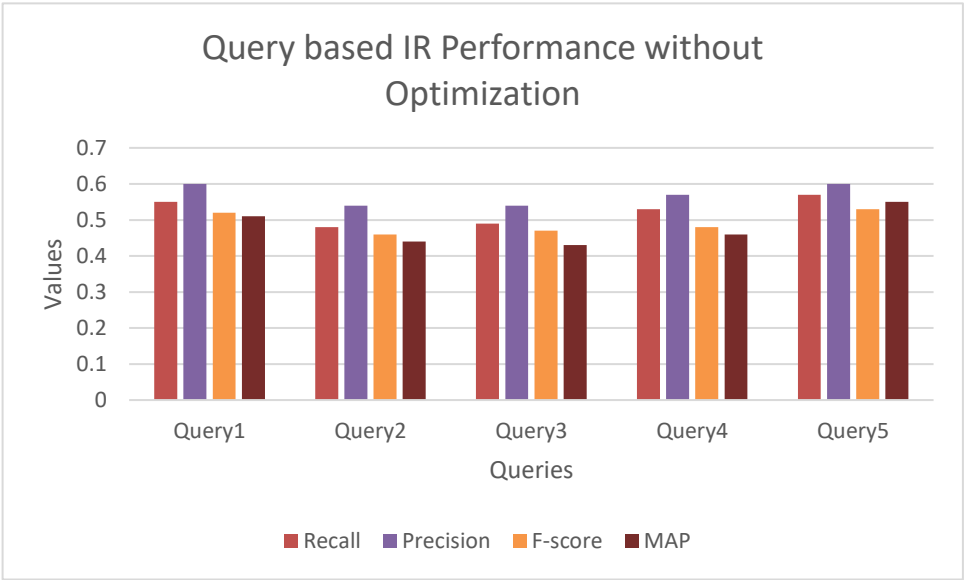

Figure 5: IR Result evaluation performance for N-Query

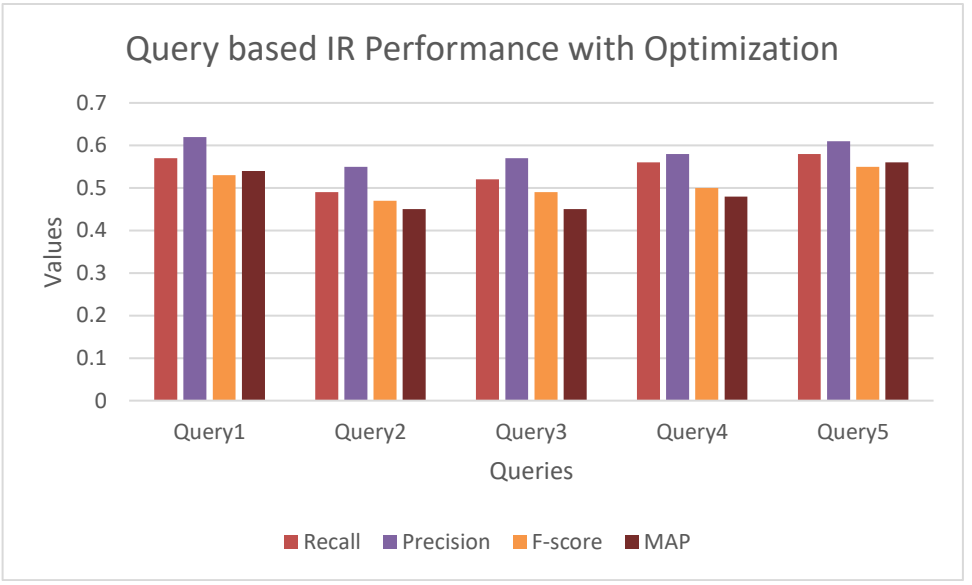

Figure 6: IR Result evaluation performance for N-Query

Table 3 and Table 4 depicts the results of several retrieval as per the query applied shown using f-measure, recall, precision, and MAP. New quarterly highs of 0.58 for recall, 0.61 for precision, 0.55 for f-measure, and 0.56 for MAP were achieved for Query5 with feature optimization approach. With an f-Measure of 0.47 and a MAP of 0.45, the final results for Query2 are precision of 0.55 and a recall of 0.49. Several measures, including f-measure, recall, precision, and MAP, all attained levels of 0.49, 0.52, 0.57, and 0.45 during Query3. See the visual examination of query expansion in Figure 6. When comparing studies conducted in Query1, Query2, Query3, Query4 and Query5 performance parameters including f-measure, recall, precision, and MAP all showed significant improvements in Query5 with feature optimization as shown in figure 5 and figure 6.

Table 5: State of Art Comparative Analysis

Reference Model	MAP
IRS [22]	0.20
IAOCOOT [32]	0.3945
Wikipedia [32]	0.3166
WordNet [32]	0.2901
Proposed Model	0.715

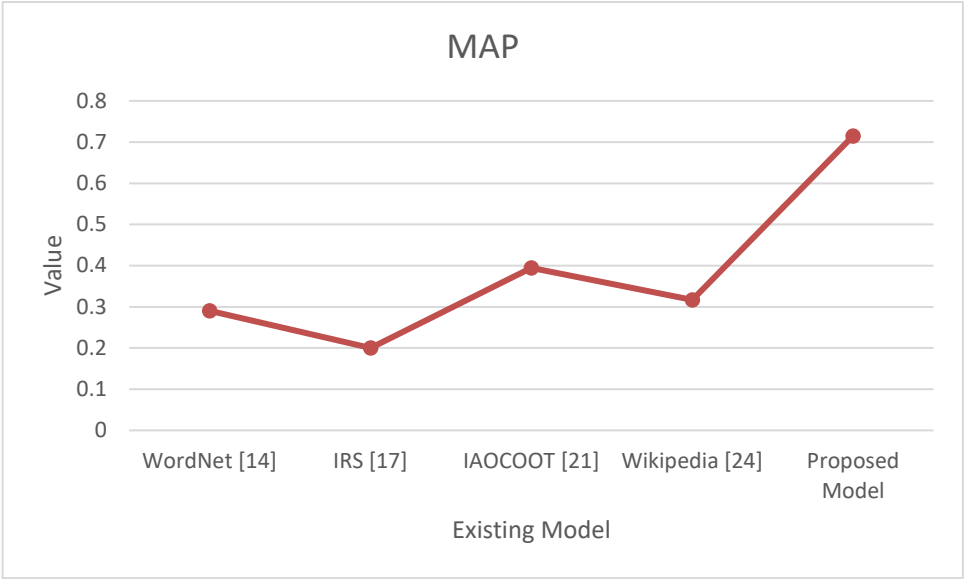

Figure 7: Comparison with state of art model

Table 5 provides a full analysis of the proposed MAP results and compares them with the most recent models. As seen in Figure 7, the suggested model outperformed rival state-of-the-art models in MAP findings. The ontology-based semantic information-based model reduces ambiguity, which increases the proportion of relevant documents recovered by an IR system while boosting the performance of suggested information retrieval.

5. Conclusion

Conventional semantic information retrieval methods may converge slowly when presented with enormous amounts of complex content. This paper explored an ontology framework for enhancing semantic information retrieval systems through query expansion, feature optimization, and machine learning. Our study aimed to address the limitations of traditional information retrieval systems by integrating advanced techniques to improve query accuracy and relevance. Using semantic association, the indexing strategy was able to better capture terms associated with documents. Thus, the two parts of the proposed system coordinated to ensure that the appropriate papers were located for the appropriate individuals at the appropriate times. Through the application of ontology and semantics approaches, TREC datasets showed improved performance. The window function computed, retrieved the state parameters of the ideal semantic information choice, and fused its data levels to gain information fusion outcomes utilizing unstructured semantic data processing. Feature optimization using machine learning models enabled us to prioritize and select the most relevant terms from the expanded query. This approach reduced noise and improved the precision of the retrieval system. Applying machine learning algorithms to refine feature selection and query expansion processes demonstrated substantial improvements in retrieval performance. The models effectively captured complex patterns and relationships, enhancing the system's ability to handle diverse queries. In comparison to earlier studies, the suggested

strategy outperformed the competition when we monitored the overall amount of document obtained from major search engines and used metrics like as f-measure, precision, MAP and recall. There was a 10% improvement in accuracy after broadening the query. The proposed framework showed promising results in terms of retrieval accuracy and relevance. By leveraging ontologies and advanced optimization techniques, the system was able to provide more precise and contextually appropriate results compared to traditional methods. The proposed framework not only addresses existing limitations but also opens new avenues for future research and development in this field. Future work should focus on enhancing the scalability of the framework to handle large-scale ontologies and real-time query processing. Improving the system's ability to understand and interpret user context more effectively will further refine query expansion and retrieval accuracy.

References

1. Yuvaraj, D., Alnuaimi, S. S., Rasheed, B. H., Sivaram, M., & Porkodi, V. (2024). Ontology Based Semantic Enrichment for Improved Information Retrieval Model. International Journal of Intelligent Systems and Applications in Engineering, 12(15s), 70–77.
2. Anil Sharma and Suresh Kumar. 2023. Machine learning and ontology-based novel semantic document indexing for information retrieval. Comput. Ind. Eng. 176, C (Feb 2023). https://doi.org/10.1016/j.cie.2022.108940.
3. Hamane, Zakaria & Samih, Amina & Fennan, Abdelhadi. (2023). Ontology Matching Using Deep Learning. Journal of Theoretical and Applied Information Technology. 101.
4. Kumar, R., Sharma, S.C. Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval. J Supercomput 79, 2251–2280 (2023). https://doi.org/10.1007/s11227-022-04708-9.
5. Chebil, Wiem, Mohammad Wedyan, Moutaz Alazab, Ryan Alturki, and Omar Elshaweesh. 2023. "Improving Semantic Information Retrieval Using Multinomial Naive Bayes Classifier and Bayesian Networks" Information 14, no. 5: 272. https://doi.org/10.3390/info14050272.
6. Ali, A. (2022). Review of Semantic Importance and Role of using Ontologies in Web Information Retrieval Techniques. International Journal of Computer and Information Technology (2279-0764), 11(1). https://doi.org/10.24203/ijcit.v11i1.240
7. Abhishek Kumar Shukla, Sujoy Das, Pushpendra Kumar, Afroj Alam, "Relevance Feedback and Deep Neural Network-Based Semantic Method for Query Expansion", Wireless Communications and Mobile Computing, vol. 2022, Article ID 6789044, 11 pages, 2022. https://doi.org/10.1155/2022/6789044
8. Alothman AF, Wahab Sait AR. Managing and Retrieving Bilingual Documents Using Artificial Intelligence-Based Ontological Framework. Comput Intell Neurosci. 2022 Aug 25;2022:4636931. doi: 10.1155/2022/4636931. PMID: 36059407; PMCID: PMC9436537.
9. Sharma, Anil & Kumar, Suresh. (2022). Shallow Neural Network and Ontology-Based Novel Semantic Document Indexing for Information Retrieval. Intelligent Automation & Soft Computing. 34. 1989-2005. 10.32604/iasc.2022.026095.
10. Anand, Sanjay & Kumar, Suresh. (2022). Ontology-based Soft Computing and Machine Learning Model for Efficient Retrieval. 10.21203/rs.3.rs-2328237/v1.
11. Hu, M. (2022). Research on Semantic Information Retrieval Based on Improved Fish Swarm Algorithm. Journal of Web Engineering, 21(03), 845–860. https://doi.org/10.13052/jwe1540-9589.21313.
12. Shakeri, M.; Sadeghi-Niaraki, A.; Choi, S.-M.; AbuHmed, T. AR Search Engine: Semantic

Information Retrieval for Augmented Reality Domain. Sustainability 2022, 14, 15681. https://doi.org/10.3390/su142315681.

13. Data-English documents. Text REtrieval conference (TREC) english documents. (n.d.). Retrieved from https://trec.nist.gov/data/ docs_ eng. html. Accessed 27 May 2022; https://microsoft.github.io/TREC-2019-Deep-Learning/.
14. Kumar, Ram & Sharma, Subhash. (2022). Smart Information Retrieval using Query Transformation based on Ontology and Semantic-Association. International Journal of Advanced Computer Science and Applications. 13. 10.14569/IJACSA.2022.0130446.
15. Trabelsi, M., Chen, Z., Davison, B.D. et al. Neural ranking models for document retrieval. Inf Retrieval J 24, 400–444 (2021). https://doi.org/10.1007/s10791-021-09398-0.
16. Jain, Shivani & K.R., Seeja & Jindal, Rajni. (2021). A fuzzy ontology framework in information retrieval using semantic query expansion. International Journal of Information Management Data Insights. 1. 100009. 10.1016/j.jjimei.2021.100009.
17. Lamurias A, Jesus S, Neveu V, Salek RM, Couto FM. Information Retrieval Using Machine Learning for Biomarker Curation in the Exposome-Explorer. Front Res Metr Anal. 2021 Aug 19;6:689264. doi: 10.3389/frma.2021.689264. PMID: 34490412; PMCID: PMC8417071.
18. Mahalakshmi, P. & Fathima, N.. (2021). An Art of Review on Conceptual based Information Retrieval. Webology. 18. 21-31. 10.14704/WEB/V18SI02/WEB18009.
19. Kulmanov M, Smaili FZ, Gao X, Hoehndorf R. Semantic similarity and machine learning with ontologies. Brief Bioinform. 2021 Jul 20;22(4):bbaa199. doi: 10.1093/bib/bbaa199. PMID: 33049044; PMCID: PMC8293838.
20. L. Yan-ji, "Construction of Chinese-English Cross-language Information Retrieval Model Based on Dictionary Learning," 2021 International Conference of Social Computing and Digital Economy (ICSCDE), Chongqing, China, 2021, pp. 100-104, doi: 10.1109/ICSCDE54196.2021.00032.
21. Wang, Zhe & Zhao, Yingying & Dong, Hai & Xu, Yulong & Lv, Yali. (2021). Improved Algorithm Based on Decision Tree for Semantic Information Retrieval. Intelligent Automation & Soft Computing. 29. 419-429. 10.32604/iasc.2021.016434.
22. L. Kim, E. Yahia, F. Segonds, P. Véron, and A. Mallet, “i-Dataquest: A heterogeneous information retrieval tool using data graph for the manufacturing industry,” Computers in Industry, Vol. no.132, pp.103527, 2021.
23. N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models.” arXiv preprint arXiv:2104.08663. 2021.
24. Esteva, A., Kale, A., Paulus, R. et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. npj Digit. Med. 4, 68 (2021). https://doi.org/10.1038/s41746-021-00437-0.
25. S. F. N. B. S. Kamaruddin, F. Mohd, M. P. Hamzah, F. Harun, N. R. Zainol and N. I. M. Daud, "Information Retrieval for Malay Text: A Decade Review of Research (2008–2019)," 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP), Kuala Lumpur, Malaysia, 2021, pp. 2-7, doi: 10.1109/CAMP51653.2021.9498034.
26. Liji S K, Muhamed Ilyas P, “Review and Analysis of Different Approaches to Semantic Level Question Answering and Information Retrieval”, International Journal of Science and Research (IJSR), Volume 10 Issue 1, January 2021, DOI: 10.21275/SR21121141135
27. Rawal, Samarth & Baral, Chitta. (2020). Multi-Perspective Semantic Information Retrieval. https://doi.org/10.48550/arXiv.2009.01938
28. Maristella Agosti, Stefano Marchesin, and Gianmaria Silvello. 2020. Learning Unsupervised Knowledge-Enhanced Representations to Reduce the Semantic Gap in Information Retrieval. ACM Trans. Inf. Syst. 38, 4, Article 38 (October 2020), 48 pages. https://doi.org/10.1145/3417996.

29. P. Lai, N. Phan, H. Hu, A. Badeti, D. Newman and D. Dou, "Ontology-based Interpretable Machine Learning for Textual Data," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-10, doi: 10.1109/IJCNN48605.2020.9206753.
30. Silva, Fabiano & Maia, José. (2020). Query Expansion in Text Information Retrieval with Local Context and Distributional Model. Journal of Digital Information Management. 17. 10.6025/jdim/2019/17/6/313-320.
31. Qi, Y., Zhang, J., Xu, W. et al. Salient context-based semantic matching for information retrieval. EURASIP J. Adv. Signal Process. 2020, 33 (2020). https://doi.org/10.1186/s13634-020-00688-1
32. ALMarwi, H., Ghurab, M. & Al-Baltah, I. A hybrid semantic query expansion approach for Arabic information retrieval. J Big Data 7, 39 (2020). https://doi.org/10.1186/s40537-020-00310-z.
33. D. Yasin, A. Sohail and I. Siddiqi, "Semantic Video Retrieval using Deep Learning Techniques," 2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2020, pp. 338-343, doi: 10.1109/IBCAST47879.2020.9044601.
34. Rawal, S. (2020). Multi-Perspective Semantic Information Retrieval in the Biomedical Domain. ArXiv, abs/2008.01526.
35. Elnagar, Samaa & Yoon, Victoria & Thomas, Manoj. (2020). An Automatic Ontology Generation Framework with An Organizational Perspective. 10.24251/HICSS.2020.597.
36. Selvalakshmi, B., Subramaniam, M. Intelligent ontology based semantic information retrieval using feature selection and classification. Cluster Comput 22 (Suppl 5), 12871–12881 (2019). https://doi.org/10.1007/s10586-018-1789-8.
37. Shah, Ritesh & Guide, Prof. (2019). Ensemble based Machine Learning using Ontology Information Extraction for Information Retrieval. International Journal of Innovative Technology and Exploring Engineering. 9. 3952-3959. 10.35940/ijitee.B7011.129219.
38. P. Liu, Z. Zheng and Q. Su, "Cross-Language Information Retrieval Based on Multiple Information," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, 2018, pp. 623-626, doi: 10.1109/WI.2018.00-26.
39. M. N. Asim, M. Wasim, M. U. Ghani Khan, N. Mahmood and W. Mahmood, "The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval," in IEEE Access, vol. 7, pp. 21662-21686, 2019, doi: 10.1109/ACCESS.2019.2897849.
40. Kamran Munir, M. Sheraz Anjum, The use of ontologies for effective knowledge modelling and information retrieval, Applied Computing and Informatics, Volume 14, Issue 2, 2018, Pages 116-126, ISSN 2210-8327, https://doi.org/10.1016/j.aci.2017.07.003.
41. M.J.H. Mughal, "Data mining: Web data mining techniques, tools, and algorithms: An overview," Information Retrieval, 9(6), 2018.
42. Sridevi, U. K., P. Shanthi, and N. Nagaveni. "Deep Model Framework for Ontology-Based Document Clustering." In Handbook of Research on Investigations in Artificial Life Research and Development, edited by Maki Habib, 424-435. Hershey, PA: IGI Global, 2018. https://doi.org/10.4018/978-1-5225-5396-0.ch019