International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 39625-39630

A Framework for Semi Semantic Ontology Based Document Clustering in Geospatial Domain

A. SASI KUMAR

Assistant Professor School of Computing Sciences Vels University Pallavaram Chennai-600 117 Tamil Nadu India E-mail: askmca@yahoo.com

S.P. TAMIZHSELVI

Teaching Fellow Faculty of Information and Communication Engineering Department of CSE Anna University Chennai-600 025 Tamil Nadu India E-mail: tamizh8306@gmail.com

G. SUSEENDRAN

Assistant Professor School of Computing Sciences Vels University Pallavaram Chennai-600 117 Tamil Nadu India E-mail: suseendar_1234@yahoo.co.in

Abstract
Geospatial information (GI) is the information describing the location and names of features on the earth surface.. In a distributed environment, Semantic Web Service makes solutions for spatial annotation, discovery, extraction, composition and invocation. The main objective of this work is to develop the algorithm using ontology concepts for semantically cluster the dam documents. The proposed framework designs the semi semantic ontology based document clustering algorithm. This approach semantically clusters XML documents. The proposed algorithm is evaluated and validated using F-Measure. The result has been evaluated by using various clustering algorithm.

Keywords: Geospatial Information system (GIS), Ontology, Semi Semantic Annotation, Semantic Web, Clustering.

1. INTRODUCTION

1.1 Geographic Information System

Geographic information system [1] is used to analyze the process and maintain all types of spatial data. It is an emerging field for research work. GIS has powerful tool to organize complex spatial environment. Translation of implicit geographic data into an explicit map is defined as GIS [2]. To analyze the real world problems GIS is the new technology which integrates the geographical features with tabular data. The result of the GIS analysis may be categorized into two type's namely derivative results and interpolated results.

In GIS, the particular location is identified by using the two co-ordinates such as x and y. (i.e., latitude and longitude). These co-ordinates are stored in a vector format. Different types of GIS models are also applicable for describing the well defined data structures. GIS deals with geospatial data such as dams, geographic images, satellites and telecommunications [2]. There are various advantages are available in the concept of GIS namely, it accepts the input as digitized maps, rescaling the geographic data is also possible, GIS has its own RDBMS, this GIS retrieves the solution for simple and complex queries to determine the pattern at any given point, GIS provides the results visually either in the form of maps or graphs.

1.2 Ontology

The sequence data are maintained in heterogeneous formats by public databases for different functionalities with their own naming conventions and structures. The various data formats used in sequence information display are: Genpept, FASTA, XML, ASN.1, and Plaintext. The data can be downloaded in any format for processing the information.

Ontology is a Greek term which gives the philosophical meaning as a description of what exists [15]. Ontology is a vocabulary of entities, classes, properties, functions and their relationships. The entities and relationships among the vocabulary is described by the ontology. Ontologies for computing applications are schemas for metadata. Ontologies are meant to provide an understanding of the static domain knowledge that facilitates knowledge sharing and reuse. The four different types of ontologies are:

i) *Domain ontologies*: Represents target domain related information for various domain sources, of engineering, medicine and more [4].
ii) *Generic or Common Sense ontologies*: Generic ontology captures general knowledge about time, space, events and more.
iii) *Method ontologies*: Method ontology describes the type of ontology in specific task, used in diagnosis of medical and clinical domains.
iv) *Metadata ontologies*: Meta ontology describes the content of on-line information sources.

In the existing approaches in order to access the data, one needs to go directly to the actual database, generated queries, most likely, using some standard query language such as SQL, and then submit it to get the required data. This approach is not convenient due to two reasons:

- First, the user is not able to access the database without proper access privileges the database is useless.
- Secondly, the user must have knowledge about a query language supported by that specific database.

By using ontologies, the system finds a solution to this problem. One approach is to provide features that support generating queries on the ontology. Ontology tools can aid in the task of mapping and merging information from different domains, obtaining finally a semantically integrated model.

1.2.1 Ontologies for Data Integration

Ontologies have been extensively used in data integration systems [22] because they provide an explicit and machine-understandable conceptualization of a domain. They have been used in one of the three following ways:

Single ontology approach:
In this approach source schemas are directly related to a shared global ontology that provides a uniform interface to the user. However, this approach requires that all sources should have nearly the same view on a domain, with the same level of granularity. A typical example of a system using this approach is SIMS [21].

Multiple ontology approach:
Each data source is described by its own local ontology separately. Instead of using a common ontology, local ontologies are mapped to each other. For this purpose, additional representation formalism is necessary for defining the inter ontology mappings.

Hybrid ontology approach:
Hybrid ontology approach is combination of single and multiple ontology approaches. First, a local ontology is built for each source schema, which, however, is not mapped to other local ontologies, but to a global shared ontology. New sources can be easily added with no need for modifying existing mappings. Our layered framework is an example of this approach. The single and hybrid approaches are appropriate for building central data integration systems, the former being more appropriate for Global as View (GaV) systems and the latter for Local as View (LaV) systems [9]. Uses of ontologies in data integration are:

(i) *Metadata Representation*: Metadata (i.e., source schemas) in each data source can be explicitly represented by a local ontology, using a single language.

(ii) *Global Conceptualization*: The global ontology provides a conceptual view over the schematically-heterogeneous source schemas.

(iii) *Support for High-level Queries*: Given a high-level view of the sources, as provided by a global ontology, the user can formulate a query without specific knowledge of the different data sources. The query is then rewritten into queries over the sources, based on the semantic mappings between the global and local ontologies.

(iv) *Declarative Mediation*: Query processing in a hybrid peer-to-peer system uses the global ontology as a declarative mediator for query rewriting between peers.

(v) *Mapping Support:* A thesaurus, formalized in terms of ontology, can be used for the mapping process to facilitate its automation.

1.3 Geospatial Ontology

Ontologies give the specifications of a shared conceptualization and apply to provide the semantic for geospatial data sources explicitly [11]. Ontology enables automated semantic match making decisions. Natural-resource decision-support tasks within the domain support geospatial web services such as semi automatic annotation, discovery and composition.

2. RELATED CONCEPTS

2.1 Geospatial Semantic Web

The Semantic Web for geographic information, called Geospatial Semantic Web by Egenhofer [12], is a way to process requests involving different kinds of geospatial information. This requires the capture and analysis of such information, grouping data according to criteria that extrapolate their syntactic context. According to the author, this process requires the development of multiple spatial and domain ontologies, their representation in a way that computers can implement and process the spatial ontologies and processing of queries considering these ontologies and the evaluation of results based on the required semantics. All of this leads to the search fora geospatial information retrieval framework that relies on ontologies, allowing users to retrieve desired data, based on their semantics.
In spite of extensive research, the Semantic Web is far from becoming a reality. Although several standards have been developed and adopted, there are too many views, interests and needs of people that publish and share content in the Web. Consensual vocabularies and ontologies are hard to establish and maintain. So far, most retrieval engines are restricted to text, and other kinds of media pose countless challenges to the effective implantation of the Semantic Web [10].

2.2 Semantic Annotations

“To annotate" means to add comments, to comment. An annotation is used to describe a textual content by means of formal concepts (e.g., using entities in an ontology). An annotation means, a set of metadata that provide a reference to each annotated entity by its unique Web identifier, like a URI. In other words, annotations formally identify resources (in the text, called digital content) through the use of concepts and the relationships among them, and can be processed by a machine. A way to promote interoperability is to use the entities of a domain ontology as those concepts. For example, an annotation may relate the word *orange* that occurs in a text to an ontology that identifies this word as an abstract concept *fruit* (as opposed to *color*).
The annotation process should be as automatic as possible, since a manual process can be slow and subject to errors. This remains a challenge that has been addressed by a number of

research projects. However, most of the proposed mechanisms consider annotations only of textual content, not taking into account other kinds of content. In the geospatial domain, there is also other information to consider, e.g. satellite images, maps, graphs, data from sensors. There is a scarcity of mechanisms to annotate these data, motivating our research.
This work combines characteristics of metadata and annotations into semi semantic annotations: metadata fields are filled with ontology descriptions and terms are used to describe these fields. Based on this, and following, we refer [8] the semantic annotations as follows.

Annotation Units.
An *annotation unit* a is a triple<s,m,v>, where s is the subject being described, m is the label of a metadata field and v is its value or description. Annotation. An *annotation* A is a set of one or more annotation units.

Semantic Annotation Units.
A *semantic annotation unit* is a is a triple <s,m,o>, where s is the subject being described, m is the label of a metadata field and o is a term from a domain ontology.

Semantic Annotation.
A *semantic annotation* SA is a set of one or more semantic annotation units.

Annotation Schema and Content.
An annotation (or semantic annotation) has a schema and content, or instances. The schema is it structure, given by its metadata fields; the content corresponds to the values of these fields.

In fact, annotation units describe data using natural language; semantic annotations use ontology classes and can be processed by a machine. Natural language content of annotations is also part of an ontology: we use instances (individuals) of the ontology classes.

2.3 Document Clustering

Clustering is an unsupervised approach used for exploratory data analysis to classify data into groups. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data and to subdivide a set of items in such a way that similar items fall into the same cluster, whereas dissimilar items fall in different clusters. Document clustering plays an important role in information retrieval and text mining for extracting useful information from huge amount of documents[5]. Document clustering is used in information retrieval for grouping related documents represented in unstructured and semi-structured formats. Clustering of sequence data is important to understand the patterns and structures.

3. THE PROPOSED FRAMEWORK

There is various inference tools are available for annotating, discovering, composing and invocating the geo-spatial web services. To provide dynamic service composition and semantic web service the semi semantic ontology based document clustering framework is proposed. The proposed architecture is shown in Figure 1.

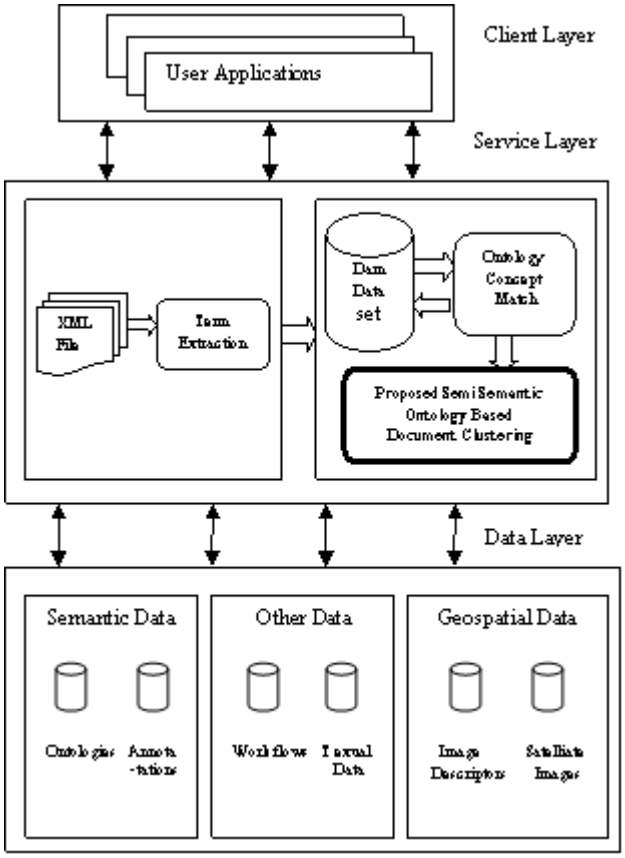

Figure 1: Architecture of Semi Semantic Ontology Based Document Clustering in Geospatial (Dams) Domain

The client layer is responsible for processing a user request, then it to be processed by the middle layer and presenting the returned result. The service (middle) layer provides services such as: textual and geospatial data management and ontology management. Ontology management is ontology Web service responsible for handling ontologies. It provides a wide range of operations to store, manage, search, rank, analyze and integrate ontologies.
The data layer contains geospatial data include satellite images, region boundaries, crop information. Ontologies provide semantics. Other data include information on properties, products and so on.
There are various components are defined in the geospatial ontology. They are semantic annotation, semantic discovery and execution components. Ontologies are mainly used to access the information and assign the data.
Analyze the semi-structured data descriptions to provide the geo-spatial web services, which will generate the semantic annotations. This result is registered in the semantic discovery and execution. This registration is used to extend the number to provide semantically described web services.

3.1 Semi automated ontology approach

This semi automated ontology approach consists of two main processes namely term extraction and Concept mapping.

Term Extraction

In term extraction process the domain terms from the dataset is extracted using XQuery. Dam's terms are identified as the relevant attribute with high ranking, which can be considered for clustering semantically similar documents. It is found from the literature gene are associated with dams terms. The gene names are extracted from the documents as domain term for mapping with the ontology. The snapshot of XQuery used for extracting domain term is shown in Table 3.1.

Table 3.1 Snapshot of XQuery for extracting Keywords

XQuery expression	Description
XQuery db2-fn:xmlcolumn('INFO.DAM_DETAILS')//Textseqid_accession/text()	Retrieve Single field
Xquery for $pro in db2-fn:sqlquery('Select damname From dam_details where damid=21') let $s1:=$pro/Dam-ref_name/text() let $s2:=$pro/Dbtag for $i in $s4 where $i/Dbtag_db/text()='damid' return($s1,$s2,$Dbtag_tag/Object-id/Object-id_str/text())	Retrieving spatial dam tag

3.2 Ontology Based Document Clustering

Ontology based clustering phase the extracted domain term is mapped with ontology concepts using vector space model. The term relativity between documents is calculated using concept similarity by assigning term weighing. The similarity metrics used for representing domain terms is given below. Given documents (*d1,d2,d3,.....dn)* with extracted domain terms (*t1,t2,t3…tn*) , the terms are mapped with the ontology concepts (*c1,c2,c3…..cn*.) The terms are assigned weights wi based on the concepts and represented in vector space model for clustering documents.

3.3 Semi automated Ontology based Document Clustering

The basics concepts of ontology is designed and used as input for developing the 'ontology based document clustering approach' which semantically grouped from the XML documents. The ontology designed manually is automated in this phase. In term extraction process the domain terms dam names are extracted from the documents. The concept match process the domain term dam name is searched in the ontology to map with other concepts, if dam name not exists in ontology the dam information is fetched from dam's dataset and the ontology is updated automatically.

The Clustering evaluation phase the clusters are evaluated using F-Measure and the proposed semi semantic ontology based document clustering algorithm is compared with other approaches like Precision and Recall.

Procedure
Semi_Semantic_Ontology_Based_Document_Clustering( )
// clustering of spatial dams as XML documents using Ontology based approach
Feature_analysis_Selection( )
Let Selected_features_arr be the set of selected features from XML document
//design schema for integrated ontology
Ont_Schema ← Design_schema()
for each feature in Selected_features_arr do
If feature is a attributes of spatial dam then
//Check the feature is in attributes of spatial dam main concept of ontology
If !(feature € ontology) then
//check the feature attributes of spatial dam occurs as synonyms in any other concept
//in the ontology
If !(checkforsynonyms(feature)) then
//add feature into attributes of spatial dam main concept of ontology
Ontology ←feature
//compute GO annotations for entity
GO_annotation_arr[ ] ← GO(attributes of spatial dam)
For each go in GO_annotation_arr do
Ontology ←go
ConceptMapping(concept, go)
end for each
end if
end if
end if
end for each
add other relevant details specified in table 3.1
using set of object properties and data properties provide mapping among concepts
//clustering process using integrated ontology
Weight_arr[]
←Compute_similarity_metric(ontololgy_concept_corpus)
clusters_arr[ ] ←hierarical_clustering(Weight_arr)
End procedure

4. RESULTS AND DISCUSSION

This section discusses and analyses the experimental results for semi semantic ontology based document clustering approach for semantically grouping spatial dam's documents. The evaluation metric F-Measure is discussed for validating the ontology based document clustering approach.

4.1 Cluster Validation using F-measure

The proposed ontology based clustering approach is validated using Precision, Recall and F-measure. A set of 5 clusters were considered for evaluation is shown in Figure 4.1.

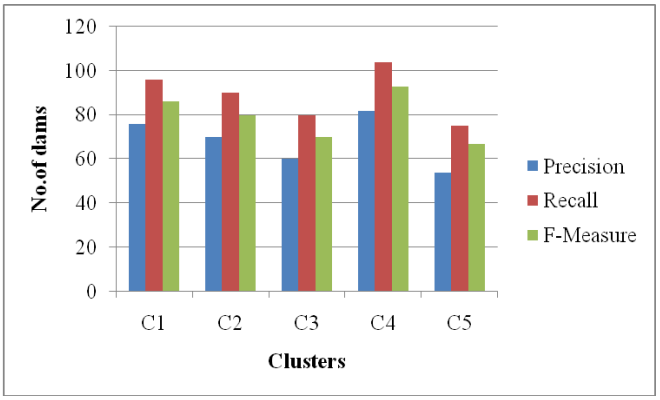

Figure 4.1 F-Measure Analysis for Semi Semantic Ontology Based Document Clustering Approach

On Analysis the average Precision and Recall metrics for clusters are 86% and 95%. The proposed approach of clustering using ontology was found to cluster relevant semantically similar documents.

This section discusses and analyses the experimental results for semi semantic ontology based document clustering approach for semantically grouping dam's documents. The evaluation metric F-Measure is discussed for validating the ontology based document clustering approach.

5. CONCLUSION AND FUTURE WORK

The results of this experiment we inferred that, the proposed semi semantic ontology based document clustering approach is found to be efficient in grouping semantically relevant xml documents with F-measure compared with precision and recall approaches. This work concludes stating that ontology based document clustering group functionally relevant documents in xml format is accurate in terms of semantic relevance. In future, this work can be extended to spatial network and it will be deployed in cloud computing. The cloud computing provide spatial data (Software) to the user as a service (SaS).

REFERENCES

[1] http://gis.nic.in/aboutus.html

[2] http://searchsqlserver.techtarget.com/definition/GIS

[3] M. Agosti, and N. Ferro, "A formal model of annotations of digital content", ACM Transaction on Information Systems, Vol.26., No.1, Article 3, 2007.

[4] Amal Zouaq, Roger Nkambou, "Building Domain Ontologies From Text For Educational Purposes", IEEE Transactions On Learning Technologies, vol. 1., 2008.

[5] Andreas Hotho, Alexander Maedche, and Steffen Staab, 2010, "Ontology-based text document clustering", [Online document], cited: http://www.aifb.kit.edu/images/2/2b/2002_19_Hotho_Text_Clustering_1.pdf, 2010.

[6] T. Berners-Lee, J. Hendler, and O. Lassila,"The Semantic Web", Scientific American, pp. 34-43, 2001.

[7] N. Carla Geovana Macario, and Claudia Bauzer Medeiros, "Specification of a framework for semantic annotation of geospatial data on the web", SIGSPATIAL, ACM, Volume: 1, Issue. 1, pp. 27-32, 2009.

[8] N. Carla Geovana Macario, Sidney Roberto de Sousa, and Claudia Bauzer Medeiros, "Annotating Geospatial Data based on its Semantics", ACM, GIS' 09, 2009.

[9] I.F. Cruz, H. Xiao, F. Hsu, "Peer-to-Peer Semantic Integration of XML and RDF Data Sources", 3rd International Workshop on Agents and Peer-to-Peer Computing (AP2PC 2004), 2004.

[10] A.C.P.L.F. De Carvalho, "Grand challenges for computer science research in brazil", 2006-2016, Workshop report, Brazilian Computer Society, 2006.

[11] Dumitru Roman, Eva Klien, and David Skogan, "SWING-A Semantic Web Services Framework for the Geospatial Domain".

[12] M.J. Egenhofer, " Toward the semantic geospatial web", In Proc. of the ACM GISt'02, pp. 1-4, 2002.

[13] Emerson Muraro., C. Greice Mariano, P. Nadia Kozievitch, Jurandy Almeida. A. Jefersson dos Santos, Ricardo Torres, Bruna Alberton., P.C. Leonor Morelatto, "A Framework for Semantic Annotation of Phenology Image Components".

[14] R. Fileto, L. Liu, C. Pu, E.D. Assad, and C.B. Medeiros, "POESIA: an ontological workflow approach for composing web services in agriculture", The VLDB Journal, Vol.12, Issue 4, pp.352-367, 2003.

[15] Georgios Gkoutos, V., Eain Green, CJ., Ann-Marie Mallon., John Hancock, M., and Duncan Davidson, "Using ontologies to describe mouse phenotypes**",** Genome Biol. 6(1): R8, 2005.

[16] Huan Liu., Lei Yu., 2005, "Toward Integrating Feature Selection Algorithms For Classification And Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 17, no. 4, 2005.

[17] C. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid, "The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing", In Geographic Information Science: Third International Conference, Gi Science, Adelphi, Md, USA, pp. 125-139, 2004.

[18] E. Klien, "A rule-based strategy for the semantic annotation of geodata", Transactions in GIS, 11(3):437-452, 2007.

[19] E. Klien, and M. Lutz, " The role of spatial relations in automating the semantic annotation of geodata", In Proceedings of the Conference of Spatial Information Theory (COSIT'05), volume 3693, pages 133-148, 2005.

[20] Nadia Zerida, and Jin Yao, "Towards Compromising Structural and Bag Of Words Approaches for Clustering Heterogeneous XML Documents", In Proceedings of the Second International Conference

on Advanced Engineering Computing and Applications in Sciences, 2008.

[21] G.Z. Pastorello Jr, J. Daltio, and C.B.Medeiros, "Multimedia Semantic Annotation Propagation", In Proceedings 1st IEEE Int. Works on Data Semantics for Multimedia Systems and Applications (DSMSA)-10th IEEE Int. Symposium on Multimedia (ISM), 2008.

[22] J. Sharmila, and Subramani. "Ontology Based Data Integration in Federated Databases and It's Issues", International Journal of Scientific & Engineering Research, Vol. 3, Issue 6, 2012.

[23] Sonam Pal, and Rakesh Sharma, "An Intelligence Decision Support System for Establishment of New Organization on Any Geographical Area Using GIS", International Journal of Advanced Research in Computer Science and Software Engineering,Vol.3,Issue 8, 2013.

[24] M. Thangamani, P. Thangara, "Integrated Clustering and Feature Selection Scheme For Text Documents", In Proceedings of the Journal of Computer Science 6 (5): 536-541, 2010.

[25] R.Thomas Gruber, "A Transaction Approach to Portable Ontology Specifications", Knowledge System Laboratory Technical Report KSL 92-71, 1993.

[26] Tran Thai Bin, Thilo Wehrmann, Steffen Gebhardt, Verena Klinger, Juliane Huth, VO Quoc Tuan, and Claudia Kuenzer, "Ontology Based Approach for Geospatial Semantic Web".

[27] H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hubner, "Ontology-Based Integration of Information-A Survey of Existing Approaches", In Proceedings of the IJCAI-01, Workshop on Ontologies and Information Sharing, 2001.

AUTHORS PROFILE:

Dr. A. Sasi Kumar received the MCA degree from Bharathiar University, Tamil Nadu, India and M.Phil in Computer Science from Alagappa University, Tamil Nadu, India. He received Ph.D degree in Computer Science from Anna University, Chennai, Tamil Nadu, India. He is currently an Assistant Professor in School of Computing Sciences at Vels University, Chennai, Tamil Nadu, India. His research interests include Data Mining, Big Data, Ad-hoc Networks, Mobile Computing and Cloud Computing.

S.P. Tamizhselvi received the B.E and M.E degree in Computer Science and Engineering from Anna University, Chennai, India. She is currently pursuing the Ph.D degree in Faculty of Information and Communication Engineering at Anna University, Chennai, Tamil Nadu, India. Her research interests include Cloud Computing, Mobile Computing, Data Mining and Ad-hoc Networks.

Dr. G. Suseendran received the M.Sc Information Technology and M.Phil degree from Annamalai University, Tamil Nadu, India and Ph.D degree in Information Technology from University of Madras, Tamil Nadu, India. He is currently an Assistant Professor in School of Computing Sciences at Vels University, Chennai, Tamil Nadu, India. His research interests include Ad-hoc networks, Data Mining, Knowledge-based systems and Web Information Exploration.