Asian Journal of Information Technology 15 (12): 2037-2047, 2016
ISSN: 1682-3915

Markov-Trivial Similarity Tree-Based Ontology Model for Geospatial Information Retrieval Systems for Grid Computing

[1]K.S. Kannan and [2]R. Saravanan
[1]Department of Computer Science and Engineering, Madurai Institute of Engineering and Technology, Pottapalayam, Sivaganga, Tamil Nadu 630611, India
[2]Department of Computer Science and Engineering,
RVS Educational's Trusts Group of Institutions, Dindigul, Tamil Nadu, India

Abstract: Grid computing allows the sharing of resources from the heterogeneous and distributed locations. An extensive study of the scientific research area in grid computing declares the effective retrieval systems design is the prerequisite for an efficiency improvement. Information Retrieval (IR) systems require the periodic updating due to the rapid increase in a number of multimedia data. Generally, extraction of information in IR system based on the matching of appropriate given queries. The retrieving information from a keyword or matching string based IR is insufficient and limited to critical information by the user. The raising up of geospatial data in semantic makes the IR systems as real-time development. The geospatial data can be used in many scientific fields such as agriculture, land use and climate change. The review of real semantic web describes the problem of poor updating. The large amounts of geospatial data are archived in multiple data center. Caching improves the retrieval performance in the widely distributed environment whereas the performance is poor in the large data set. In the case of spatial data, the geospatial semantic web identifies which parts of geospatial information need to receive semantic specifications in order to achieve interoperability. The duplication and redundant nodes exist in the two or more nodes during the tree construction process caused the irrelevance results in response to the queries. The simultaneous verification of overlap and the weight adjustment in proposed scheme in Globus toolkit environment enhances the relevance results. We use the ranking of trivial similarity measure based ontology structure to improve the efficiency of the data retrieval. Ranking of similarity measures we assures the sorting of the list. To avoid the repetition of the distributed query results from the sorted list we introduce Markov-Trivial-Tree (MTT) based index prediction process to capture the repeated results. The comparative analysis between the proposed MTT-based ontology structures proves that it offers better results than the traditional methods regarding the precision, recall, F-measure, and accuracy for diverse large dimension datasets.

Key words: Geospatial data, Markov-trivial-tree-based index, Information Retrieval (IR), ontology, ranking, semantic web, similarity measure, distributed environment

INTRODUCTION

Aggregation of discrete resources is the major requirement for large-scale real-time applications and such materialized network of them is referred as grid computing. Due to the large availability of resources in diverse locations, the aggregate of information from media (television, radio, internet etc.,) is an attractive process to enhance the knowledge. The Internet is one of the media which is used to get more information related to domain easily. An activity of gathering information from available resources refers Information Retrieval (IR). IR contains various processes such as query (formal statements) from the user to the system, matching of the query with the objects in the resources. With different relevancy, presentation of results with high relevance. IR focus how the related information from the semi-structured data namely, web pages, documents, images and video is located. The presence of information resources turns the research field into the Multimedia Information Retrieval (MIR). The file sharing in MIR uses peer to peer technology in which centralized solutions are dominated. The rise of demand in Knowledge Management (KM) solutions to perform various tasks in an organization such as document/workflow management, web conferencing and decision support system. But the existence of KM solutions for long time extraction is difficult due to the centralized architecture.

Corresponding Author: K.S. Kannan, Department of Computer Science and Engineering, Madurai Institute of Engineering and Technology, Pottapalayam, Sivaganga, 630611 Tamil Nadu, India

The maturity in web service technology increases the availability of functions and geospatial resources. The descriptive information available in geospatial data is structured. The discovery of knowledge and the data processing methods over web services is a major challenge in the research industry which raises the Geoprocessing web. The Geoprocessing web contains protocols, capability of sourcing and processing of real-time geospatial data sources. The evolution of knowledge-based web mining tool extends IR to automated location IR system. The Volunteered Geospatial Information (VGI) availability adopted the provenance based human computation approach. Geospatial services provide the geoprocessing algorithm that handles the small part of overall geoprocessing and large aggregated processing. The follow-up of interface standards requires the achievement of interoperability. The independence of information stored in multiple devices enabled the application independent caching.

The time required to provide the response is maximum in traditional geospatial data processing methods. To minimize the time, web search engine made feasible by utilization of indexing and caching techniques. An alternative query processing methods on the basis of a combination of self-indexed compressed text and caching of posting lists. The query distribution algorithms among search engines which include the query spawning, number of terms per query and length of the query in order to highlight the principal factors. The inclusion of optimal cache replacement policy enhances the data access applications by updating the data entry.

The search engines based on traditional methods such as keywords based search algorithms which are unable to transform the raw data into the knowledge oriented representation which is an inconvenient task in relevant information extraction. The demerits are overwhelmed by the integration of semantic web with the ontology. The interconnection between the IR, Ontology and Semantic Web (SW) assures the new generation of the web. The cooperation among the intelligent agent Software in order to collect the relevant information. The textual data processing considers the estimation of semantic likeness between words. On the basis of likeness principle, similarity measures developed. These measures describe the taxonomical features. The conceptual heterogeneity in the semantic web based model leads to difficult in interoperability offering among devices. The interoperability nature enlarges the data representation called Semantic Information Layer (SIL). The mapping path maintains the link between the SIL and data source, query implementation for data retrieval and user interactions.

An integrated approach that combines the orthogonal IR techniques produced dissimilar results. The combination of IR-based methods such as Vector Space Model (VSM), probabilistic Jensen and Shannon (JS) model and Relational Topic Modelling (RTM) enhances the traceability recovery. Multimedia applications require the ability of fast similarity search in a large-scale data set. The inclusion of high dimensional features into low hamming distance space considered means such as tree-based methods. The retrieval of similar resources in the Multi-Agent Systems (MAS) is a difficult problem. Research works present the three layer architecture and data model reduced the time complexity. Among many tree-based information retrieval methods, hashing algorithms are mostly preferred for less complexity. Hash algorithms are classified as unsupervised methods (Locality Sensitive Hashing (LSH) and supervised methods (Boosting Similarity Sensitive Coding (Boosting-SSC). The problems in traditional algorithms are insufficient information retrieval, the offering of critical information in the keyword-based searching, periodical updating knowledge, the discovery and utilization of data from the multiple data centers in multi-disciplinary research. To overcome these problems, Markov-Trivial-Tree-based ontology structure creation to access the geospatial data is discussed in the study. The novel contributions of proposed MTT-based Information Retrieval system are listed as follows:

- The proposed system employs the trivial similarity measurements (Paramount, angle and string-based similarity computation) to predict the semantic relationship between the query and semantic vectors
- The redundant nodes and duplication entries in the concept-baseds imilarity estimation affect the accuracy. This paper constructs the Markov-Trivial-Tree (MTT) by simultaneous overlap verification
- The sequential weight adjustment for the overlap, greater similarity values effectively maximizes the precision, recall and accuracy values for various documents effectively

Literature review: This study, describes the various related works about the techniques used to improve the performance of Information Retrieval (IR). The raising up of computer technologies increases the processor speed and storage capacity. Sanderson et al described the brief history of the Information Retrieval (IR) (Sanderson and Croft, 2012) system. They presented the text retrieval conference, rank learning methods, web search and query logs. Multimedia Information Retrieval (IR) task to overcome the semantic gap in the query. Benavente *et al.* (2013) combined the textual pre-filtering with the image re-ranking to overcome the semantic gap between the queries. E-lecturing in the educational approaches was more popular in multi-media IR systems. Hence, an amount of lecture video data in World Wide Web (WWW) grows rapidly. Yang and Meinel (2014) presented an approach for automated video indexing and search in large scale lecture dataset. The file sharing in a peer to peer communication dominated by the centralized solutions. Tigelaar *et al.* (2012) opened the door to the large-scale development of real world peer to peer IR systems. The stimulation provided in order to overcome the problems caused by the centralized solutions. Advertise and discovery of shared geospatial data required the catalog services. Yue *et al.* (2011) developed the Cyberinfrastructure perspective of expansion in geospatial data availability. They structured the semantic descriptions for geospatial data. The linking of geospatial into the cloud extended the query languages into an SPARQL in which the whole range of construction involved.

Kyzirakos *et al.* (2012) presented the Strabon, an RDF which utilized the geospatial query languages namely, SPARQL and GeoSPARQL. The identification of location based information from World Wide Web (WWW) became more popular research area in geospatial data mining. Li *et al.* (2012) reported the efforts towards the automated location information retrieval by the development of knowledge based web mining tool refers Cyber Miner. Online web services include a maximum number of geospatial resources and processing functions. Zhao *et al.* (2012) enhanced the utilization performance of geospatial resources with the combination of light weight protocols, crowdsourcing capability refers geoprocessing web. Collaborative efforts made the web search technique as successful in file sharing. The creation of geospatial data by using the collection of Volunteered Geospatial Information (VGI). Celine (2013) focused the adoption of provenance based human computation approach for the VGI consolidate. The advances in Remote sensing technologies increased the earth orbit satellites. Kyzirakos *et al.* (2014) developed the wildlife monitoring system with the combination of ontologies, linked geospatial data. They redeveloped the traditional National Observatory of Athens (NOA) for real-time wildfire monitoring system. The three approaches namely, information-centric networking, cloud computing and open connectivity raised the new network topologies.

Ahlgren *et al.* (2011) integrated the virtualization features with the networking levels. They provided the open connectivity issues on the basis of transport mechanisms. The evolution of indexing and caching techniques made the query response time as feasible. Arroyuelo *et al.* (2012) presented an alternative query processing method on the basis of self-indexed text and caching of post lists. The query distribution in IR systems possessed with the query spawning, number of terms per query and length of the query. The selection of suitable resource to perform the specific job through the query processing consumes more time and cost. Kannan presented the efficient Globus QoS-driven job scheduling approach (Taghavie *et al.*, 2012) for grid environment. The scheduling of jobs within the specified time increases the reliability of the system. The enhancement of query processing is required during the extension of scheduling to Geospatial systems. Taghavi *et al.* (2012) suggested the best search engine capabilities with the analysis of user's queries and investigation of trends. The utilization of bandwidth in the wireless environment was an attractive research in IR systems. Akon *et al.* (2012) proposed the optimal cache replacement policy in which the injected clients updated the new data. IR based on keywords provided the fewer capabilities for the capture of conceptualizations against the user needs.

The interface between the applications with the Information-Centric Networking (ICN). Tagger *et al.* (2013) provided the middleware layer used in the development of more advanced ICN protocols. The problems addressed in keyword based models were conceptual search, literal strings based IR systems. Fernandez *et al.* (2011) extended the classical keyword-based IR model into the ontology based IR model which addressed the challenges of the heterogeneous web environment. They integrated the advantages of traditional keyword-based and ontology-based models. The traditional IR systems based on keyword search methods were suffered by two problems namely, unable to transform the raw data into knowledgeable representation, difficult in the extraction of relevant information from the large collection of documents. These led to the development of semantic web with ontology existence. Jain and Singh (2013) presented the detailed survey report for IR systems, ontology structure and the semantic terms. The cooperation between the participated devices was the investigating parameter in IR systems. Yang and Chang (2011) presented the advanced system to collect the information by estimation of cooperation between the devices participated in the IR system. The semantic likeness between the words in IR systems was an important parameter in textual based language processing and knowledge acquisition. Sanchez *et al.* (2012) surveyed and modeled the ontology-based approaches. They also presented the new ontology based model on the basis of taxonomical features.

The rise of distance learning environment includes the multimedia educational resources as an important role. Yu *et al.* (2012) introduced the new video annotation and browser platform for adopting linked data technology by two tools such as annotation and Sugar tube. Semantic annotation of video resources enabled by annotation. Browsing of semantically linked educational video resources governed by sugar tube. To implement better semantic web environment, evaluation of web search tool was required. Bouramoul *et al.* (2012) proposed the new semantic approach for evaluation of IR systems. They increased the selectivity of search tools and improved the evaluation of web searching tools. The redesign and reengineering were required in system evaluation due to the huge storage of data in heterogeneous distributed environments. Song *et al.* (2012) presented the enlarged data representation model refers Semantic Information Layer (SIL). The data retrieval and interaction with end users were performed by the maintenance of links with mapping path. The recovery of traceability of the links among software artifacts was the difficult process in ontology-based IR models. Gethers *et al.* (2011) exploited the integral approach that combined orthogonal techniques utilized in following models namely, vector space models, probabilistic model and Relational Topic Modelling (RTM). The application of compression-based similarity measures on diverse data types introduced the problems for medium to large datasets. Cerra and Datcu (2012) proposed the similarity measure on the basis of compression in dictionaries. They defined the content based image retrieval system on the basis of Fast Compression Distance (FCD) which reduced the complexity of the system. The manual management of traceability information was error prone and time consuming. Lucia *et al.* (2012) proposed the traceability recovery methods on the basis of similarity between the texts which produced the software artifacts.

The conjugation of artifacts with high textual similarity (Lucia *et al.*, 2012). Multimedia applications required similarity search in a large dataset has a great importance. The retrieval of similar neighbors accomplished by image retrieval within the Hamming distance refers the Spectral Hashing (SH). The Euclidean distance was used to construct the graph based Laplacian in SH. Li *et al.* (2013) optimized the graph Laplacian in SH .The learned graph represents the similarity between the samples for effective binary code learning. Large scale image based applications required the compressed binary image descriptor. Gordo *et al.* (2013) presented the document based image descriptor on the basis of multi-scale run length. The relevance estimation between the query and response documents was absent in traditional concept-based systems. Do *et al.* (2015) proposed the semantic representation of retrieved documents by a domain knowledge extraction and application ranges. The keyword based search methodologies in semantic representation led to the irrelevance results and offered the low precision, recall values. Rao *et al.* (2015) employed the Hadoop Map Reduce to preprocess the large domain data. They also performed the concept-based similarity estimation to reduce the dimensionality effectively. The semantic relationship exists between the words was a necessary key parameter in the sentimental analysis. Agarwal *et al.* (2015) utilized the concept extraction algorithm for semantic feature extraction. The simultaneous important concepts selection and redundant concepts elimination influenced the machine learning a model to enhance the precision, recall and F-measure values. But the similarity estimation extends into the angle and string-based that enhanced the semantic presentation. This paper employs the trivial similarity measure to find the semantic relationship between the query and semantic vectors. Also, the duplication entries in the two or more concept trees overcome by the sequential weight adjustments during the tree construction that enhance the precision, recall, accuracy values effectively.

MATERIALS AND METHODS

This study, presents the detailed description about proposed Markov-Trivial-Tree (MTT)-based similarity measurement in real-time query processing of geospatial model. The flow chart of proposed system is shown in Fig. 1. The proposed model contains various processes as follows:

- Information retrieval model
- Semantic Vector (SV) creation
- Similarity Measurement (SM)
- Markov-trivial-tree construction

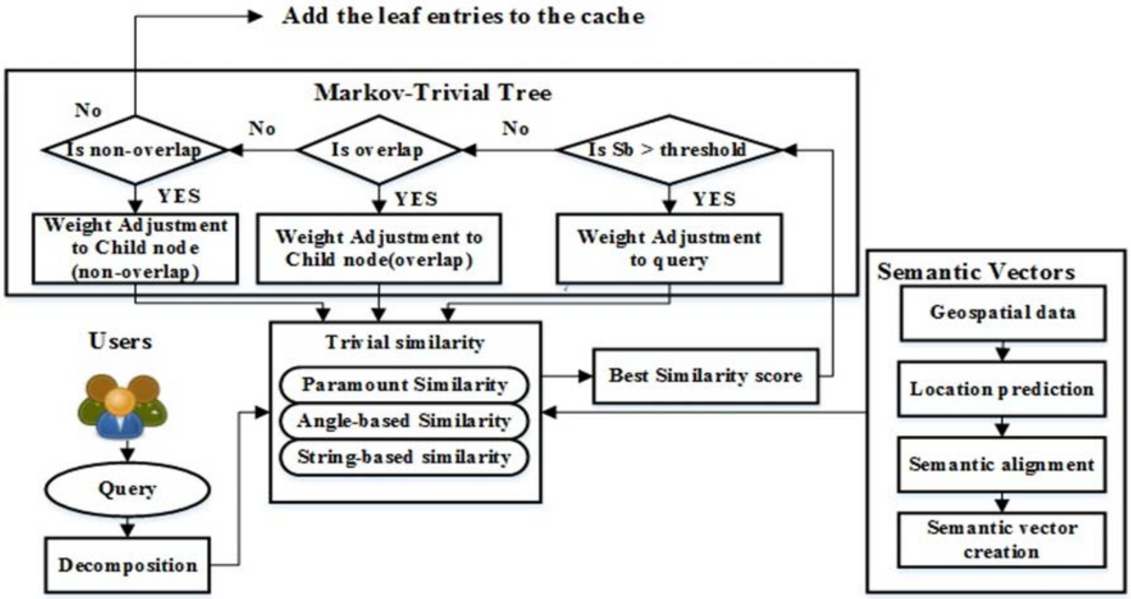

Fig. 1: Flow diagram of proposed method

Information retrieval model: The selection of proper model is to formalize the retrieving of suitable information from geospatial data. In general there are three models namely, Boolean, vector space and probabilistic model exist in the Information Retrieval (IR) approaches. The Boolean model is based on the keyword manipulation. Document (D) and Query (Q) are the two inputs for Boolean model. The pairing mode of Boolean returns only the documents related to queries. A probabilistic model is based on the mathematical model to get a unified representation extracted concepts.

But the probabilistic and Boolean-based models are based on the assumption such that independence of extracted variables which is not verified during the implementation leads to the similarity of inaccuracy. Hence, the vector space model is used in this paper in order to increase the accuracy of similarity measurement. In vector space model, N-dimensional vector space created which contains documents and queries as represented by:

$$D=(d_1,d_2,\dots d_m) \tag{1}$$

$$Q=(q_1,q_2,\dots q_w) \tag{2}$$

The reasons behind the selection of vector space model is described as follows:

- Consistency in the representation of documents and queries
- Similarity function ordering, and easy weight function adjustment to improve the search results

Semantic Vector (SV) creation: The Semantic Vector (SV) creation is to project the query to ontology concepts by using the relations such as synonymy and hyponymy. The semantic vector extraction is initiated from the ontology extraction. The main structure or semantic vector extracted from the knowledge database termed as “raw ontology”. The elaborate converts this raw term into reference ontology or manual adjustment. Then, the location data from the database converted into ontology instances by population process. The query to the ontology instances creates the vectors. The recovery of semantic vectors based on the weighing techniques. The Term Frequency (TF) and Inverse Document Frequency (IDF) are the two factors for weighing techniques.

Term Frequency (TF): The measure of the frequency of the word in geospatial data termed Term frequency (TF).

Inverse Document Frequency (IDF): The measure of the importance of a location term in total weight refers to Inverse Document Frequency (IDF). The TF and IDF measure provide the good approximations for location importance in geospatial data. Let M be the un corpus (structured set of texts) documents, then ensemble of un corpus defined:

$$M=[D]=(d_1,d_2,\dots d_m) \tag{3}$$

The multiplicity (cardinality) of ensemble representation is as follows:

$$\text{Card}(M)=\sum_{i=1}^{m}\text{card}(d_i) \tag{4}$$

The potential value of each document ∈ d_i, I→[1, m] is denoted as:

$$\text{Pot (M)} = \sum_{d\in m} m(\omega) \quad (5)$$

Then, the TF and IDF are calculated as:

$$T_f = \frac{\text{Pot (M)}}{\text{Card (M)}} \quad (6)$$

$$\text{IDF} = \log \frac{n}{|\{d_i : \omega \in d_i\}|} \quad (7)$$

On the basis of weighing function obtained from Eq. 6 and 7 Semantic Vector (SV) module calculates the documents and query vectors. The weighted coefficients of document and query (d_i, q_{ij}) influences the semantic vector (S_k) creation in following steps:

- The coefficient d_{ij} of document defines the weight of term i in document j described as in Eq. 6
- The coefficient of query vector defines the weight of the ith term in all documents by in Eq. 7

The generated semantic vector is then passed to the cache module and check whether the semantic vector is presented in the cache. If yes, then the vectors pass to the ranking module which calculates the score based on the query. The generated semantic vectors are not in the cache module then it follows the similarity measurement process. The matching of query weighting and document weighting factors investigated in similarity measurement module.

Similarity Measurement (SM): The relevancy of documents with the query is investigated by calculating a score from trivial similarity measures as follows: paramount, string-based and angle-based. The best similarity score is computed from the trivial similarity comparison. The algorithm a for similarity measurement is listed as follows.

Alogrithm A: Trivial similarity algorithm

```
Input: Document vectors ($d_i$), semantic vectors ($S_k$)
Output: Best similarity score ($S_b$)
Initialize the set as E = {$d_i$, $s_i$}
//Paramount similarity
For each vector
If ($d_i$! = $s_i$)
$X_i = \lambda* (d_i*s_i)-d_i*s_i$
 $t_{i_1} = d_i^2$
 $t_{i_2} = s_i^2$
Similarity score $S_i = X_i / (\lambda*t_{i_2})*(\lambda*t_{i_2})$
End
End for
Similarity score $S_p$ = mean [$S_i$]
//String based similarity
Initialize into array
For I = n-1…0 then
    For j = 0…m
        Init= str1.string length (I)–str 2. String length (j)
     End For
End For
 Initialize result array → 0
  For I = 0…, n
       Result [I] [0] = result [I+1] [0]+init [I] [0]
End For
For j = 0…, m
       Result [n-1] [j] = result [n-1 ][j-1]+init [n-1] [j]
End For
For I = n-2…. 0 then
       For j = n-2…..0 then
             Value 1 = result [i] [j-1]+init [i] [j]
              Value 2 = result [I+1] [j]+init [I] [j]
              Value 3 = result [i+1] [j-1]+( inti [i] [j] * 2)
               Resul t = value1<value2?value1:value 2
                 Resul t = result<value3? result:value3
                 Resul t = Similarity Score SS
           End For
 End For
Return Similarity Score ($S_s$)
 //Angle-based similarity
Initialize query: Vector → {left word, right word}
temp → null
    for left word: left
     temp = left word count
             if (temp == null) then
                Left word count = 1
               else then
                Left word count = 1+temp
              end else
              end if
             end for
              for right word: right
                 Temp = right word count
                 if (temp == null) then
                 Right word count = 1
              else
                  Right word count = 1+temp
               end if
             end for
             Initialize left vector, right vector, index → 0
             temp count → 0
            for unique word: a unique set
                Temp count = left word count
                   Left vector [index] = temp count == null → 0: temp
                    count
                    Temp count = right word count
                Right vector [index] = temp count == null → 0:
                 temp count
                 Index ++;
            end for
Return Similarity Score ($S_A$)
( If ($S_P$>$S_A$) then
            If ($S_P$>$S_A$) then
                        Best Similarity Score ($S_b$ ) = $S_p$
                         Else
                        Best Similarity Score ( $S_b$) = $S_p$
                                End If
Else then
Best Similarity Score ($S_b$ ) = $S_A$
 End
 Return Best Similarity Score ($S_b$)
```

The document and semantic vectors are given as the inputs to the trivial similarity algorithm. Initially for each vector, the paramount similarity is computed that provides the relevancies between the document and semantic vectors. Then, the direction based estimation (angle-based similarity measurement) predicts the numerical relevancies between the query and document vectors. The string relevancies are also estimated through the string-based similarity measurement. Finally, the comparison of three estimated similarity values identifies the best similarity value S_b. The score is assigned to the search engine by using the sorting of best similarity scores for each document/query vectors. The computed scores are given as an input vectors to the tree construction to improve the retrieval performance.

Markov-trivial-tree construction: The synonyms, missing words and irrelevant expansion terms limit the precision, recall and accuracy of information retrieval. Query expansion technology offers the necessary solution for performance metrics enhancement. The words selection and weight distribution are the core issues in the query expansion process.

A network has the strong semantic relationship among the interconnected nodes refers Markov network. During the retrieval process, the original query input (node) and an associated semantic relationship between the closely related query (node) are extracted through the trivial similarity measurement. The algorithm b to create the MTT is listed as follows:

Alogrithm B: Markov-trivial-tree:

```
Initialize p = 1, user threshold value ($\in$) and pointer q
Search the item ($d_i$) related to the query ($d_i$) and assign it to the semantic vectors set ($S_k$)
Initialize $q_m$ be the root node
Let $t_j$ be the child node and add to the list $L_P$
If $(Lp \neq null)$
q points the first child of $L_p$
Else
Go to step 9
Remove the node $t_j$ from the list $L_p$
Find the item $t_k$ connected with $t_j$
Find the trivial similarity score $S_b$ between the $t_k$ and tj
If $(S_b > \varepsilon)$
Adjust the weight for query by using Eq. 8 and repeat the similarity estimation
Else
Check the overlap or non-overlap
If(overlap)
Adjust the weight for tree node by using Eq. 9 and repeat the similarity measurement process
Else
Adjust the weight for tree node by using Eq. 10) and repeat the similarity measurement process
End
If $(S_b > \sigma)$
Add the $t_k$ to list $L_p$+1
Change the pointer q to the next node
Else
Change the pointer q to the next node
End
If $(q \neq L_{p-end})$
Go to step 9
Else
Change the p to p+1
 Repeat the similarity computation and weight adjustment for new list ($L_p$)
end
```

The query expansion terms deviate from the core query terms that leads to irrelevant results in the multi-dimensional query. Hence, redundant removal and semantic weight adjustment make the expansion terms as closer to core queries that improve the performance. The three consecutive processes of MTT creation are as follows:

- Weight adjustment
- Redundant node pruning
- Non-core terms weight reduction

Weight adjustment: The query terms have the high correlation between them called as core query that plays a major role in search task and corresponding weight is increased. The query weight after the adjustment (W_b) is mathematically expressed with best similarity score (S_b) and the prior query weight (q_q) as follows:

$$W_b = (1 + S_b) * \text{weight}(q_m) \tag{8}$$

The changes in query weight point the new data item during the retrieval process. The trivial similarity of the query with new weight and new data item are re-estimated to improve the precision and recall of relevant results. The same query node in two tree structures leads to the noise and complexity.

Redundant node pruning: The nodes exist in one or more tree structure are called redundant nodes. If the trivial similarity measure is zero, then there is no semantic relationship between them. The redundant node removal effectively minimizes the noise and complexity. The simultaneous keeping of overlap node in one query and removal of redundant node in other query influence the weight adjustment as:

$$\begin{gathered} W_{bnew}(t_j) = (1 + S_b(q_m, q_m) \times W_{bold}(t_j) \times S_b(q_m, t_j)) \\ t_j \in \{\text{tree}(q_m)^{\wedge}\,\text{tree}(q_n)\}^{\wedge}\, S_b(q_m, q_n) > \varepsilon \end{gathered} \tag{9}$$

The weight update in the child node influenced on the root node. Hence, the semantic correlation is again executed using the trivial similarity algorithm. The sequential weight update and the similarity estimation removes the redundant nodes that limit the noise.

Non-core weight reduction: The expansion terms associated with the initial query may not relevant to the entire query topic that leads to less precision and high topic drift. Hence, the weight is reduced to avoid the semantic topic drift and irrelevant results:

$$W_{bnew}(t_j)=(1\text{-}S_b(q_m,q_n)\times W_{bold}(t_j)\times S_b(q_m,t_j)) \\ t_j\in\{\text{tree}(q_m)^{\wedge}\text{tree}(q_n)\}^{\wedge}S_b(q_m,q_n)>\varepsilon \quad (10)$$

The revised weight computation for each similar score of core and initial query vectors effectively removes the redundant nodes from the two trees. If the similarity values are still zero after three processes, then we identify that there is no semantic relationship is exists between the query terms and does not require any weight adjustment.

RESULTS AND DISCUSSION

Performance analysis: This section presents the comparison of proposed MTT-based ontology information retrieval system with the existing models namely, semantic processing-based information retrieval, Key-Based Information Retrieval Systems(KIRS), Secure Semantic Symmetric encryption and Ranking-based systems (SBIRS) and Semantic Parser (SP) methods (without Common Sense with Common Sense (CS) and m-RMR feature selection) regarding the precision, recall, accuracy and F-measure. The MTT-based IR proposed in this study validates its performance on the various review datasets such as Movies, books, DVD and electronics with the size of 10000. The evaluation of different metrics with different document sizes variations.

Precision: The document sizes vary in the ranges of 1000, 2000, 5000 and 10000 and test the performance with 100 queries. For the different ranges, the precision values for existing graph-based methods, semantic methods and proposed MTT-based IR are measured.

Figure 2 describes the comparative analysis of proposed MTT-based IR with the graph-based and semantic-based methods regarding the precision values. The semantic-based IR offered the better precision values compared to graph-based methods. But the same entries existence on two or more nodes in the concept-trees caused the duplicate entries were more. The semantic based IR has the precision values of 90.51 and 89.85 for minimum and maximum documents respectively. But the trivial measures and the weight adjustment in proposed MTT-based IR offers 91.08 and 91.51% values which are 0.57 and 1.66% more compared to existing IR method.

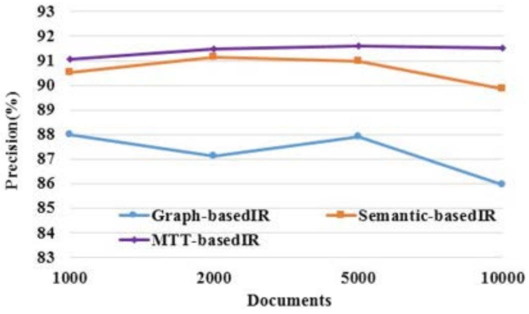

Fig. 2: Precision vs large size variation

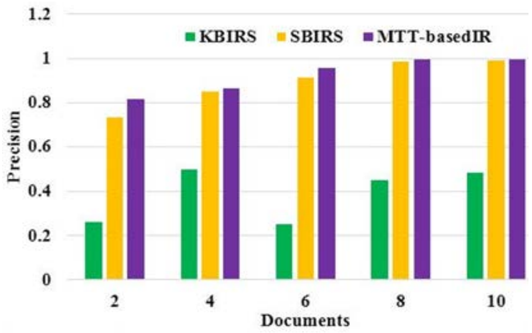

Fig. 3: Precision vs small size variations

Figure 3 presents the comparison of proposed MTT-based IR with the KBIRS and SBIRS on precision values. The duplication removal through the MTT and the effective similarity measurements improves the precision values by 10.88 and 0.8 % for minimum (2) and maximum (10) documents, respectively

Recall: The document sizes vary in the ranges of 1000, 2000, 5000 and 10000 and test the performance with 100 queries. For the different ranges, the recall values for existing graph-based methods, semantic methods and proposed MTT-based IR are measured. Figure 4 describes the comparative analysis of proposed MTT-based IR with the graph-based and semantic-based methods regarding the recall values. The semantic-based IR offered the better recall values compared to graph-based methods. But the same entries existence on two or more nodes in the concept-trees caused the duplicate entries were more.

The semantic-based IR has the recall values of 89.25 and 90.45 for minimum and maximum documents, respectively. But the trivial measures and the weight adjustment in proposed MTT-based IR offers 89.6 and 90.61% values which are 0.35 and 0.15% more compared to existing IR method.

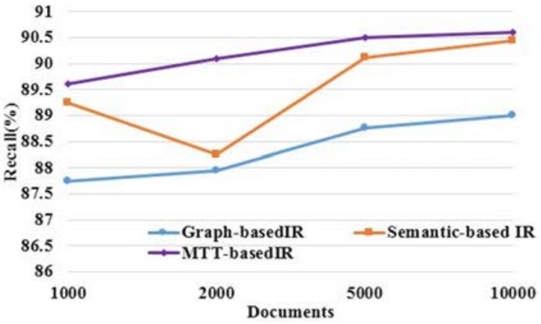

Fig. 4: Recall vs large size variations

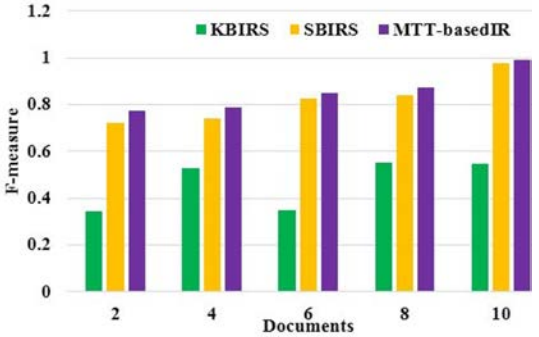

Fig. 6: F-measure vs small size variations

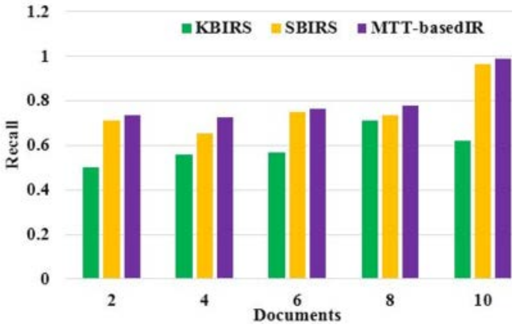

Fig. 5: Recall vs small size variations

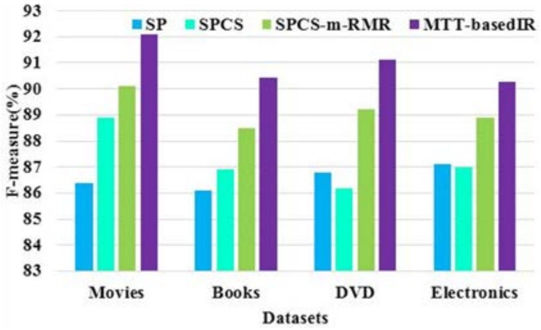

Fig. 7: F-measures vs datasets

Figure 5 presents the comparison of proposed MTT-based IR with the KBIRS and SBIRS on recall values. The duplication removal through the MTT and the effective similarity measurements improves the precision values by 3.65 and 2.49% for minimum (2) and maximum (10) documents, respectively.

F-measure: The increase in precision and recall values will directly increase the F-measure values. The comparison of proposed MTT-based IR with the KBIRS, and SBIRS regarding F-measures for minimum documents size variations conveys the performance.

Figure 6 shows the comparative analysis of proposed MTT-based IR with the existing KBIRS and SBIRS regarding the F-measure. The SBIRS provided better recall values than the KBIRS. But the trivial similarity measures enhanced the F-measure values in MTT-based IR. The F-measure values for SBIRS are 0.723 and 0.977 for minimum (2) and maximum (10) documents respectively. The proposed MTT-based IR offers the F-measures are 0.774 and 0.993, respectively. The performance of proposed MTT-based IR is enhanced by 7.09 and 1.65% for minimum and maximum documents respectively compared to SBIRS.

Figure 7 discusses the comparative analysis of proposed MTT-based IR with the Semantic-Parser (SP) without common sense, Semantic-Parser with Common Sense (SPCS) and SP with the m-RMR feature selection (SPCS) regarding the F-measures variations with various datasets utilization. The SPCS-m-RMR provided the 90.1, 88.5, 89.2, 88.9% for movies, books, DVD and Electronics documents respectively which was more than the SP and SPCS. But, the redundancy removal in proposed MTT-based IR offers the F-measures of 92.1, 90.45, 91.11 and 90.26%, respectively. The MTT-based IR enhanced the F-measures by 2, 1.95, 1.91 and 1.36%, respectively

Accuracy: The semantic relationship identification between the query and related document vectors and the weight adjustment during the tree construction provides the both redundancy/duplication removal that enhances the accuracy.

Figure 8 shows the comparative analysis of proposed MTT-based IR with the existing KBIRS and SBIRS regarding the accuracy. The SBIRS provided better accuracy than the KBIRS. But, the trivial similarity measures enhanced the accuracies in MTT-based IR. The F-measure values for SBIRS are 0.835 and 0.962 for minimum (2) and maximum (10) documents respectively. The proposed MTT-based IR offers the F-measures are 0.862 and 0.982, respectively. The performance of

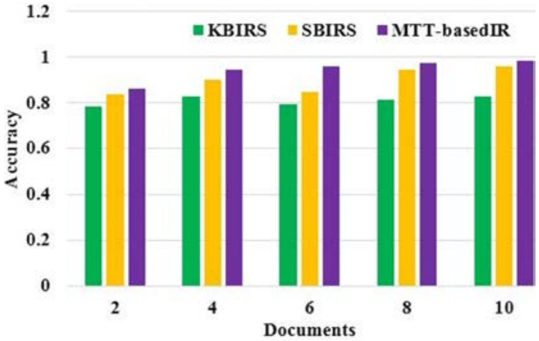

Fig. 8: Accuracy analysis

proposed MTT-based IR is enhanced by 3.23 and 2.08% for minimum and maximum documents, respectively compared to SBIRS.

CONCLUSION

This study, reviewed the problems in design methods of geospatial Information Retrieval (IR) systems for Grid computing such as the retrieving information from a keyword or matching string based IR are insufficient due to the critical information by the user, poor updating and poor performance for large data set such as geospatial data set. The achievement of interoperability between the devices also poor in the identification of which part of the spatial information needed semantic specifications. The duplication and redundant nodes exist in the two or more nodes during the tree construction process caused the irrelevance results in response to the queries. This study proposed Markov-Trivial-Tree (MTT)-based ontology model for geospatial query processing in order to solve the above issues. The simultaneous verification of overlap and the weight adjustment in proposed scheme in Globus toolkit environment enhanced the relevance results. We used the ranking of trivial similarity measure based ontology structure for improvement of IR efficiency and sorting. The proposed MTT-based index model effectively avoided the repetition of the distributed query results from the sorted list. Moreover, the comparative analysis between the proposed MTT-based ontology structures and the traditional methods regarding the precision, recall, accuracy, F-measure proved the effectiveness in the retrieval of information to support the an effective grid computing.

REFERENCES

Agarwal, B., S. Poria, N. Mittal, A. Gelbukh and A. Hussain, 2015. Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach. Cognit. Comput., 7: 487-499.

Ahlgren, B., P.A. Aranda, P. Chemouil, S. Oueslati and L.M. Correia et al., 2011. Content, connectivity, and cloud: Ingredients for the network of the future. IEEE. Commun. Mag., 49: 62-70.

Akon, M., M.T. Islam, X.S. Shen and A. Singh, 2012. A bandwidth and effective hit optimal cache scheme for wireless data access networks with client injected updates. Comput. Networks, 56: 2080-2095.

Arroyuelo, D., V.G. Costa, S. Gonzalez, M. Marin and M. Oyarzun, 2012. Distributed search based on self-indexed compressed text. Inf. Process. Manage., 48: 819-827.

Benavent, X., S.A. Garcia, R. Granados, J. Benavent and D.E. Ves, 2013. Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection. IEEE. Trans. Multimedia, 15: 2009-2021.

Bouramoul, A., M.K. Kholladi and B.L. Doan, 2012. An ontology-based approach for semantics ranking of the web search engines results. Proceedings of the International Conference on Multimedia Computing and Systems, May 10-12, 2012, IEEE, Tangier, Morocco, ISBN: 978-1-4673-1518-0, pp: 797-802.

Celino, I., 2013. Human computation VGI provenance: Semantic web-based representation and publishing. IEEE. Transac. Geosci. Remote Sens., 51: 5137-5144.

Cerra, D. and M. Datcu, 2012. A fast compression-based similarity measure with applications to content-based image retrieval. J. Visual Commun. Image Represent., 23: 293-302.

Do, N.V., T.A.P. Nguyen, H.K. Chau and T.T.T. Huynh, 2015. Improved semantic representation and search techniques in a document retrieval system design. J. Adv. Inf. Technol., 6: 146-150.

Fernandez, M., I. Cantador, V. Lopez, D. Vallet, P. Castells and E. Motta, 2011. Semantically enhanced information retrieval: An ontology-based approach. Web Semantics Sci. Services Agents World Wide Web, 9: 434-452.

Gethers, M., R. Oliveto, D. Poshyvanyk and A.D. Lucia, 2011. On integrating orthogonal information retrieval methods to improve traceability recovery. Proceedings of the 27th IEEE International Conference on Software Maintenance, September 25-30, 2011, IEEE, Williamsburg, Virginia, ISBN: 978-1-4577-0662-2, pp: 133-142.

Gordo, A., F. Perronnin and E. Valveny, 2013. Large-scale document image retrieval and classification with runlength histograms and binary embeddings. Pattern Recognit., 46: 1898-1905.

Jain, V. and M. Singh, 2013. Ontology based information retrieval in semantic web: A survey. Int. J. Inf. Technol. Comput. Sci., 5: 62-69.

Kyzirakos, K., M. Karpathiotakis and M. Koubarakis, 2012. Strabon: A Semantic Geospatial DBMS. In: International Semantic Web Conference. Mauroux, P.C., J. Heflin, E. Sirin, T. Tudorache and J. Euzenat et al. (Eds.). Springer Berlin Heidelberg, Heidelberg, Germany, ISBN: 978-3-642-35176-1, pp: 295-311.

Kyzirakos, K., M. Karpathiotakis, G. Garbis, C. Nikolaou and K. Bereta et al., 2014. Wildfire monitoring using satellite images, ontologies and linked geospatial data. Web Semant. Sci. Serv. Agents World Wide Web, 24: 18-26.

Li, P., M. Wang, J. Cheng, C. Xu and H. Lu, 2013. Spectral hashing with semantically consistent graph for image indexing. IEEE. Transac. Multimedia, 15: 141-152.

Li, W., M.F. Goodchild, R.L. Church and B. Zhou, 2012. Geospatial data mining on the web: Discovering locations of emergency service facilities. Proceedings of the International Conference on Advanced Data Mining and Applications, December 15-18, 2012, IEEE, USA., ISBN: 978-3-642-35527-1, pp: 552-563.

Lucia, A.D., A. Marcus, R. Oliveto and D. Poshyvanyk, 2012. Information Retrieval Methods for Automated Traceability Recovery. In: Software and Systems Traceability. Huang, J.C., G. Orlena and A. Zisman (Eds.). Springer, Berlin, Germany, ISBN: 978-1-4471-2239-5, pp: 71-98.

Rao, P.S., M.K. Prasad and K.T. Reddy, 2015. An efficient semantic ranked keyword search of big data using map reduce. Int. J. Database Theory Appl., 8: 47-56.

Sanchez, D., M. Batet, D. Isern and A. Valls, 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Syst. Appl., 39: 7718-7728.

Sanderson, M. and W.B. Croft, 2012. The history of information retrieval research. Proc. IEEE., 100: 1444-1451.

Song, F., G. Zacharewicz and D. Chen, 2012. An ontology-driven framework towards building enterprise semantic information layer. Adv. Eng. Inform., 27: 38-50.

Tagger, B., D. Trossen, A. Kostopoulos, S. Porter and G. Parisis, 2013. Realising an application environment for information-centric networking. Comput. Networks, 57: 3249-3266.

Taghavi, M., A. Patel, N. Schmidt, C. Wills and Y. Tew, 2012. An analysis of web proxy logs with query distribution pattern approach for search engines. Comput. Stand. Interfaces, 34: 162-170.

Tigelaar, A.S., D. Hiemstra and D. Trieschnigg, 2012. Peer-to-peer information retrieval: An overview. ACM. Trans. Inf. Syst. TOIS., Vol. 30,

Yang, H. and C. Meinel, 2014. Content based lecture video retrieval using speech and video text information. IEEE. Trans. Learn. Technol., 7: 142-154.

Yang, S.Y. and Y.Y. Chang, 2011. An active and intelligent network management system with ontology-based and multi-agent techniques. Expert Syst. Appl., 38: 10320-10342.

Yu, H.Q., C. Pedrinaci, S. Dietze and J. Domingue, 2012. Using linked data to annotate and search educational video resources for supporting distance learning. IEEE. Transac. Learn. Technol., 5: 130-142.

Yue, P., J. Gong, L. Di, L. He and Y. Wei, 2011. Integrating semantic web technologies and geospatial catalog services for geospatial information discovery and processing in cyberinfrastructure. GeoInformatica, 15: 273-303.

Zhao, P., T. Foerster and P. Yue, 2012. The geoprocessing web. Comput. Geosci., 47: 3-12.