Smart Information Retrieval using Query Transformation based on Ontology and Semantic-Association

Ram Kumar*, S. C. Sharma
Indian Institute of Technology Roorkee, India

***Abstract*—A notable problem with current information retrieval systems is that the input queries cannot express user information needs properly. This imprecise representation of the query hampers the effectiveness of the retrieval system. One method to solve this problem is to transform the original query into a more meaningful form. This paper proposes an ontology-based retrieval system that transforms initial user queries using domain ontologies and applies semantic association during the indexing process. The proposed system performs a semantic matching between an ontologically enhanced query and index to capture query-related terms. To show the performance of the proposed system, it is evaluated using standard parameters like precision, recall, and NDCG. In addition, the authors presented a comparison between the proposed and existing retrieval systems on three test datasets. Experimental results on these datasets indicate that the use of ontology and semantics has significantly increased the retrieval efficiency obtained by baseline. This work highlights the importance of ontology and semantics in information retrieval.**

Keywords*—*Ontology; semantics; information retrieval; query transformation; indexing

I. Introduction

In this digital age, the fast rate of data generation created a massive collection of data. It becomes a challenging task for computer users to extract relevant information from this large data collection. World Wide Web [1] is one such collection with billions of such web pages. Information Retrieval System (IRS) enables web users to find the desired information from such extensive resources. Search engines are commercially available IRS, which play a vital role in finding information online. Nowadays, the search engine has become the primary means to find websites. The simple Web search engine retrieves information from Web pages using keyword matching between queries and documents. The Web is rapidly growing, and different users seek focused information; now, it has become a must for search engines to utilize semantic techniques to satisfy users' information needs [2]. Using Semantic search, the machine can also understand the user's interest and the context in which the search is issued; this will help in providing the most relevant information to the user's search need. Traditional search engines need to adopt new changes to find exact information from the Web.

Ontology-based information retrieval is a small step in this direction. This can meet those challenges that were not met by traditional retrieval systems. The authors propose an ontology-based retrieval system for finding the most relevant search results.

The Internet is a collection of interlinked documents (billions in numbers) distributed over the most extensive network. In the last two decades, the web has grown exponentially. The large scale of the Web made it almost impossible to retrieve desired information without any tool. That's why internet user searches the web for the topic of their interest using a search engine. A search engine [3] is a software program or tool to find information from many web pages distributed over the internet. A search engine searches the document for entered set of keywords and returns a list of results in links to relevant resources. These links are Uniform Resource Locators (URLs) of those documents where any or all of the searched keywords can be found. A search engine provides results quickly using high-speed systems working globally known as index servers. The searched URLs are accessed using a program called a web browser.

Sir Tim Berners-Lee defines the Semantic Web [4] as an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. Semantic search is a search where humans and machines try to find concepts behind terms used by different users. The Semantic Web supports more efficient discovery, automation, integration, and data reuse. Semantic Web languages have been developed to describe knowledge using a new W3C standard. These are RDF(S) (Resource Description Framework/Schema), OWL (Web Ontology Languages), and OIL, DAML+OIL. Most of these standards relied on Ontology for releasing the dream of the semantic web. Gruber [5] defines ontology as a specification of a shared conceptualization.

Ontology always includes a vocabulary of representational concept labels to describe a shared domain. These concept labels are usually called terms (lexical references) associated with entities. Ontology is one of the essential concepts used in the semantic web infrastructure. These days, researchers use the Semantic Web and Ontology to manage data in information retrieval systems. The Semantic Web simplifies and improves knowledge-intensive applications through ontology by addressing weaknesses in information retrieval, matching data, and data integration on the current Web. The semantic web aims to provide an extra machine-understandable layer, simplifying programming and maintenance efforts for knowledge-based web services.

*Corresponding Author.

This paper proposes an ontology-based retrieval system that transforms initial user queries using domain ontologies and applies semantic association during the indexing process. The main contributions of this paper are:

- Firstly, this paper proposes an ontology-based information retrieval system.
- Secondly, this paper implements the semantic approach in the indexing of documents.
- And finally, it also demonstrates how ontology-based search systems outperform the baseline.

The rest of the paper is organized as, in Section II, the authors present related work, Section III shows the proposed methodology, architecture, and algorithm. The experimental setup is given in Section IV. Section V presents the results and its discussion; we conclude in Section VI.

II. Related Work

By establishing successful approaches to overcome the difficulties in providing a more specific description of a user's information need, Ontology and Semantics took a giant step forward. Furthermore, it has also outperformed traditional IRS in terms of retrieval results. Researchers have adopted various techniques to transform query terms for performing retrieval. Some of them are given below.

Maxat Kulmanov et al. [6] employed fuzzy ontology in this study to aid with query transformation for an IRS. This procedure generates a dictionary of concepts from a given domain and an exterior ontology and then assigns fuzzy memberships via ConceptNet. They have used ConceptNet Global Ontology to determine fuzzy membership. By adding semantic weights to every one of the Concept-Net Ontology's semantic relations, the researchers created fuzzy membership for all these relations.

The researcher of this study [7] provides an overview of the methods for computing similarity using ontologies and incorporating them into machine learning techniques; they discuss how ontology embeddings and semantic similarity measures can leverage the background knowledge contained in ontologies, as well as how ontologies could provide restrictions that enhance machine learning methods.

The reference [8] established a mechanism for semantic query expansion based on domain ontologies. It expanded on synonyms, hypernyms, and similar words. Adding a similarity function to a system improved the quality of the formed query and the search engine results. Using a programming language domain enabled the system to be evaluated manually by impartial and experienced personnel.

Prilipsky et al. [9] proposed a hybrid Personal Knowledge Management (PKM) system to extract and store helpful information. The software-assisted PKM is the most effective one. Many software tools are used to retrieve concept mapping, tagging, flashcards, and hyperlinking. These tools are added to the traditional PKM to make it flexible and extendable. The advantage of the state-of-art work is a single solution using PKM instead of integrating all the functionalities from different tools.

Kim et al. [10] proposed an i-Dataquest prototype for a graph-based information retrieval system. I-Dataquest prototype contains three steps, data pre-processing, query pre-processing, relevance evaluation, and feedback. The graph data integrates the query with all syntactic and semantic extensions to retrieve the required answer. The PAINT'R dataset is selected, which is the same as the data of the manufacturing company.

Esposito et al. [11] proposed a hybrid query transformation approach. This approach is used in Information retrieval-based QA systems. It is based on lexical resources and word embeddings. In the query expansion for the question-answer system, the answer for the user-defined question is retrieved from an already formulated database having a particular required domain. This method modifies the natural language questions to enhance proper semantics to get the required sentences. The author used WordNet to produce appropriate nouns and verbs for the user questions. Then as per the sense of the question, the answer is retrieved using the Word2Vec model, which is created using a semantic similarity metric. The main drawback is that this approach uses only nouns and verbs for evaluating the query; it does not consider other syntactic categories, such as adjectives and adverbs.

III. Methodology

The problem with current IR systems is that the input queries are generally too short and too ambiguous to express the user's information needs. Such imprecise representation of users' information needs directly affects retrieval performance. In other words, it can be said that a simple query can't satisfy users' information needs.

A. Problem Formulation

Vocabulary mismatch between the query and documents. Let's consider a document collection with D relevant documents for a user query Q to understand this problem. For representing a single concept, Q and D may use different vocabulary. A traditional retrieval system performs only keyword matching between query and document. It does not detect similarity between Q and D. The authors address this problem using semantic and ontologies, which provide shared meaning for two different terms.

B. Proposed Information Retrieval System

To solve the above-stated IR problem, the researchers propose a novel ontology-based retrieving system as shown in Fig. 1. The steps of the proposed system are as follows:

1) The user's information need is specified by a user query (typically made of keywords) entered via the user interface.

2) The initial query is processed using domain ontology and query processing operations. Same operations are applied to document collection by semantically association module for Indexing purposes.

3) Index building from the document source is an offline process performed by the indexing module.

4) The transformed query is a semantic representation of user information needs.

5) This query is executed by searching using a semantics module to retrieve a set of relevant documents. Fast matching between query keywords and documents terms is done by the index structure.

6) The set of retrieved documents is then ranked according to the estimated relevance with respect to the term matching score.

7) The user then examines the set of ranked documents; he might point to a subset of the documents as useful and thus provide feedback to the system.

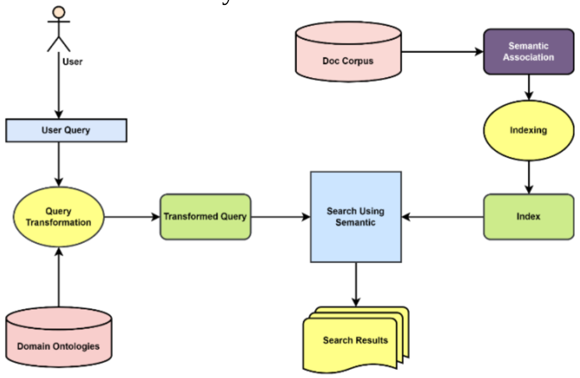

Fig. 1. An Ontology-based Semantic Information Retrieval System.

1) Query processing: To discover the documents that meet the user's information requirements, the query must first pass through a pre-processing module, which converts it into a more precise or machine-readable format. The matching module receives this type of question, known as a processed query. A typical search engine query consists of several terms. A list of terms with weights can also represent such a query. A standard information retrieval system will provide a high percentage of pages relevant to the user's query after matching. If the sender of the inquiry finds a page relevant to the topic for which the query was submitted, it is considered relevant.

2) Indexing: The IRS indexes millions of web pages containing a comparable number of distinct terms. Indexing can be defined as a process that collects, parse and store data to facilitate fast and accurate information retrieval. During indexing, a search engine records the words and phrases from downloaded pages; then, it prepares an index based on this data. It stores terms in an inverted file structure known as an inverted index. An inverted index stores the positions of text for each occurrence of a term. Another reason for indexing web pages by search engines is that it carried out processing like lexical analysis similar to the query processing phase, which improves the performance of search engines. In full-text indexing, virtually every word in the document is employed as an index term.

Indexing is an integral part of every search engine because it optimizes the query performance by improving the response times considerably. Along with Indexing, search engines also perform ranking, which is an attempt to see how good an approximation to “importance” can be obtained from just the link structure of the web. The index is built from text documents by the indexing module. Preparing an index is an offline process that parses text documents into tokens. Various text operations are performed on these tokens, transforming them into indexing terms.

3) Search using semantics: In this module, query keywords are matched with index terms. It retrieves those documents from the index that contain given query terms. It is typically a standard search for processed query terms in an index of documents. The degree of matching between a page and a query, called the similarity, can be measured by the number of terms they share. A similarity score between query and index terms is calculated to rank returned documents. A simple approach is to match query keywords with index terms and return the URLs of those documents that contain matching terms. This keyword or syntax matching between query and document can be improved with semantic matching.

The computational processing required for an NLP-based query having a probabilistic weighted model is more than an unweighted, Boolean matching model. Ranking scores all retrieved documents according to a relevance metric. One of the fundamental difficulties in information retrieval, the scientific/engineering subject that underpins search engines, is ranking query results. The task is to rank or sort the documents in D according to some criterion such that the "best" results appear first in the result list displayed to the user, given a query q and a collection D containing documents that fit the query. Traditionally, ranking criteria have been described in documents relevance to a query's expressed information demand. These graded documents are returned to the user using a user interface.

Algorithm 1: Pseudo Code for The Proposed OBS-IRS

```
Input     :  User Information need
Output    :  Retrieved relevant documents
             start
Step 1    :  Formulates the initial query q and submits it via the
             user's interface
Step 2    :  Query transformation module process q using
             domain ontology
Step 3    :  q is transformed to a more machine-readable form
             q'
Step 4    :  QT utilizes domain knowledge from ontology
             dataset
Step 5    :  Build Index I;
Step 6    :  Process documents from Doc Corpus using
             semantic association
Step 7    :  Apply Index preparing method Inverted Index on
             fetched pages
Step 8    :  for terms extracted from doc with semantic data
Step 9    :  Semantic Search
Step 10   :  Match q' with index terms
Step 11   :  Start for
Step 12   :  each term in q'
Step 13   :  if (q' term and index term have the same meaning)
Step 14   :  Retrieve document according to the matching
             function
Step 15   :  End for
Step 16   :  Apply Ranking on documents
Step 17   :  Return search results
             end
```

C. Query Transformation

The initial user query is not able to represent his information need properly. Our system uses three main query transformation techniques.

1) Query expansion: The IR use of Query Expansion (QE) has been the subject of research [12]. As seen in manual search, it can be seen that the user reformulates their search query. They do it because they didn't get the exact result of their original query. In the QE, the IRS improves the user's query by automatically expanding it. This can be done in several manners, like providing suggestions by guessing the user's intention according to the user's past behaviour.

This technique adds additional terms to the user's initial query based on local and global information resources analysis. This analysis focuses on finding semantically related terms to the original query. These target resources can be the whole document collection, the initially retrieved documents set, or documents from the computer. Expansion of queries [13] with matching terms improves performance in terms of recall. However, any method must find similar terms carefully during this process because it may lose gain in terms of precision.

2) Query refinement: Researcher use this refinement technique to improve the matching between queries and documents. The process of reflecting user needs with high accuracy is called query refinement [14]. In this process, the feedback by the user plays an important role. It generates a new query after applying the refinement process. Research on query refinement is not as dominant as query expansion.

3) Query suggestion: Query suggestion is also a part of the query transformation module. The most common form is spell checking during query processing by any search engine. The user is offered replacements to the initial query. These alternatives are more specific to the user's information need. Query suggestions provide more detailed descriptions of the search concept. This technique uses the extensive query history collected by web applications [15]. To implement query suggestion, the retrieval system generates a new query. It is different from the QE because it does not always add new terms to the initial query. Table I shows the query transformation approaches concerning their behavior on user feedback.

TABLE I. SUMMARY OF QUERY TRANSFORMATION METHODS

	Query expansion	Query refinement	Query suggestion
Generate new query		Yes	Yes
Expand original query	Yes		
User feedback before modification	Yes	Yes	
User feedback for final modification			Yes

IV. EXPERIMENTAL SETUP

For the experiment, the authors used Terrier [16] retrieval system. It is an open-source tool for IR experiments developed by Glasgow University, UK. Three standard datasets are also used for comparing the results of both approaches.

A. Datasets

The first dataset, ANTIQUE [17], is a non-factoid community question answering dataset. It is a collection of 2,626 open-domain non-factoid questions from a diverse set of categories. The Cranfield [18] is a small curated dataset extensively used in information retrieval experiments. There are 226 queries (search terms), 1400 documents, and 1837 (evaluations) in the dataset. The TREC MRT dataset [19] helps fill evaluation gap issues in IR. Due to the sensitive nature of medical records, data constraints are the overarching factor for Medical Records.

B. Evaluation Metrics

The authors used three standard evaluation parameters to evaluate our results. The detailed information about these measures is given below.

1) Precision (P): It is the fraction of retrieved documents that are relevant. It measures the quality of the results [20]. It is also known as positive predictive value. It can be calculated at different values, denoted as Precision at k (shortened as P@k). It can be calculated using Eq (1).

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \tag{1}$$

2) Recall (R): It is the ratio of relevant documents that are retrieved. It can be seen as a measure of the quantity of documents corresponding to an information need [21]. It is also known as sensitivity. Being ratio, it has a value between 0 and 1. It can be calculated using both equations given below as Eq. (2).

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)} \tag{2}$$

3) Normalized Discounted Cumulative Gain (NDCG): To represent non-binary relevance, the use of cumulative gain or specifically normalized discounted cumulative has been increasing. It is a widely accepted evaluation parameter in the IR community [22]. It can also be calculated at a given rank cutoff (e.g. ndcg_cut_10). It can be calculated for k, top search results. For query, j form a set of queries Q, consider R(j,d) as relevance score is given to document d, it is mathematically calculated by Eq. (3).

$$NDCG(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{log_2(1+m)} \tag{3}$$

V. RESULT AND DISCUSSION

To show the performance of the proposed and baseline system, the authors evaluated both on three datasets. The parameter used to judge the performance are precision, recall, and NDCG. The results were taken at different levels such as 10, 20, 50, and 100 top documents. The detailed discussion for each parameter is given in the following subsection.

A. *Precision Analysis*

Fig. 2 shows results for the precision on the ANTIQUE dataset. Here you can see that the proposed combination of semantics and ontology outperformed the baseline results. For p@10, the value for the ANTIQUE dataset was 0.2356 in the baseline model, whereas it was 0.2452 for the proposed system. For P@20, values of baseline and proposed were 0.2322 and 0.2396. The value of @50 was 0.2145 and 0.2292. At the 100 documents, the precision values of baseline and proposed were 0.2012 and 0.2086. This shows that the proposed approach performed better for all values of precision. Fig. 3 shows results for precision on the Cranfield dataset. For this dataset also, the proposed retrieval outperformed bassline retrieval. Similarly, in Fig. 4, you can see the precision results for the TREC MRT dataset; our system was also better in terms of P@10, P@20, P@30, and P@100 results.

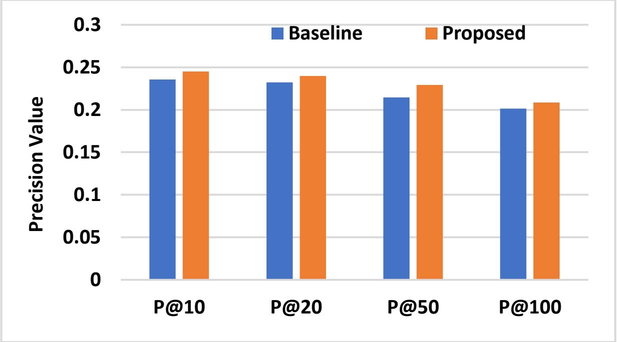

Fig. 2. Precision Comparison between Baseline and Proposed on ANTIQUE.

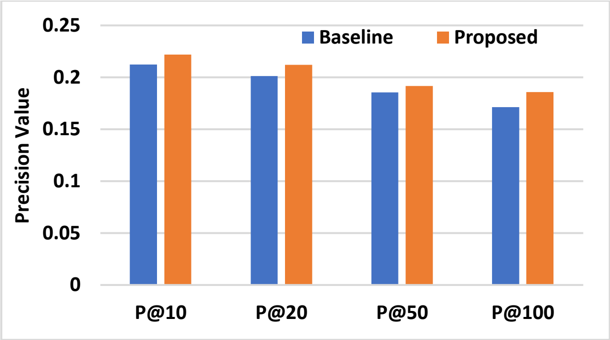

Fig. 3. Precision Comparison for Baseline and Proposed System on Cranfield.

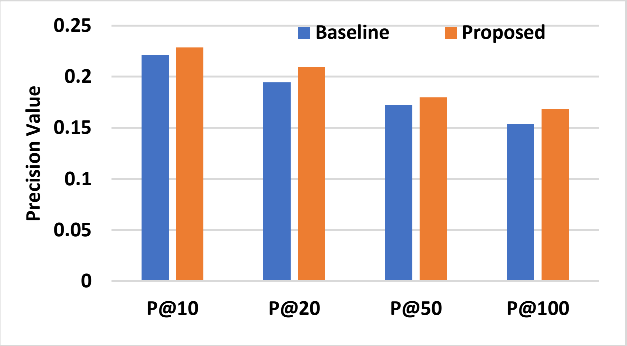

Fig. 4. Precision Comparison for Baseline and Proposed System on TREC MRT.

B. *Recall Analysis*

The results for recall measures on the ANTIQUE dataset are shown in Fig. 5. Here you can see that the proposed system outperformed the baseline. For R@10, the baseline achieved a value of 0.3234, and the proposed system achieved 0.4321. The values for baseline and proposed at R@20 were 0.3745 and 0.4023. The best value of R@50 is 0.4421 given by the proposed approach. For R@100, baseline achieved 0.5122, but our semantics method gave 0.5622. In Fig. 6 comparison between baseline and proposed on Cranfield is presented. The graphical representation in Fig. 7 shows that the proposed methods have beaten the baseline for TREC MRT results.

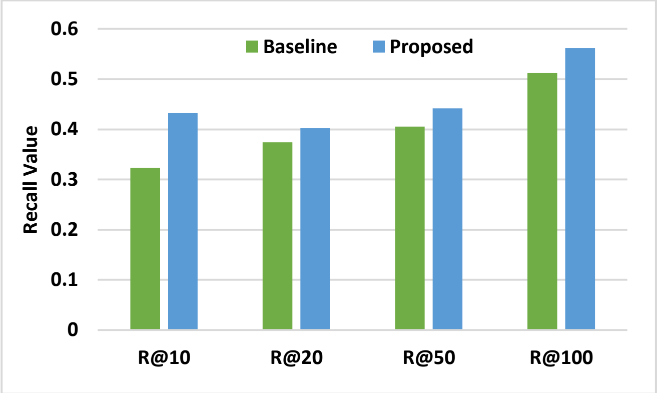

Fig. 5. Recall Comparison between Baseline and Proposed on ANTIQUE.

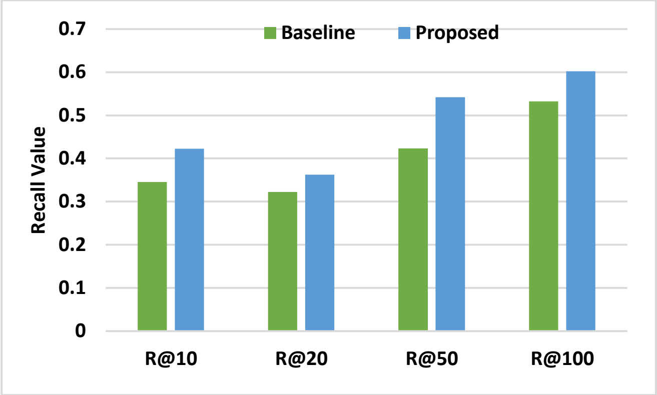

Fig. 6. Recall Comparison between Baseline and Proposed on Cranfield.

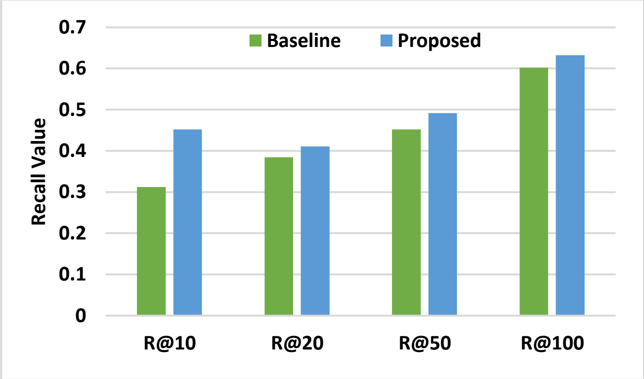

Fig. 7. Recall Comparison between Baseline and Proposed on TREC MRT.

C. NDCG Analysis

From Fig. 8, you can see NDCG results for the ANTIQUE dataset, for cut10 proposed approach gave 0.2168, whereas for baseline value was 0.2101. In comparing cut20 results, the authors found that the difference between baseline and proposed is less than 0.0028. For cut50, the value for the ANTIQUE dataset was 0.2045 in the baseline model, whereas in the proposed system corresponding value was 0.2128. For cut100 value for the proposed was 0.2162 higher than the baseline. Fig. 9 and Fig. 10 show NDCG results for the Cranfield and TREC MRT dataset, respectively; from these comparisons, the authors found that the ontology method outperformed the baseline.

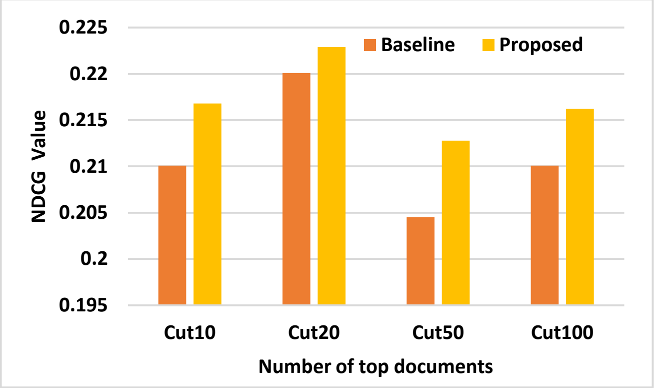

Fig. 8. NDCG Comparison between Baseline and Proposed on ANTIQUE.

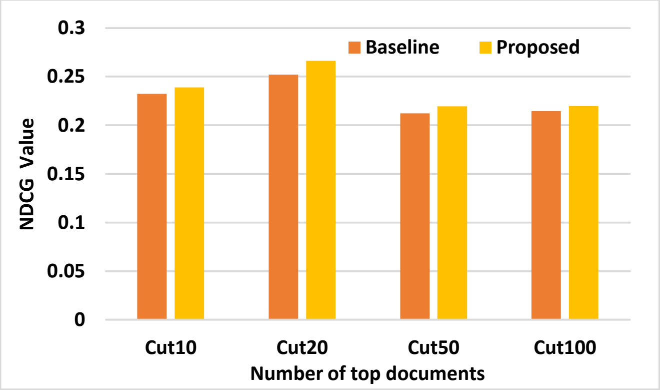

Fig. 9. NDCG Comparison between Baseline and Proposed on Cranfield.

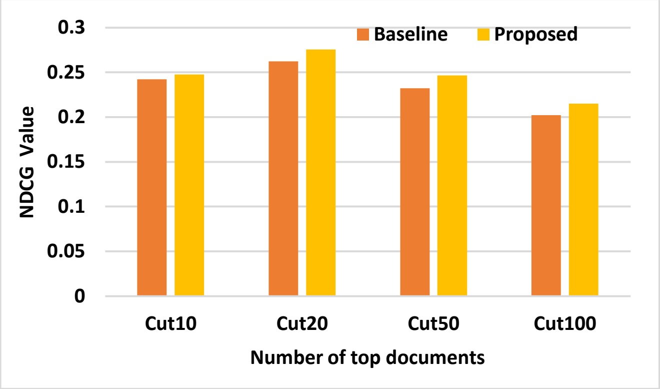

Fig. 10. NDCG Comparison between Baseline and Proposed on TREC MRT.

VI. Conclusion

The results of ontology-based query transformation have shown that if any IRS uses domain ontologies for query processing, its retrieval performance improves. Similarly, semantic association in the indexing process helped to capture the related terms from documents. Hence both modules of the proposed system abetted to retrieve relevant documents as per the user's information need. The different comparisons between the proposed and the baseline results proved that the author's hypothesis is correct. The use of ontology and semantics techniques delivered better results on all three datasets. The proposed system achieved 4% high precision at 100 top documents for ANTIQUE dataset. The recall values on Cranfield dataset at top documents increases by 13%. The NDCG parameters values at top 100 retrieved documents is 6% higher than baseline results. So, it is undeniably an efficient system for retrieving relevant documents from the extensive collection of unstructured data.

The authors conclude that query transformation using ontology has a high impact on retrieval performance. Use of semantic matching and domain-specific knowledge helped IR users to find documents that satisfy their information need. The authors hope that commercial search engines will utilize the full benefit of domain ontology and semantics in near future.

References

[1] M.J.H. Mughal, “Data mining: Web data mining techniques, tools, and algorithms: An overview,” Information Retrieval, 9(6), 2018.

[2] R. Kumar and S. C. Sharma, “Information retrieval system: An overview, issues, and challenges,” International Journal of Technology Diffusion (IJTD), 9(1), pp. 1–10, 2018.

[3] F. Liang, C. Qian, W. G. Hatcher, and W. Yu, “Search engine for the internet of things: Lessons from web search, vision, and opportunities,” IEEE Access, 7, 104673-104691, 2019.

[4] Tim Berners-Lee, James Hendler, and Ora Lassila. "The semantic web," Scientific American, Issue 284, Vol. no. 5, pp. 34-43, 2001.

[5] Thomas R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, 5(2), pp. 199-220, 1993.

[6] Maxat Kulmanov, Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf, "Semantic similarity and machine learning with ontologies," Briefings in bioinformatics, 22(4), pp. bbaa199, 2021.

[7] Shivani Jain, K. R. Seeja, and Rajni Jindal. "A fuzzy ontology framework in information retrieval using semantic query expansion." International Journal of Information Management Data Insights, 1(1) pp. 10009, 2021.

[8] J. Singh, M. Prasad, K. O., E. M. Joo, A.K. Saxena, and C.T. Lin, “A novel fuzzy logic model for pseudo-relevance feedback-based query expansion.” International Journal of Fuzzy Systems, 18(6), pp. 980–989, 2016.

[9] R.E. Prilipsky and M.A. Zaeva, “A Hybrid System for building a Personal Knowledge Base,” Procedia Computer Science, Vol. no.169, pp.96-99. 2020.

[10] 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.

[11] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, “Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering,” Information Sciences, Vol. no. 514, pp.88-105, 2020.

[12] Claudio Carpineto, and Romano Giovanni, "A survey of automatic query expansion in information retrieval." ACM Computing Surveys (CSUR), 44(1) pp. 1-50, 2012.

[13] D. K. Sharma, R. Pamula, D. S. Chauhan, “A contemporary combined approach for query expansion,” Multimedia Tools and Applications, 2020.

[14] H. Scells, G. Zuccon, and B. Koopman, “Automatic Boolean query refinement for systematic review literature search,” In The World Wide Web conference, pp. 1646-1656, 2019.

[15] Makoto P. Kato, Tetsuya Sakai, and Katsumi Tanaka, "When do people use query suggestion? A query suggestion log analysis," Information Retrieval, 16(6), pp. 725-746, 2013.

[16] C. Macdonald, R. McCreadie, R.L. Santos, I. Ounis, “From puppy to maturity: Experiences in eveloping terrier,” Proc. of OSIR at SIGIR, pp.60– 63, 2012.

[17] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft, "ANTIQUE: A non-factoid question answering benchmark," In European Conference on Information Retrieval, pp. 166-173. 2020.

[18] Cyril Cleverdon, "The Cranfield tests on index language devices," In Readings in information retrieval, pp. 47-59. 1997.

[19] Nicola Ferro and Peters Carol, “Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF.” Vol. 41. Springer, 2019.

[20] L. Azzopardi, P. Thomas, and N. Craswell, “Measuring the utility of search engine result pages: an information foraging based measure,” In The 41st international ACM SIGIR Conference on research & development in information retrieval, pp. 605-614. 2018.

[21] 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.

[22] T. Sakai, “On the instability of diminishing return IR measures,” In European Conference on Information Retrieval, pp. 572-586, Springer, Cham. 2021.