OSII: An Efficient Multi-Domain Information Retrieval Framework Using Ontology based spatial Inverted Index

P. Sunil Kumar Reddy[1*] and Dr.P.Govindarajulu[2]

[1]Research Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati, India.
[2] Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, India.

Abstract
In recent years, the information retrieval of multi domain data access has become a critical process. Information retrieval is the process of retrieving relevant information based on user query. The conventional information retrieval methods were employed in multiple domains such as medical, instrumentation, mechanical, electrical, and software. Nevertheless, it had limitations such as index based matching methods, reduced accuracy on matching information, and time consumption for voluminous information. In order to solve these issues, this research work aims at developing an ontology based spatial inverted index list mechanism for multi-domain information retrieval. In the initial stage of the proposed work, user loads the query as input and it is preprocessed to eliminate redundant words for keyword extraction. Words are analyzed for similarity and then it's evaluated. Finally, the ontology based ranking methodology is applied to rank the similar information based on user query. Here, the rank is evaluated by integration of both semantic searching and ontology based searching. The performance of proposed work is analyzed with existing works.

Index Terms— information retrieval (IR), multiple domain information, preprocessing, similarity matching

1. Introduction

Ontology plays a significant role in retrieval of similarity document. In fact, it is the foundation of semantic web and specification of shared data. It creates a defined ontology [1] language and it is based on shared knowledge. Sharing and reuse are the necessary properties of ontologies among humans as well as machines. Ontology search mechanism plays a significant role in the discovery in semantic web applications. The applications are agriculture, military, and medical. Various ontologies are constructed by different type of applications. Creating new ontology is a challenging task due to time and cost concerns. Some challenges [2] faced in conventional ontology are

- Discovering the domain ontology for reuse
- Incorporating the concepts into domain using information oriented methodologies
- Different domain knowledge from discovered ontology
- A general forum to share the domain information

Knowledge discovery from a large volume of unstructured data is an arduous task, even though large number of techniques are available to extract data. Knowledge-Based similarity is utilized to measure and determine the degree of similarity between words via information which are derived from semantic web. The similarity matching technique is used to compute the similarity measures between the documents. Clustering takes place based on the similarity measures which provides meaningful data. The semantic annotation describes the semantic content of information and retrieval of the queries. It needs the general framework to represent the understanding of semantic meaning to retrieve the query and standardization of their representation. It can compare two words between their similarities of query and retrieve the query information.

Information retrieval

Information retrieval is the process of recovering data form [3] specific database. It is very important to retrieve relevant information from the bulk of web documents based on their application or user requirements. It decreases the process of identifying similarity between the information documents with respect to keyword [4] obtained from the query. This information retrieval process is most widely used in the applications of software development [5].

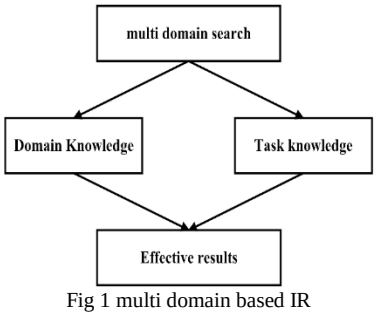

Fig 1 multi domain based IR

Limitation

Ontology search engine handles the multi domain keyword based search mechanism. It is a challenge to accurately represent the user requirements based on of their recall rates, and accuracy rate [6]. This system does not implies the use of ontology reasoning with the aim of searching

Manuscript received April 5, 2018
Manuscript revised April 20, 2018

and retrieving [7] complex information based on complex query.

Objectives

The objectives of the work are as follows:

- To implement a multi domain based searching model to handle different types of data.
- To retrieve or extract the answer from unique area by using ontological database
- To enable the customers for query searching even though they are not familiar with ontology.
- To develop an efficient information retrieval system with reduced complexity and increased accuracy to improve the performance of query processing.

The paper is organized as follows: The detailed description of the related works on information retrieval and ontology based methodologies are discussed in section II. The implementation of proposed work is described in section III. The comparative analysis of proposed method with existing methods is presented in section IV. Finally, the conclusions of proposed work is presented in section V.

2. Related Works

In this section, the existing methodologies of ontology, information retrieval process are discussed along with their merits and demerits. Bobillo [8] stated the reduction of fuzzy ontology perceptive to identify with conventional amenable idioms. This work also analyzed the waves of making an allowance for some concepts and roles with optimal reasoning. Correction was done with the fuzzy ontology language Fuzzy OWL 2 and demarcated some optimizations by implementing selected instructions to catalog some fuzzy ontology features as crisp or non-crisp inevitably. The limitation of this method was that it cannot reflect the human behavior in the ontology comparisons. Classical two valued semantics cannot unswervingly be able to indefinite or elusive smithereens of facts, which are in-built to a number of practical complications. Dou, et al. [9] focused on the importance of ontology, which incorporated with semantic mining. Ontologies were very useful in semantic data mining to bridge semantic gaps between the data, applications, data mining algorithms, and data mining results. It moved the motivation to ontology based clustering in the text mining methods. The context based relations could be improved using ontology based clustering. So the outliers could be easily predicted and time complexity could be reduced in efficient manner. Aljawarneh, et al. [10] introduced a framework for efficiently querying and analyzing big geospatial information which was plugged on top of geospark. The work was further included with types of queries, machine learning, and data mining services with awareness for geospatial analysis. Thaker and Goel [11] presented an ontology driven query processing system. Ontology approaches can be used to find the answers of the query. The information was retrieved by ontology and some additional information was obtained by their adjacent key words. A car ad dataset was used for this experiment. A better result was achieved for unstructured and ungrammatical data. The system required modification in extracted information without offering any additional knowledge. It also required to create ontology automatically from user queries. Kasbe, et al. [12] described the ontology based semantic information retrieval and semantic web framework. The advantage of this work was that it was quite generic, and easily adapted to new domain like library domain to validate the method. It was capable of handling the specific domain. The work required improvement in different domain. Palangi, et al. [13] developed a concept based paragraph vector formation for the effective information retrieval. This semantic vector was formed by applied recurrent neural networks (RNN) with Long Short-Term Memory (LSTM) cells. The meaning of the words in a sentence was extracted and embedded in the semantic vector. The stop and unused words were automatically eliminated in the analysis phase. The LSTM-RNN was prepared in an uncertainly supervised manner using back propagation through time on consumer click-via data logged by web search engine. The paragraph vector was matched with user query information in the semantic conceptual manner. This approach was effective for natural processing language. It required more embedding time and space for generating semantic vector. A theoretical analysis of message complexity of each operation in GeoTrie proved their scalability. The property of GeoTrie was allowed to alleviate potential bottle neck on the root node. The work was extended to handle $n-$dimensional range queries over massive dataset. Nodarakis, et al. [14] offered a novel methodologies for categorizing multidimensional data by using $AKNN$ queries in single batch based procedure in Hadoop. Map reduced methodology was utilized for classifying the multidimensional data. The work was extended and improved to increase their efficiency and flexibility. Guo, et al. [15] introduced a new semantic matching for IR via nonlinear word transportation framework. It was included with two element in the model which made the dissimilar from semantic based models and operative for IR,

- Word alignment effect
- Flexibility in model definition

The results were compared with some retrieval model and it was planned to discover the dissimilar model differences with moving framework. Yih, et al. [16] proposed a novel semantic parsing structure for query responding by knowledge base. Semantic analyzing decreased the query processing and expressed as staged search problem. By

using this dataset, the innovative entity connecting scheme and deep convolutional neural network model were matched with the queries and the sequences were established. Singh and Jain [17] studied the basic concepts of Information Retrieval through Semantic Web. It was very important to retrieve the relevant information from bulk web documents. This work stated the complete overview of the semantic web. It focused various research areas like information retrieval, architecture, and prototype systems. It explained the basic concepts of RDF, URL, Ontology, and OWLIR architecture. It described the prototype search engine Swoogle and their various operation mechanism. Dong and Hussain [18] introduced SASF crawler in order to imply discovering, Formatting, and indexing mining service information using combined methodologies of semantic intensive crawling and ontology learning. The crawling was very useful to avoid ambiguity in the heterogeneous environment. Ontology offered high performance level of searching. The ontology of a keyword id extracted from the vocabulary based ontology and concept-metadata matching algorithm. This method worked in probability and semantic based mining. The advantage of this work was that it provided very relevant services for the user query. The limitations of the system was the concepts of the vocabulary ontology list should be filled manually for every terms. Harispe, et al. [19] presented the unifying framework which was aimed to improve the semantic measures and highlighted their equivalences. The main advantage of the framework was it relied on identification of core elements which was commonly used in the design of SSMs. Akmal, et al. [20] presented the ontology based approach which was identified the similarity between two classes by feature based similarity measure. The outcome of correlation between different semantic similarities and human judgment of web based search was suggested multiple similarity measures to validate ontologies. Müller, et al. [21] stated the challenges in finding ontology concepts for general queries. This work studied various domains and related ontologies. It focused the common ontology concepts that intersect between two domains. In order to distinguish the uniqueness of queries between interrelated domains, it was not possible to create combined ontology of several domains. For example, the drug and agriculture domains were different in nature. But it could be related in some aspects. This research focused such work between inter cross domain. The generic concepts were derived between two ontology concepts and results were verified. The advantage of this work was integrating various concept corpus. Saini, et al. [22] focused on the different models and techniques for information retrieval. This work surveyed various indexing and probability methods that enabled to focus our concentration on information retrieval within the certain range. It analyzed the merits and demerits of such searching techniques based on the measures of time and space complexity. This work also surveyed the application area of IR like Web Search Engine, Multimedia search, Digital Library, and Information Filtering. This work provided all the basic information about the semantic web information retrieval system. Itoh [23] focused on the Semantic Textual Similarity via incorporating word alignment information. It was strongly suggested that it was not enough to measure the similarity and the common phrases between the web documents. This work also concentrated on abbreviation used in the text and bigram alignment. Then the web documents were extracted with various word alignment and commonality was compared with similarity score using regression method. Anuar, et al. [24] Introduced a semantic algorithm to associate feature in terms of conceptual similarity. The algorithm conveyed whole new similarity evaluation conception in the domain of data retrieval. It dealt with problem of trademark which was similar. It handled the issue based on three operations similarity comparison measure. The new concept introduced in this paper was hash indexing. This was done to reduce the computational time at the time of searching. Then the hash index value was planned to list of features from the dataset by using mapping function. Tversky's contrast theory was used to compute the similarity distance between the two objects. This algorithm provided better results in finding the similar trademarks from the database of trademark. But this trademark refers in the meaning of short phrase. It needed improvement for accurate results. Caldarola, et al. [25] developed the ontology reuse based on heterogeneous matching technique. There was several experimentations in specific digital eco systems knowledge base. The system was effectively demonstrated based on linguistic matching could be automatize the selection of most relevant reference model which belongs to specific amount of manual work. Bechhofer and Matentzoglu [26] stated the reduction of fuzzy ontology perceptive to intellectual with conventional amenable idioms. This work also analyzed the waves of making an allowance for some concepts and roles with optimal reasoning. Correction was done with the fuzzy ontology language Fuzzy OWL 2 and demarcated some optimizations by implementing selected instructions to catalog some fuzzy ontology features as crisp or non-crisp inevitably. The limitation of this method cannot reflect the human behavior in the ontology comparisons.

3. Proposed Method

This section illustrates the overall working procedure of proposed work of multi domain information retrieval process. Initially, the user loads the query to language parser engine. Here, the preprocessing technique is applied to remove the stop words. Further, the enriched semantic search keyword engine is used to identify the similar

information based user query. The cosine similarity is applied to calculate the similarity measures. Then, the query related information is retrieved and is stored in the big data server. Finally, the retrieved information is forwarded to the users.

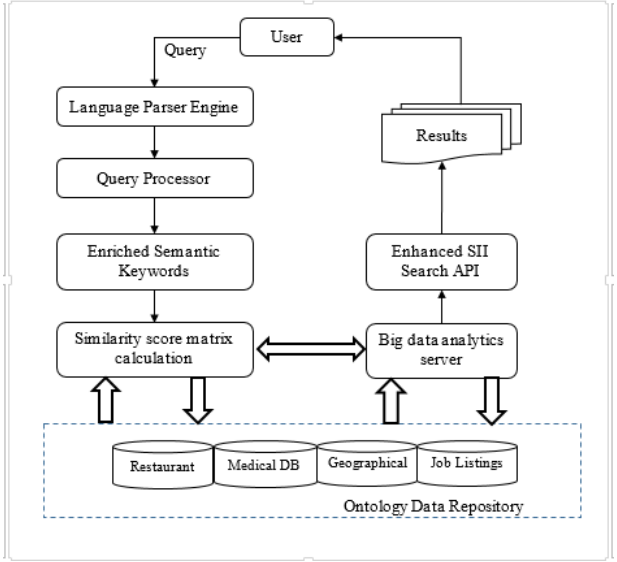

Fig. 2 Overall flow

3.1 Language parser Engine

The input document is parsed and tokens are created using dictionary matching. These tokens are matched with semantic grammar to create patterns. If information is parsed based on a single pattern, it uses single pattern matching algorithm. If information is parsed based on multiple pattern, it uses the multiple pattern string matching algorithm.

3.2 Preprocessing

Preprocessing stage is an important process to reduce the redundancy and improve the work efficiency. Each value of the input dataset [27] is preprocessed and the resultant information is stored in the database. The main intension of stop-word removal process is to obtain the key words or key features from the text document and to improve the relevancy among word and their category. A stop word is a word that has frequent occurrences in the text. Words have little effect on identifying classes of articles, and it also removed. Stop words has little effect on identifying the classes of articles, and it also removed. Stemming is the process to reduce the number of unique terms and it is generally useful to stem terms of their roots. The part of the speech tagging procedure is used to break each document into sentences and utilize the Stanford parser to extricate noun, verb, adverb, and the adjective expressions to evacuate the non-word tokens, for example, numbers, HTML labels, and accentuation. A large portion of the words utilized as a part of words that are extricated by utilizing the information Recovery (IR) technique. These words are called as 'Stop words' that convey no data (i.e., pronouns, relational words, and conjunctions). After preprocessing, the stop words removal and stemming processes are performed on the query information. It will increase the accuracy of entire process. Here, irrelevant information or irrelevant queries are eliminated. The semantic search is executed by utilizing WordNet library to determine the synonyms or relevant information of each preprocessed query.

3.3 Ontology construction and word search

A semantic key word search is used to search and collect the relevant information based on user queries. At this stage, the queries are split into multiple words to search the similar information based on their domain knowledge. A semantic search engine capable of keeping the semantic data on web resources and its handle to solve the complicated user queries. Semantic search include the methodologies of semantic web and search engine for increasing the search results. The word similarity evaluated between two words, based on their attributes, standard deviations, entropy, and information gain.

The mean value is evaluated by equation 1,

$$\mu_i = \frac{\sum_{j=0}^{n} F_j}{n} \quad (1)$$

Standard deviation is computed by

$$\sigma_i = \sqrt{\frac{\sum_{j=0}^{n} (F_j - \mu_i)^2}{n}} \quad (2)$$

Entropy(H_i),

$$H_i = \sum_j^n P(F_j)\, log_2 P(F_j) \quad (3)$$

Information gain $(I(Wo_1, Wo_2))$,

$$I(wo_1, wo_2) = H(wo_1) - H(wo_1 | wo_2) \quad (4)$$

3.4 Similarity calculation

The best result is identified from the dataset, then the keywords are matched with query keywords based on the similarity estimation. The deviance in the input dataset are also included to improve the information in the files. In this paper, Cosine similarity is utilized to estimate the similarity matching between the words of their space and evaluate the cosine angle between them. Cosine similarity is commonly used for positive space while outcome always lies between [0, 1]. Cosine similarity is most suitable for the high-dimensional positive spaces. The queries can be calculated based on cosine angle in n dimensional space between query vector $(W_{d,t})$.

Symmetric uncertainty,

$$Sym(d1,d2) = \sum_{i,j}^{l,m} 2I \frac{(w_i,w_j)}{H(w_i)+H(w_j)} \quad (5)$$

The working procedure of the proposed algorithm is represented as follows:

Ontology based spatial inverted index

To rank a document collection with regard to query q and identify the top r matching document

Step 1: calculate query vector (w_q, t) for each query term t in q

Step 2: for each and every document d in the group

set $S_d \leftarrow 0$

For each query term t

Step 3: Compute or read w_d, d

Set $s_d \leftarrow S_d + w_{q,t} \times w_{q,t}$ (6)

Step 4: Compute or read w_d

Set $S_d \leftarrow {}^{S_d}/_{W_d}$ (7)

Step 5: Categorize the r greastest S_d values and return the corresponding documents

Generate inverted index based ontology

Step 6: To use an inverted index to rank a document collection with regard to a query q and determine the top r matching document

Step 7: An Ontology O_d for each document d and set $O_d \leftarrow 0$

Step 8: for each query term t in q

Step 9: Compute w_q, t and fetch the inverted list for t

For each pair (d, f_d, t) in the inverted list

Step 10: Analyze w_d, t

Set $O_d \leftarrow O_d + w_d, t \times w_d, t$ (8)

Step 11: Read the array of w_d values

Step 12: for each and every $O_d > 0$,

set $S_d \leftarrow {}^{O_d}/_{W_d}$

Step 13: Find the r greastest S_d values return the corresponding documents

To create an inverted index by merge (semantic and ontology) based techniques

Step 14: To build an inverted index by merge based techniques

Step 15: Until all information have been processed

Step 16: Initialize an in memory index and using a dynamic structure for the vocabularies

Step 17: static coding scheme for inverted lists

Step 18: Store lists either vocabularies or a static coding scheme for inverted lists

Step 19: Store lists either in dynamically resized arrays or linked blocks

Step 20: Read the information and insert $(d, f_{d,t})$ pointers into the in memory index, continuing until all allocated memory is consumed

Step 21: Flush the temporary index to disk, including its vocabulary

Step 22: Combine the set of partial indexes to form a single index, compressing the inverted lists if required.

3.5 Ontology based Ranking

The informations are retrieved by rank based on the user queries. In this approach, search engine creates the initial set of ranking and query which helps the user to selects the related documents within that rank based information. The ranking is the method that determines ordering of results of search query. The search requires matching and ranking, wherein matching the information and selection to be scored. The ranking determines the degree of matching using some information of relevance. The ranking is performed after the semantic mapping is executed. The rank will be calculated based on the score of multi domain information.

The ontology based ranking includes TF and IDF score values,

$$Score(q,d) = \propto * Score_{BM25}(q,d) + (1-\propto) * Score_{topic}(q,d) \quad (9)$$

Where,

$Score(q,d)$ – Relevance score of information which is based their query q

$\propto$ – Trading factor

$Score_{BM25}(q,d)$ –Score of TF and IDF

$Score_{topic}(q,d)$ – Score of topic

TF and IDF values are estimated by similarity matching and it uses the information retrieval procedures. The score of BM25 is assigned to the document d and term t is analyzed by,

$$Score_{BM25}(t,d) = IDF_t * TF_{BM25}(t,d) \quad (10)$$

The measure of IDF_t and $TF_{BM25}(t,d)$ are evaluated by,

$$IDF_t = log\left({}^{N}/_{N_t}\right) \quad (11)$$

Here, N – total number of documents

N_t –Number of document that contains term t

$$TF_{BM25}(t,d) = \frac{f_{t,d} \times (k_1+1)}{f_{t,d} + k \times \left((1-b) + b \times \left({}^{l_d}/_{l_{avg}}\right)\right)} \quad (12)$$

$f_{t,d}$ –Frequency term

l_d –Length of document

l_{avg} – Average length of document

k and b – Constant value as 1.2 and 0.75

The scores,

$$Score_{BM25}(q,d) = \left(\frac{\sum_{t \in q} IDF_t * TF_{BM25}(t,d)}{\sum_{t \in q} IDf_t * (k_1+1))}\right) \quad (13)$$

$$Score_{topic}(t,d) = P({}^{d}/_{t}) \propto P({}^{d}/_{t})P(d) \quad (14)$$

Where,

$$Score_{topic}(q,d) = \frac{\sum_{t \in q} \sum_{k=1}^{K} \phi_{kt} * \theta_{dk}}{length\ (q)}$$

Where the length (q) – number of terms in the query
The TF and IDF values are integrated to calculate the relevance score of the document, and it is represented below,

$$Score_{q,d} = \propto * \frac{\lambda_{t \in q} IDF_t * TF_{BM25}(t,d)}{\sum_{t \in q} IDF_t * (k_t + 1)} + (1 - \propto) * \sum_{t \in q} \frac{Score_{topic}(t,d)}{length(q)} \quad (15)$$

Rank correlation,

$$R_w = 1 - \frac{\partial \sum_i^I (p_i)^2}{n(n^2 - 1)} \quad (16)$$

Table 1: Notation description

Notation	Description
μ_i	Mean of I[th] Feature
F_j	Feature Value for i[th] instance of j[th] Instance
n	Number of Instance
σ_i	Standard deviation of I[th] Feature
δ_i	Rank of Deviation
∂_i	Rank of Symmetric uncertainty
y	Different attackers

4. Performance Analysis

The section illustrates the analysis of both existing and proposed techniques. The techniques are analyzed with the performance measures of accuracy, precision, recall, F-measures, execution time, and processing time.

System configuration

Software tools execute in specific computer environments. It is important for the users to be aware of the specific requirements and their selected tool. The proposed work requires Intel (R) Pentium(R) processor and 4GB memory. And also it used the 64 bit operating system and $\times 64$ based processor. The system configuration must be flexible to include a wide variety of machines which are ranging from PC.

Example for Query Processing

Q. 1. Define computer programming using JAVA?
Initially, the query length is five.
Step 1: preprocessing
Stop-word and wh words are removed
Computer programming using JAVA?
Query length is four.
Step 2: Query split into separate word
Computer, Programing, JAVA
Step 3: meanings are found by split words
Computer: processer, CPU, mainframe, processor, Super computer, work station, PC, Laptop, electronic machine, data processor, calculator, digital computer.
Programming: scheduling, planning, scheming, designing, software design, software development.
JAVA: programming Language, coffee cup, syntax, computer language.
When the user does not split the query, the accuracy is increased but the execution time is increased. Here, the user splits the query and ontology based ranking methodology is used to found the query related information. Ranking of information is based on both semantic similarity and ontology based searching methodology. So accuracy is increased and decreased with execution time.

Q 2. What is the use of data mining?
Initially, the query length is seven.
Step 1: preprocessing
Stop-word, wh words and redundant words are removed.
So, the query length is decreased. The query length is three.
Step 2: Query split into separate word
Use, data, mining
Step 3: meanings are found by split words
Use: utilize, employ, apply, usage, benefit, profit, function, exploit, practice, purpose, service
Data: information, facts, evidence, material, proof, knowledge, documentation
Mining: removal, withdrawal, taking out, drawing out, excavating, quarrying
When the user does not split the query accuracy increases but the execution time increases too. Here, the user splits query and ontology based ranking methodology is used to found the query related information. Ranking of information is based on both semantic similarity and ontology based searching methodology. Hence accuracy is increased and decreased with execution time.

4.1. Precision

Precision is determined by the ratio of the sum of retrieved information which are relevant to the search query to the sum of retrieved information. The precision of the retrieved information can be obtained as,

$$Precision = \frac{Total\ number\ of\ relevant\ information\ retrieved}{Total\ number\ of\ information\ retrieved}$$

4.2 Recall

Recall is the relation between the sum of retrieved documents to the sum of the relevant documents. Accuracy is determined by the relation between sum of relevant documents that are retrieved to the sum of documents. It is also referred to as sensitivity and it obtained by using the following equation,

$$\text{Recall} = \frac{Total\ \text{number of documents retrieved}}{Total\ \text{number of relevant documents}}$$

$$Accuracy = \frac{total\ number\ of\ relevant\ documents\ that\ are\ retrieved}{total\ number\ of\ documents}$$

Fig. 3 shows the analysis of precision, recall, and F-measures. Proportion of positive prediction that are correct is termed as precision and recall is the proportion of positive cases identified correctly. The processed technique provides better result because in the preprocessing stage, unwanted information is eliminated so the performance values are increased in proposed techniques.

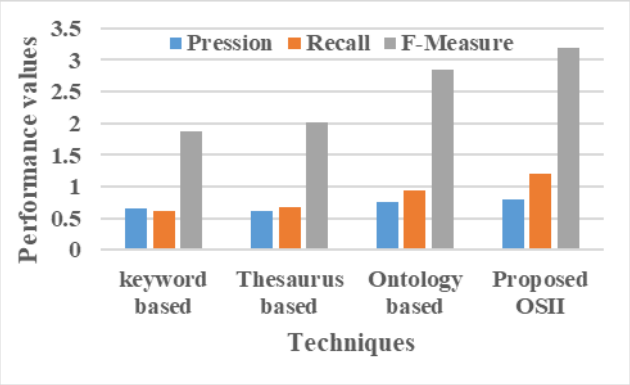

Fig 3 Analysis of precision, recall and F-measures

4.3. Average Precision

The average precision (AP) is applied to identify the quality of search system in the information retrieval process. The average precision for the set of related document is obtained as the mean precision of entire documents.

$$AP = \frac{1}{n}\sum_{i=1}^{n} Precision(P_i) \quad (17)$$

Where P_i is the set of relevant documents

A. F-measure

F-measure is a group of both recall and precision values of a document set for information retrieval.

$$F_n = \frac{2P_nR_n}{P_n+R_n} \quad (18)$$

Rank Score

Rank Score is obtained from the frequency of words that are present in the document.

Semantic score

Semantic score is obtained from the occurrence of relevant documents that are matched in the document.

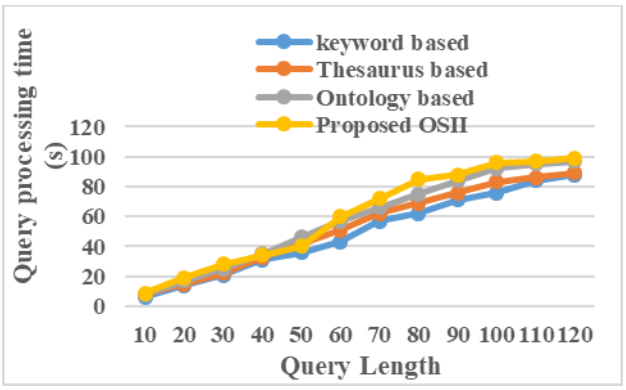

Fig. 4 % of correctness

Proportion of predictions that are correct is termed as correctness. Fig 4 shows the % of correctness of both the existing and proposed techniques. The key based techniques offers the minimum result compared to other techniques. The average correctness values are improved when compared to other existing techniques.

4.4 Rank Accuracy

Rank accuracy is defined as ratio of sum of retrieved documents that are correctly ranked to the sum of total documents.

$$Rank\ accuracy = \frac{No.of\ retrieved\ documents\ that\ are\ correctly\ ranked}{total\ number\ of\ ranked\ documents}$$

4.5 % of Document Retrieval

The percentage of document retrieval is evaluated by the ratio of amount of related documents that are extracted to the total number of documents

$$\%\ of\ document\ retrieval = \frac{No.of\ relevant\ documents\ that\ are\ retrieved}{Total\ number\ of\ documents} * 100$$

4.6 Execution Time

Execution time is defined as the time taken by query information to perform the particular task.

$$execution\ time = ending\ time\ of\ the\ process - initial\ time\ of\ the\ process$$

Fig 5 shows the execution time of both existing and proposed techniques. The execution time is measured in seconds and data size is measured in GB. The execution time is reduced compared to other existing techniques.

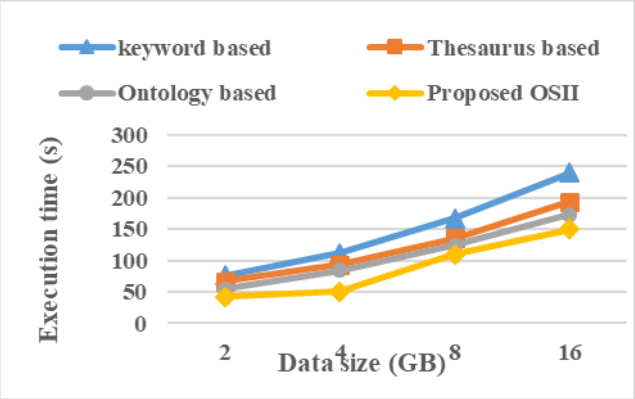

Fig. 5 Execution time

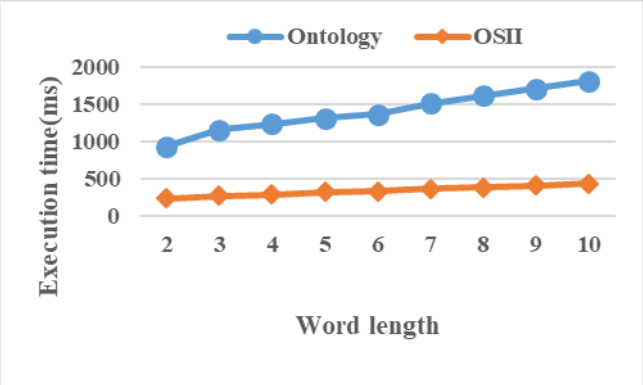

Fig. 6 word length

Fig. 6 represents the execution time for different word length. Word length is defined as the count of bits which are present in the word. Execution time refers to the reading time of user query. Execution time should be increased for an optimum algorithm. Here, the result shows that when the number of words increases the time increases too. The existing ontology techniques take more time compared to the proposed OSII technique. Hence, the proposed OSII technique offers better result.

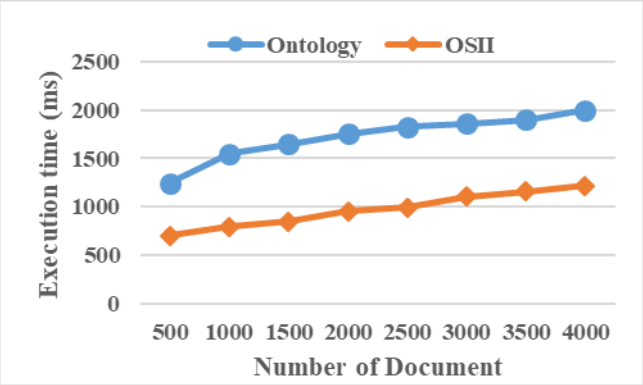

Fig 7 Number of documents

Fig. 7 shows the execution time for number of documents. Execution time should be increased for an optimum algorithm. Here, the result shows that when the number documents increase the time increases too. The existing ontology techniques take more time compared to the proposed OSII technique. Hence, the proposed OSII technique offers superior result.

5. Conclusion and Future Work

This research work developed a novel technique namely Ontology Based Spatial Inverted Index Algorithm based on multi domain information retrieval process. The main objective of this work was to improve the accuracy and reduce time complexity. The preprocessing technique is used to eliminate redundant words like wh and other stop words. The semantic search is executed by using WordNet library to determine the synonyms or relevant information of each preprocessed query. The extracted keywords are used to identify the similarity between two words. Then, the relevant information was retrieved from the database and is aligned via the ontology based ranking mechanism. The ranking of information is based on both semantic similarity and ontology based searching methodology. Thereby, the precision and recall values are increased and the execution time is reduced compared to other existing techniques. The prominent advantage of this technique was to retrieve the information for multiple domains with high accuracy.

References

[1] K. Guo, R. Zhang, and L. Kuang, "TMR: towards an efficient semantic-based heterogeneous transportation media big data retrieval," *Neurocomputing*, vol. 181, pp. 122-131, 2016.

[2] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45-59, 2016.

[3] S. Vigneshwari and M. Aramudhan, "Social information retrieval based on semantic annotation and hashing upon the multiple ontologies," *Indian Journal of Science and Technology*, vol. 8, p. 103, 2015.

[4] P. Sunil Kumar Reddy, and, and P. Govindarajulu, "Adjacent Search Outcomes with Keywords," International Journal of Computer Science and Information Technologies, vol. 7, pp. 312-317, 2016.

[5] D. Binkley and D. Lawrie, "Information retrieval applications in software maintenance and evolution," *Encyclopedia of software engineering*, pp. 454-463, 2010.

[6] P. Sunil Kumar Reddy, and, and P. Govindarajulu, "Implementation of Nearest Neighbor Retrieval," 2017.

[7] P. Sunil Kumar Reddy, and, and P. Govindarajulu, "A Keyword Search Based Enhanced Spatially Inverted Index List For The Health Data Retrieval " IJCSNS

International Journal of Computer Science and Network Security ,, vol. 8, 2018.

[8] F. Bobillo, "The role of crisp elements in fuzzy ontologies: the case of fuzzy OWL 2 EL," *IEEE Transactions on Fuzzy Systems*, vol. 24, pp. 1193-1209, 2016.

[9] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," in *Semantic Computing (ICSC), 2015 IEEE International Conference on*, 2015, pp. 244-251.

[10] I. M. Aljawarneh, P. Bellavista, A. Corradi, R. Montanari, L. Foschini, and A. Zanotti, "Efficient spark-based framework for big geospatial data query processing and analysis," in *Computers and Communications (ISCC), 2017 IEEE Symposium on*, 2017, pp. 851-856.

[11] R. Thaker and A. Goel, "Domain Specific Ontology based Query processing System for Urdu Language," *International Journal of Computer Applications*, vol. 121, 2015.

[12] S. Kasbe, P. Shahane, R. Kasar, and M. Navale, "Ontology Based Information Retrieval System Using Multiple Queries For Academic Library," 2017.

[13] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, *et al.*, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, pp. 694-707, 2016.

[14] N. Nodarakis, E. Pitoura, S. Sioutas, A. Tsakalidis, D. Tsoumakos, and G. Tzimas, "kdann+: A rapid aknn classifier for big data," in *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV*, ed: Springer, 2016, pp. 139-168.

[15] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "Semantic matching by non-linear word transportation for information retrieval," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 701-710.

[16] S. W.-t. Yih, M.-W. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," 2015.

[17] G. Singh and V. Jain, "Information retrieval (IR) through Semantic web (SW): an overview," *arXiv preprint arXiv:1403.7162*, 2014.

[18] H. Dong and F. K. Hussain, "Self-adaptive semantic focused crawler for mining services information discovery," *IEEE Transactions on Industrial Informatics*, vol. 10, pp. 1616-1626, 2014.

[19] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain, "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain," *Journal of biomedical informatics*, vol. 48, pp. 38-53, 2014.

[20] S. Akmal, L.-H. Shih, and R. Batres, "Ontology-based similarity for product information retrieval," *Computers in Industry*, vol. 65, pp. 91-107, 2014.

[21] B. Müller, A. Hagelstein, and T. Gübitz, "Life Science Ontologies in Literature Retrieval: A Comparison of Linked Data Sets for Use in Semantic Search on a Heterogeneous Corpus," in *European Knowledge Acquisition Workshop*, 2016, pp. 158-161.

[22] B. Saini, V. Singh, and S. Kumar, "Information retrieval models and searching methodologies: Survey," *Information Retrieval*, vol. 1, p. 20, 2014.

[23] H. Itoh, "RICOH at SemEval-2016 Task 1: IR-based Semantic Textual Similarity Estimation," *Proceedings of SemEval*, pp. 691-695, 2016.

[24] F. M. Anuar, R. Setchi, and Y.-K. Lai, "Semantic retrieval of trademarks based on conceptual similarity," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, pp. 220-233, 2016.

[25] E. G. Caldarola, A. Picariello, and A. M. Rinaldi, "An approach to ontology integration for ontology reuse in knowledge based digital ecosystems," in *Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems*, 2015, pp. 1-8.

[26] S. Bechhofer and N. Matentzoglu, "The OWL API," *COMP60421: Ontology Engineering for the The Semantic Web*, 2014.

[27] "Multi domain Knowledge Database," 2017. Available: https://github.com/harpribot/awesome-information-retrieval

Author's Biography:

P. Sunil Kumar Reddy received his MCA from Bharathiar University in 2004 and M.Phil. Computer Science from Madurai Kamaraj University. Currently, he is Pursuing his Ph.D. in SV University Tirupati. His research areas are Databases and Data Mining.

Dr. P. Govindarajulu is a Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, India. He received his M.Tech. from IIT Madras (Chennai), Ph.D. from IIT Bombay (Mumbai). His area of research: Databases, Data Mining, Image Processing, Intelligent Systems, Software Engineering, and Parallel Computing.