DESIDOC Journal of Library & Information Technology, Vol. 40, No. 2, March 2020, pp. 89-96, DOI : 10.14429/djlit.40.2.15213

Comparative Analysis of Information Retrieval using Ontology Based vs Traditional Information Systems in Food Science Domain

Padmavathi T.

CSIR-Central Food Technological Research Institute, Mysuru - 570 020, India
E-mail: padmavathit@yahoo.com

ABSTRACT

The current methods of searching and information retrieval are imprecise, often yielding results in tens of thousands of web pages. Extraction of the actual information needed often requires extensive manual browsing of retrieved documents. In order to address these drawbacks, this paper introduces an implementation in the field of food science of the ontology-based information retrieval system, and comparison is made with conventional information systems. The ontology of Food Semantic Web Knowledge Base (FSWKB) was built using the Protégé framework which supports two main models of ontology through the editors Protégé-Frames and Protégé-OWL. The FSWKB is composed of two heterogeneous ontologies, and these are merged and processed on a separate server application making use of the Apache Jena Fuseki an SPARQL server offering SPARQL endpoint. The experimental results indicated that ontology-based information systems are more effective in terms of their retrieval capability compared to the more conventional information retrieval systems. The retrieval effectiveness was measured in terms of precision and recall. The results of the work showed that traditional search results in average precision and recall levels of 0.92 and 0.18. The ontology-based test for precision and recall has average rates of 0.96 and 0.97.

Keywords: Information retrieval; Ontologies; Information systems-traditional; Food science; Semantic web technologies; Apache jena fuseki.

1. INTRODUCTION

Huge data availability in each subject area makes it extremely difficult for users to integrate and understand web - based data using traditional information systems and search engines. The principle purpose behind this circumstance is that search engines are based on searching for keywords that are not suitable for accurate retrieval of information. In spite of the fact that the web has turned out to be considerably more intuitive in recent years using social networking platforms, the basic standards stay unaltered, driving traditional information retrieval systems (IRS) to do just keyword matching. The challenge is to initially upgrade the current web in which machines are just equipped for presenting data stored and are not capable of understanding it. It is therefore important to influence the machine in order to understand the content of documents and also the user question in order to be able to better link the web of data. The solution is to support and implement Semantic Web Technologies and principles pioneered by the founder of the World Wide Web, Tim Berner's Lee[1].

Increasing data availability in each field makes the incorporation and interpretation of web-data by conventional information systems and search engines exceedingly difficult for users. The main goal is that search engines are focused on keywords that are not appropriate for correct knowledge recruitment.

The problems related to information retrieval as described above can be overcome by using an ontology-based search approach. Ontologies are commonly used for the specification and explication of concepts and relationships related to a given domain[2]. In Food Science & Technology (FST), as in any other domain, prior knowledge is of utmost importance in the discovery of new knowledge, so the knowledge should have as much expressive power as possible. In the proposed study, ontology is used to represent the background knowledge about the FST to represent the set of concepts, relations and attributes of that domain. The domain knowledge thus represented helps to improve the relevancy of information retrieval from the database and SPARQL query language for retrieving meaningful information from the ontology.

The scope of this work was to create an information system containing the knowledge recorded in the knowledge base of Food Science which will facilitate searching for any query.

This research work established an information system that was intelligent than keyword-based approach to achieve accurate results from the point of view of the user. In conceptualizing and formalizing the information structure of the FST domain an ontological approach was chosen. That doesn't mean, though, that there were no limits. it is difficult to move radically from

Received : 16 October 2019, Revised : 06 March 2020
Accepted : 20 March 2020, Online published : 27 March 2020

the existing information system to semantic based approach, but instead it involves a gradual approach incorporating the benefits of conceptual and ontological approaches. Another major limitation is the minimal ability of ontological formalisms to express themselves. Although they are far more powerful than the thesauri, there are still many important aspects which cannot be modeled in present-day ontology languages. This research work evaluated the following hypotheses:

- Ontologies allow domain information to be stored in a way that is much more sophisticated than thesauri. Therefore, we conclude that a major improvement in the effectiveness of retrieval can be calculated by using ontologies in IR systems. The more accurate an ontology models the application domain, the greater the advantage in the efficacy of the retrieval.

The remaining sections of the paper are organised as follows. In Section 2, related works is discussed. In Section 3, ontology-based information system is explained. In Section 4, the system is validated by ontology searching with test cases. In Section 5, the system performance by comparing the effectiveness of our method between Protégé versus Fuseki; traditional keyword search and ontology search is evaluated. The paper concludes by briefly describing future works.

2. LITERATURE REVIEW

Compared to internet search engines, Kim[3] assessed the efficiency of a Web retrieval system based on ontology. Their study showed that ontologies that provide a domain conceptualisation could not only be used to improve accuracy, but also to reduce the search time.

Mustafa[4] suggests a system for the retrieval of semantic knowledge for enhancing the accuracy of results retrieved. In order to capture the meaning of particular concept(s), thematic similarity technique was used for the information retrieval. Source(s) metadata information is stored as RDF triples. Queries were executed against RDF triples rather than text attributes of metadata. The experiments indicated tremendous improvement in accuracy as compared to currently available ontology based retrieval algorithms.

Bikakis[5] proposed a hybrid search method to overcome the disadvantages of traditional keyword and semantic document annotation and retrieval searches. It supports both manual and automatic annotation of documents using ontologies to help the user increase the resulting list to obtain results of high quality. A user - based assessment showed that in terms of precision and recall, the hybrid search was better than the keyword - based and semantic - based search.

In the comparative study of semantic versus keyword based search engines by Tumer[6] results from Google, Yahoo and Msn were analysed against the Hakia semantic search engine. In this study, ten queries from dissimilar subjects and four sentences with similar meaning but with different syntax were considered. To evaluate these two types of search engines, the accuracy and normalised recall ratios were computed at different intervals. The study showed that results from Yahoo were more accurate while Google results had better normalised recall ratio.

To improve the efficiency of online information retrieval, Kang and Jao[7] proposed smart agricultural information search technology based on ontology. To make the structure of the ontology of agriculture, the knowledge of the agricultural domain and the retrieval of semantics are analysed. The results showed that the use of agricultural ontology technology in the retrieval of agricultural information enhances the smart retrieval of agricultural information but also significantly improves the accuracy and reliability of the retrieval of agricultural information.

By collecting and analysing vegetable e-commerce domain information on the web, Teng and Ming[8] developed a system for information retrieval. The ontology consists of certain types of classes of vegetables and the hierarchy of classes of vegetables. During the information retrieval process, the domain ontology helped to index and infer information. The model implemented had more features than the web information retrieval engines based on keywords. The results showed that the ontology-based information retrieval model's recall and precision ratio is higher than the keyword-based information retrieval model.

Kamran and Sheraz[9] explored the tools and techniques available for executing database search queries against ontology based systems. They also analysed the current ontology-to-database transformation and mapping methods with regard to data loss and semantics, structural mapping and applicability of domain knowledge. The results showed that ontology and relational models can bridge the gap by using ontologies to generate accurate search requests.

Binbin[10] has introduced a domain ontology information retrieval model to expand ontology through query expansion into the traditional data retrieval model to improve performance. The method consists of two stages: ontology document processing and ontology document retrieval. Document processing extracts valuable knowledge from unstructured text messages and creates a mapping relationship between the terms of the document and concepts based on the ontology of the domain. The user input search terms to the search interface in the ontology document retrieval, which excludes stop words and preserves only the name and the verb. The term word extraction is used to produce semantic conceptual words and phrases. They used genetic algorithm approach to determine the best weighting factor for the retrieval system.

3. ONTOLOGY-BASED INFORMATION SYSTEM

The ontology of the Food Semantic Web Knowledge Base (FSWKB) is designed using the Protégé software, which supports two main ontology models through the Protégé-Frames and Protégé-OWL editors. The ontology is populated using Protégé to check the application's performance, as it can be exported to standard formats including RDF, RDFS, OWL, and XML Schema.

The ontology test library consisted of data sets of different sizes and different domains. The focus was on scalability, i.e., the ability of matches to deal with data sets of an increasing number of elements. Scalability was evaluated for two seed

ontologies (bibliography ontology and food ontology) of different sizes. The ontology size is indicated in Table I.

Table 1. Ontology size (Bibliographic and food ontology)

Test set	Bibliographic ontology	Food ontology
classes+prop	341	962
Instances	112	214
Entities	508	209

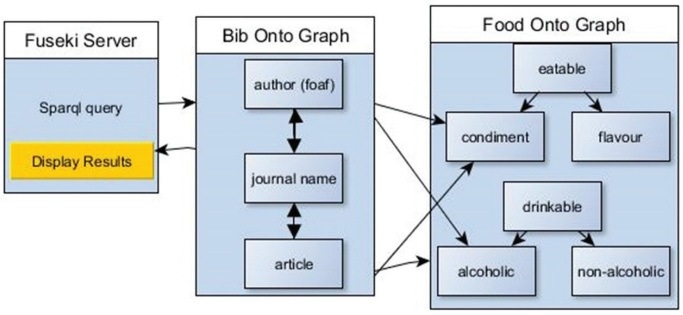

Figure 1. FSWKB framework.

3.1 FSWKB Components

FSWKB has three main components as shown in Fig.1. In the diagram first dataset is food ontology and the second dataset is bibliographic ontology which together represents the Knowledge Base (KB).

3.2 Implementation

Ontology processing was carried out using the Apache Jena on a separate server application. Apache Jena Fuseki is an endpoint SPARQL server. An SPARQL endpoint can be understood as an interface that can be accessed by users (human or application) by using SPARQL query language to query an RDF data store.

After building the knowledge base, it was important to query the system and examine the responses. The FSWKB can generate new relationships based on the data and possible inconsistencies in the (integrated) data. When a user sends a search request, it takes the query and collects the facts from the FSWKB. After the query is processed based on the given facts, it generates new knowledge and sends a response to the user through the web user interface.

Ontology processing was implemented by making use of Apache-Jena-Fuseki server for accessing the data in the knowledge base. It supports data retrieval from multiple graphs and creates communication with the Web client. The two ontologies have been merged using this server to transform the whole knowledge base to RDF triples, i.e., subject, predicate, and object. The information is accessed via SPARQL query language.

3.3 Query RDF Graphs

The SPARQL queries and (SPARUL) updates are sent to Fuseki using simple HTTP requests to get responses in various formats (JavaScript Object Notation (JSON), XML and Comma Separated Values (CSV) for instance). To evaluate the system, the following simple query was tested in a knowledge base. The result of the query in Fuseki displays 25 triples as shown in Fig. 2.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?Article ?Descriptors
WHERE {
 ?class a owl:Class.
```

```
1  prefix owl: <http://www.w3.org/2002/07/owl#>
2  prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3
4  SELECT DISTINCT ?class ?label ?description
5  WHERE {
6    ?class a owl:Class.
7    OPTIONAL { ?class rdfs:label ?label}
8    OPTIONAL { ?class rdfs:comment ?description}
9  }
10 LIMIT 25
```

QUERY RESULTS

Raw Response | Table

Search: Show All entries

	class	label	description
1	<http://purl.org/ontology/bibo/Document>	"Document"^^	"A document (noun) is a bounded physical representation of body of information designed with the capacity (and usually intent) to communicate. A document may manifest symbolic, diagrammatic or sensory-representational information."^^
2	<http://purl.org/ontology/bibo/Thesis>	"Thesis"^^	"A document created to summarize research findings associated with the completion of an academic degree."

Figure 2. Results of query in Fuseki.

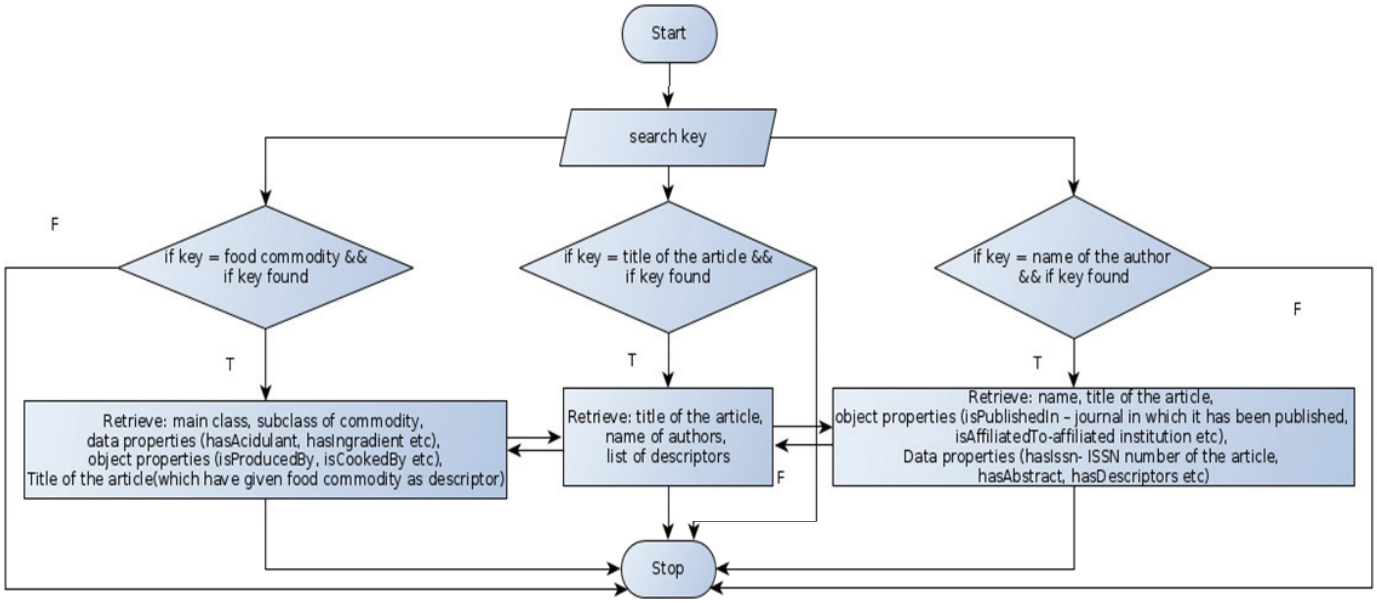

Figure 3: Searching process established for the system.

Table 2. Comparison with time constraints: Time taken by Protege versus Fuseki

Test Entity	Time taken in Protégé		Total time taken for search in Protégé (A+B)	Time taken by Fuseki server	Comparative Efficiency
	Search in Bibliographic Ontology (A) (in ms)	Search in Food Ontology (B) in (ms)			
Ester	0.048263	0.039179	0.087442	0.068124	1 : 1.283576
Egg	0.072359	0.034123	0.106482	0.072144	1 : 1.475965
Algae	0.037146	0.034812	0.071958	0.061824	1 : 1.163911
Sweet Pea	0.041223	0.052146	0.093369	0.071835	1 : 1.29977
Honey	0.071545	0.043123	0.114668	0.091038	1 : 1.259562
Cabbage	0.051038	0.034514	0.085552	0.073423	1 : 1.165193
Beet	0.043468	0.041265	0.084733	0.070124	1 : 1.208331
Copper	0.043123	0.071824	0.114947	0.080234	1 : 1.432649
Citric	0.03462	0.023445	0.058065	0.048231	1 : 1.203894
Skin (animal organ)	0.034156	0.077232	0.111388	0.09841	1 : 1.131877
Parsley	0.071835	0.041223	0.113058	0.081242	1 : 1.39162
Water	0.034812	0.039872	0.074684	0.062345	1 : 1.197915
Potato	0.041223	0.054245	0.095468	0.079341	1 : 1.203262
Ran Li	0.051038	0.034145	0.085183	0.065492	1 : 1.300663
Amudha Senthil	0.023861	0.045612	0.069473	0.041241	1 : 1.684561
Ravishankar	0.034572	0.034581	0.069153	0.042628	1 : 1.622244
Pilar	0.062688	0.045246	0.107934	0.082032	1 : 1.315755
Shah	0.034614	0.051037	0.085651	0.072362	1 : 1.183646
Xin Li	.072681	0.035246	0.107927	0.087302	1 : 1.236249
Prakash Halami	0.020415	0.040532	0.060947	0.048424	1 : 1.258603

```
OPTIONAL { ?class rdfs:Article ?Author}
OPTIONAL { ?class rdfs:Descriptors ?Descriptors }
}
LIMIT 25
```

4. ONTOLOGY SEARCHING

It provides a search system applied to a domain ontology based on applications where the user looks for ontology instances instead of searching for specific web pages. Initially, the searching of ontology entities was performed in Protégé limited to a specific graph only. Nevertheless, it was not possible to search for ontology entities from 2 or more graphs in protégé. Therefore, in our framework, we have built the graphs in such a way that more than 2 ontologies can be accessed in fuseki server together as a dataset that provides querying facilities via sparql queries. We have added 2 ontologies (food & biblio) to a specific dataset and searched for the entities that belong to both of them, and obtained the combined output that clearly demonstrated that we can integrate heterogeneous ontologies by the definition of the domains to which they belong and fetch combined output.

A user can query the system by choosing one of the types of queries: journal, commodity, article and author. The search process established for the system is shown in Fig. 3.

5. SYSTEM COMPARISON

The following metrics proposed by Guo[11] have been used in the present work for analysis:

- Load time: The time taken by the computer to load a data set into memory or permanent storage. Some systems do TBox or ABox inference during the initial loading which is known as reasoning time.
- Query response time: The time needed to submit a question, the result set, and the results navigated iteratively.
- Completeness and soundness: The completeness and soundness is measured in terms of recall and precision in response to a systems answer to a query.

5.1 Experimental conditions

5.1.1 Test case 1: Comparison with time constraints: Time Taken by Protege versus Fuseki

The performance of the system was tested with different entities like food commodity in food ontology, author in bibliographic ontology and the time taken by Protege versus Fuseki was compared. Table 2 shows the observed results showing different levels of method performance in different cases. The metrics are based on following criteria measured in terms of milliseconds:

- Fetch the time taken for searching an entity in bibliographic and food graph (Protégé)
- For fetching the total time taken for a particular entity, we have taken the sum of two-time values which we got earlier i.e. total time taken for searching an entity both in bibliographic and food graph (Protégé)
- Similarly, we have to fetch the time taken by fuseki server to search the same entity
- The formula calculated efficiency

$$\frac{\text{Time taken for search (bibliographic + food graph)}}{\text{Time taken for search (fuseki server)}}$$

- The results in table indicated that the values retrieved prove that the efficiency lies within the range of 1.2 to 1.6 on an average the efficiency is 1.35 which shows that searching (fuseki server) using this framework in comparison with Protégé is 1:1.35.

Figures 4 and Fig. 5 show the system efficiency measured in the form of comparative graphs i.e. time taken by Protege versus Fuseki.

5.1.2 Test Case 2: Effectiveness: Accuracy and Time Constraints

The efficiency of ontology use was measured in terms of search query results being accurate and recalled. A perfect 1.0

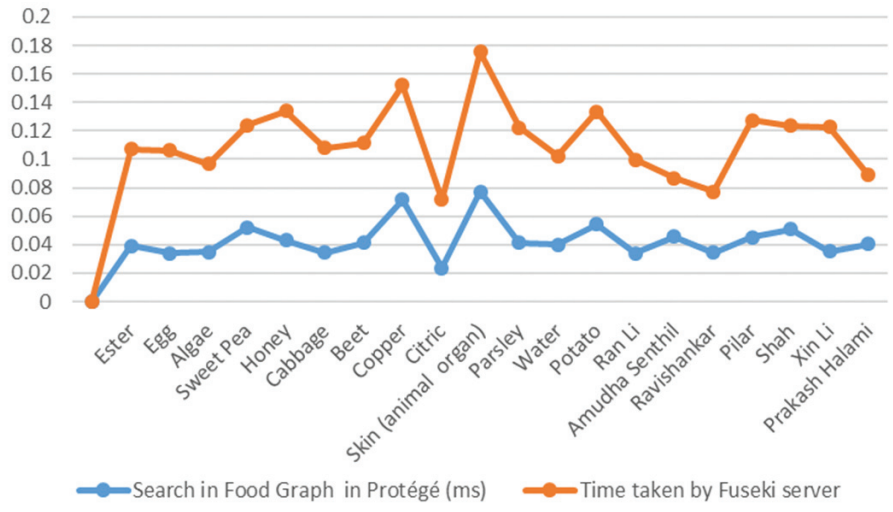

Figure 4. Comparison graph 1: Protégé vs Fuseki.

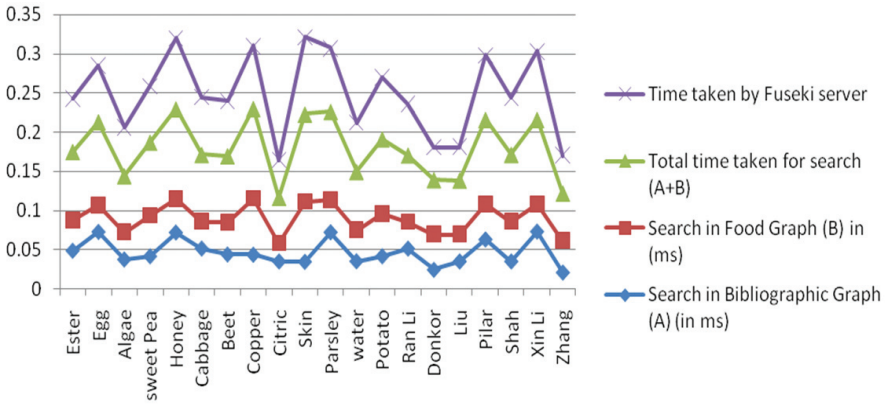

Figure 5. Comparison graph 2: Time efficiency (Accuracy).

accuracy score indicates that each search result was relevant, while a perfect 1.0 recall score indicates that all relevant documents were retrieved by the search.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

A set of 13 queries was used to retrieve information from the traditional food technology database that is being maintained at FOSTIS and the ontologies that were built for this research. The experiment compared a traditional search (Annexure '1') with ontology-based query expansion. The retrieval effectiveness was measured in terms of precision and recall. The results showed that the rates of accuracy and recall increased from 0.92 to 0.96 and 0.18 to 0.97 on average. The precision and recall are taken to know the F-Measure which was calculated using the formula F1 Score = 2*(Recall * Precision) / (Recall + Precision).

Query 1. Diabetic who must reduce their rice intake

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX food:http://ir.cftri.com/fostis/fswkb/ontologies/food.owl#
PREFIX ingredient: http://ir.cftri.com/fostis/fswkb/ontologies/contains
SELECT DISTINCT ?food
WHERE {
?food rice:contains contains:carbohydrates .
FILTER NOT EXISTS{
?food rice:contains ?contains .
?contains rice:hasGlycemicIndex ?GI .
FILTER (?GI <= 40)}
}
```

Table 4. Search result (Average)

Approach	Relevant	Retrieved	Retrieved and Relevant	Precision	Recall	F Score
Traditional Search (Keyword)	317.75	29.5	26.16	0.92	0.18	1.33
Ontology Search	317.75	342.15	340.69	0.96	0.97	2.93

6. DISCUSSSION

The FST ontology was evaluated in terms of time constraints and precision and recall rate as retrieval efficiency. The differences between the Protégé-based search individually in two graphs and Fuseki server search by combining two graphs is lesser than the time is taken for the search in two graphs. So the efficiency of the combined approach is 20 per cent - 30 per cent higher as compared to Protégé. The efficiency also depends on the system configuration.

Highly significant were the differences between traditional search and ontology-based search. Traditional search results have average rates of 0.92 and 0.18 for precision and recall. The ontology-based search has average rates of 0.96 and 0.97 for precision and recall. Consequently, the ontology-based search's precision and recall rates are higher than conventional search (Table 4). Relevant records by some entities are also higher than recovered records in some of the search results. This is because the concepts, terms and individuals that have been added in ontology are not sufficiently comprehensive.

7. CONCLUSIONS

The research discussed in this paper has shown that works in the field of information search and retrieval, especially semantic web, have not yet taken full advantage of the technology, knowledge and experience gained through several decades of work in the field of IR tradition. It is proposed that the understanding of ground ideas in both areas might sometimes have some variations, but important possibilities for study would therefore lie in combining mutually beneficial observations from both fields.

This paper has presented the FSWKB framework. The use of OWL as an implementation language Protégé as an ontology editor and Apache Jena Fuseki server for interconnectivity between two heterogeneous ontologies was shown. The functionality of these system components has been demonstrated by a query for different entities. Complex semantic relations could be managed more effectively compared to the relational database management system (RDBMS). The current method of information retrieval techniques, i.e., keyword-based search retrieves poor quality search results with low precision as they do not encompass domain knowledge or not able to consider the context of the user query. A lot of irrelevant results are retrieved. Hence, there is a need for a system that was more intelligent than the keyword-based approach to retrieve precise results which would be more relevant to the user's interest. An ontological approach was chosen in conceptualizing and formalizing the FST domain knowledge structure. In order to improve and provide more comprehensive search results to serve the needs of the user, we may consider populating the FSWKB as one of the major future work.

REFERENCES

1. Berner's Lee, T.; Lassila, H. & Hendler, J. The Semantic Web. *Sci. Am.*, 2001, **284**(5), 34-43.
2. Studer, R.; Benjamins, R. & Fensel, D. Knowledge engineering: Principles and methods. *Data Knowl. Eng.*, 1998, **25**(1/2), 161–198. doi: 10.1016/S0169-023X(97)00056-6.
3. Kim, H.H. ONTOWEB: Implementing an ontology - Based web retrieval system. *J. Am. Soc. Inf. Sci. Technol.*, 2005, **56**(11), 1167–1176. doi: 10.1002/asi.20220.
4. Mustafa, J.; Sharifullah, K. & Khalid, L. Ontology based semantic information retrieval. *In* 4th International IEEE Conference Intelligent Systems, 6-8 September 2008.

doi: 10.1109/IS.2008.4670473.
5. Bikakis, N.; Giannopoulos, G.; Dalamagas, T. & Sellis, T. Integrating keywords and semantics on document annotation and search. *In* Proc. of the 2010 International conference on the move to meaningful internet systems: Part II, 2010, Berlin, Heidelberg. pp.921-938. doi: 10.1007/978-3-642-16949-6_19.
6. Tumer, D.; Shah, M.A. & Bitirim, Y. An Empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, Yahoo, Msn and Hakia. *In* 4th International Conference on Internet Monitoring and Protection, 24-28 May 2009.
7. Kang, J. & Gao, J. Application of ontology technology in agricultural information retrieval. *In* Proc. of the 2nd International Conference on Computer and Information Application, 2012, Taiyuan, China. pp.1183-1186. doi: 10.2991/iccia.2012.292.
8. Teng-yang, T. & Ming, Z. An ontology-based information retrieval model for vegetables e-commerce. *J. Integr. Agric.*, 2012, **11**(5), 800-807. doi: 10.1016/S2095-3119(12)60070-7.
9. Kamran, Munir. & Sheraz Anjum, M. The use of ontologies for effective knowledge modelling and information retrieval. *Appl. Comput. Inf.*, 2018, **14**(2), 116-126. doi: 10.1016/j.aci.2017.07.003.
10. Yu, B. Research on information retrieval model based on ontology. *J. Wireless Com. Network*, 2019, **30**. doi: 10.1186/s13638-019-1354-z.
11. Guo, Y.; Pan, Z. & Heflin, J. An evaluation of knowledge base systems for large OWL datasets. *In* 3rd International semantic web conference, 7-11 November 2004, Hiroshima, Japan. 2004. doi: 10.1007/978-3-540-30475-3_20.

CONTRIBUTOR

Dr T. Padmavathi is Principal Technical Officer at CSIR-Central Food Technological Research Institute, FOSTIS/Library, Mysuru. She holds Doctoral Degree from Bharathiar University. Her core areas of expertise include traditional library services, semantic web technologies, computer based information services, digital library technologies, documentation, IT applications (web page design and content development) and library automation.

Annexure '1'

Table 3. Query result from traditional search (Keyword) and ontology search

Test Entity	Relevant	Retrieved		Retrieved and Relevant		Precision		Recall		FScore	
		Traditional search	Ontology search	Traditional search	Ontology search	Traditional search	Ontology search	Traditional search	Ontology search	Traditional search	Ontology search
Rice and starch	640	82	643	82	634	1	0.97	0.12	0.99	0.21	0.98
Egg and products	327	15	323	15	323	1	1.00	0.04	0.97	0.08	0.98
Algae and red	63	6	61	5	60	0.8	0.96	0.07	0.90	0.13	0.93
Pea and chick pea	672	67	670	67	670	1	1.00	0.09	0.99	0.17	0.99
Honey and product	177	5	176	5	175	1	0.98	0.02	0.97	0.04	0.97
Cabbage and red	60	16	60	14	60	0.8	1.00	0.23	1.00	0.36	1.00
Beet and sugar	78	30	76	30	76	1	1.00	0.38	0.94	0.55	0.97
Copper and catalyzed	151	41	151	41	151	1	1.00	0.27	1.00	0.43	1.00
Citric and acid	36	34	43	33	36	0.97	0.70	0.94	1.00	0.95	0.82
Wheat and product	765	33	765	33	765	1	1.00	0.04	1.00	0.08	1.00
Parsley and shelf life	12	1	12	1	12	1	1.00	0.08	1.00	0.15	1.00
Water and activity	1119	77	1116	41	1116	0.53	1.00	0.03	0.99	0.06	0.99
Potato and chips	353	29	352	29	351	1	0.99	0.08	0.98	0.15	0.98
Average	317.75	29.5	342.15	26.16	340.69	0.92	0.96	0.18	0.97	0.26	0.97