ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

An innovative approach in Information Extraction using Ontology for Geospatial Domain

A.Karthic[1]

PG Student, Department of Computer Science, Park College of Engineering and Technology, Kaniyur, Coimbatore, India[1]

Abstract: Information Extraction is an energetic area of research in several fields, such as relational databases, Ontology and data incorporation. The major issue in artificial intelligence is information management, which is critical for any intelligent system. Knowledge plays a very crucial role, and it seems to be more important for finding a solution than the algorithm. Correct knowledge needs to be identified, fetched and organized to develop an intelligent system. Ontology can support all the above tasks. In the ontology construction for any language, WordNet is used. Ontology data processing will be complexly due to the flexibility of the question asking for users. Keyword based searches retrieve too many results that may not relevant. In this work, preferred domain is 'Geography' the input is restricted to the same. The proposed method for ontology based information extraction is totally unsupervised, and it does not require any human interference. This approach reduces the amount of computations required and also it provides additional filtering

Keywords: Ontology; Information Extraction; OWL; Description logic; Relational Database.

I. INTRODUCTION

Ontology is defined as 'Explicit specification of conceptualization' [1]. To construct an ontology from the unstructured text, a numerous of supervised and unsupervised techniques have been proposed. As manual construction of the ontology is a difficult task. Ontology constitutes an approach for knowledge representation that is capable of expressing a set of entities and their relationships, constraints, axioms and the vocabulary of a given domain. Ontology is applied in domains wherever information extraction and retrieval is needed such as data management system, military intelligence, medicine, Wikipedia, social media and enterprise. As noted by Leenheer, and Moor (2005), 'No matter how expressive ontology might be, they are all, in fact, lexical representations of concepts'. Some components of Ontologies are individuals, classes, attributes, relations, rules, axioms and events. Ontology offers many structural similarities irrespective of the language in which they are expressed.

Information Extraction is the process of identification and extraction of instances of a particular class of events or relationships from text and their transformation into a structured representation (e.g. A database) and automatically extracting structured information from unstructured and/or semi-structured machine readable documents. In this work, preferred domain is 'Geography' the input is restricted to the same. Ontology uses standard machine-readable languages to explicitly declare the formal semantics of concepts and their relationships in the domain. Eventually, geographical services, such as showing a map, searching or finding a route between two locations are getting an indivisible part. In philosophy, Ontology deals with "the nature and the organization of reality" [2] or "the science studying of being". Since there is only one reality, there is Ontology that seeks what is there by identifying the types of entities that exist in reality and their relationships. Information scientists extend the concept of ontology to what information users will seek from information systems and knowledge management systems

In a domain ontology uses common machine readable languages to expressly define the formal semantics of concepts and its relationship. Ontology based information retrieval will be challenging mainly because of the ontology. So the proposed mechanism for the ontology framework for geospatial data will be powerful. Geographic information systems were designed for specialized users. In such systems, querying geographic data often requires using complex formal query languages for major tasks. To achieve this resource description Framework (RDF) is used. Resource Description Framework (RDF) for interrogation system will need a domain expert knowledge and data owner authentication as well as the appropriate ontology prediction. The amount of data searches grown rapidly. However, this growth of obtainable data may not be required data as the access is typically supported keyword based search. It leads to lots of tangential data as some term will have completely different meanings in distinct contents. The access of geographical data on the World-Wide Web and mobile devices has opened up a change. e.g. "Nimord" because this is the name of a king associate degree because the name of a software company. Presently it's troublesome to supply this data to a computer program, e.g. exclude all the data regarding the king "Nimord" while not losing information regarding the company.

II. RELATED WORK

Ontology plays an important role in many domains. Manual construction of ontology is a cumbersome task, to construct ontology automatically from the unstructured text there are many supervised and unsupervised techniques are available. Here the main focus on domain specific concept identification and construction of the concept hierarchy [3]. To handle and to extract the information k-partite graph learning algorithm is used. The graph-based learning algorithm is proposed to extract domain specific ontology from natural languages. Here the graph is defined as directed acyclic graph G (V, E), where V is a set of concept nodes and E is a set of relationship edges. The preferred domain is a health domain and the experiments conducted in Hindi and English. Some of the features of proposed systems are, the process does not require human intervention; patterns are generic which can work in any language, and it does not require existing NLP techniques such as NER or parser. Construction of ontology involves three major tasks; they are pre-processing, graph creation and concept hierarchy generation. At last it is compared with manually crafted ontology. The experiments are conducted in Hindi and English, and the performance is evaluated by comparing results ontology with manually constructed ontology for Health domain. This approach not only reduces the amount of computations required for ontology construction, but also provides an additional level of term filtering.

tOWL is an example for a specific domain, as it can be employed for representation and reasoning in a wide variety of dynamic domain [4]. Here financial domain is taken as preferred domain. The Web Ontology Language (OWL) is the most expressive standard language for modeling ontology. There was no standard way of expressing time and time-dependent information in OWL. Knowledge and constraints cannot be enforced using existing OWL-DL semantics [5]. The main goal pursued here is an extension of a fragment of OWL-DL (Description Logic) with time and temporal aspects. Consider, for instance, the temporary relation between people (CEO of the company) and a particular company (Twitter). Such relation can be described by some temporal interval (T1, T2, and T3). Until October 16, 2008, Jack Dorsey was the CEO of Twitter. On that date, Jack Dorsey stepped down, and Evan Williams became the new CEO of the company. In the existing semantic web approach based on OWL-DL, as the CEO of the company changes, there is no way of representing with time. A temporal extension of the very expressive fragment SHIN (D) of the OWL Description Logic language has been proposed which give us the result as temporal OWL language. The language has provided a concrete domain based on the set Q of rational numbers and the set of binary concrete domain predicates. Some of the previous methods such as OWL – Time and the OWL ontology for fluent (representation of change), which only address temporality to a limited extent. The tOWL language meets shortcoming of those methods.

Recent Research for recognizing and processing online handwritten words are general in Latin and Oriental scripts. Word recognition of online handwritten was addressing the problem specifically for two major Indic scripts - Devanagari and Tamil [6]. Two different techniques for word recognition based on Hidden Markov Models have been proposed, they are lexicon driven and lexicon free. Tamil word samples feature the standard symbolic order, but Devangari word samples featuring both standard and nonstandard symbol writing order. In the script, it is important to identify the basic unit of writing. From the building block of their recognition approach, the basic units which are referred as 'symbols', HMMs are built at the character level and then concatenated to form a word-HMMs. The word-HMM is then used to implicitly segment the input into letters as a byproduct of recognition. Data sets of handwritten word sample creation are the first challenge. Then they have used TablePC-based and Digimemo-based data collection to store the word samples. The various steps in recognition such as preprocessing and feature extraction, HMM based modelling of strokes and symbols. There are two algorithms proposed; the first one is for detection, in which inputs are Set of ink strokes corresponding to a word or character where n is the number of strokes and output could be List S containing indices of strokes that correspond to the data. Next algorithm used to match the lexicon entries when words are represented using BoS is based on majority voting. Here inputs will be word choices from recurrent HMM, and lexicon L and the output are Word decisions with confidence scores. Results show that the lexicon-driven approach is highly effective for Tamil and lexicon free strategy address the problem of symbol order variation. This research represents the beginning attempt of online word recognition in Indic scripts in depth.

Learning word meanings from perception is a difficult task [7]. The model is based on an affordance network, such as mapping between robot actions, robot perceptions, and the perceived effects of these actions upon objects. The extended model to incorporate spoken words from WordNet [8]. A robot is communicating with people to understand their needs and intentions. To make a robot looking like a human some interacting object such as camera, data glove and a microphone were used. The problem of bootstrapping language acquisition for an artificial system is difficult to observe efficiently. Baltazar-the humanoid torso, a 14-degree-of-freedom humanoid torso integrated with a binocular head and an arm. This robot equipped with the skills required to perform a set of simple manipulation actions on a number of objects. Some of the actions are grasped, tap, and touch. In addition to this, its perception system allows it to detect objects in front of it and extract information about them. With respect to the word which it hears from human, its perform action. All the action, the object and its property are represented by a variable number of synonyms for a total of 49 words. Such words are stored under bag-of-word's model; it is an unordered set where multiple occurrences are merged. Using ASR (Automatic speech recognizer) robot can classify speech input into a sequence of words, which is implemented as Hidden Markov model (HMM). Spoken words will be correlated with the object features and effects present in front of robot table. The instruction from the user has to be complete and meaningful. The use of the model in practical applications, such as interpreting instructions or using context in order to improve speech recognition.

III. PROPOSED METHODOLOGY

a. Dataset Prediction :

In this proposed method the main concentration on the dataset from the geospatial domain. For this process, any dataset is chosen from the UCI repository with the objective of the spatial data. The dataset will contain the details about states, city, river, road, lake and mountain. This will help to provide the answers for user search queries. The input dataset will be in the form of the RDF with the extension of OWL. In this proposed ontology based extraction, the data need to select from the domain expert. Then will be uploading the data to search engine from the given dataset and use the RDF dataset for the ontology process. At the home page, any data set can be browsed and attach which is dynamic. The purpose of this selection is, at any time any dataset can be integrated and work with the further processing. But the format must satisfy OWL and the ontology type must be RDF. It predicts the classes in the RDF file. Next have to predict the properties of the classes. Properties are categorized into two groups:

- Data Type Properties
- Ontology properties.

Once the properties are categorized, every RDF class will be shown. With this, it is easy to identify the tables in the database of Wamp server. After creation of tables and its properties the server needs to be started. RDF properties have interlinked between all the classes, and it allocates relationship among one class with another.

b. Triples Conversion :

The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specification. Using a variety of syntax notations and data serialization formats. It is similar to classical conceptual modeling approaches such as Entity-Relationship or class diagram. After identifying the class names and its properties, here in this module the content of the domain expert data set is extracted. It is complex to extract all the exact information from the data set without finding its components. By applying the N-Triples algorithm, it can split the content into the subject object predicate. It is the combination of three words that form any sentence. The RDF data model is subject-predicate-object expressions. In before approaches, the answering system ontology formation is very complexity due to the ontology formation. The subject refers to the resource, the predicate indicates traits or aspects of this subject and expresses a relationship between the subject and the object. With the server, a database is integrated with the RDF classes, that can store the elements of the data set. These expressions are known as triples in RDF terminology.

c. Input Processing & Similarity measurement :

As all the required data has been stored in the database as tables and fields, now it is necessary to allow the solution seeker to ask the queries from the user interface. Here two types of processing the information. First the input string from the user is received and stored as text (.txt) file. That particular text file is taken for processing, and it will search with ontology relations. It will process through all the tables and its fields; every class will be matched semantically which will produce similarity among them. Those similarities are shown as a table to the user that shows all the class and its numeric similarity. The Next way of processing is voice information, which is stored, and its wave signal can be viewed by them to check the processing, and the voice file is stored as the audio format for further processing. After predicting the similarity values between the data, the top most similarity valued data from the domain expert server for a given query. The proposed method has the two level ontology answering support for the user query; the text file based searches, and the audio based retrieving of the query answering support for user convenient. This will improve the accuracy of the retrieved results from server data's

IV. CONCLUSION

Resource Description framework for interrogation system will require the domain expert knowledge and data owner authentication as well as the appropriate ontology prediction. The proposed method of RDF creation using protégé for data collected from the data owners and the domain experts. It will improve the Ontology structure to get the items. The Triple Conversion to predict the data in the RDF file is implemented. Triples are used to classify the RDF file to the Subject, Object, and Predicate Classification and also introduces the Question formulation mechanism for enabling the user-defined semantic query search. In this mechanism first formulate the Question by using the Solution seekers semantic queries. Generate the rules for the web ontology. These rules are the Fuzzy rules based on the Ontology Generator. It will use the user query and Triple data to retrieve the solution of the Given Semantic Question. The proposed technique overcomes the disadvantages of keyword-based search and flexibility of query searching is improved. Ontology based Information Extraction is totally unsupervised, and it does not require any human intervention. This approach not only reduces the amount of computations required for ontology construction, but also provides an additional level of term filtering.

REFERENCES

[1] Thomas R. Gruber "Toward principles for the design of ontologies used for knowledge sharing" Int. J. Hum.-Comput. Stud. 43, December 1995, pp. 907-928.

[2] N. Guarino, "Formal Ontology in Information Systems" Proceedings of FOIS'98, Trento, Italy, Amsterdam, IOS Press, June 1998, pp. 3-15.

[3] James N. K. Liu, Yu-Lin He, Edward H. Y. Lim and Xi-Zhao Wang "A New Method for Knowledge and Information Management Domain Ontology Graph Model" IEEE Transactions on Systems, Man and Cybernetics: Systems, Vol. 43, No. 1, January 2013, pp. 115-127.

[4] Viorel Milea, Flavius Frasincar, and Uzay Kaymak, "tOWL: A Temporal Web Ontology Language" IEEE Transactions on Systems, Man and Cybernetics, Vol. 42, No. 1, February 2012, pp. 268-281.

[5] Alexander Maedche and Steffen Staab "Ontology Learning for the Semantic Web" IEEE Intelligent Systems 16, March 2001, pp.72-79.

[6] A. Bharath and Sriganesh Madhvanath, "HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 4, April 2012, pp. 670-682.

[7] Giampiero Salvi, Luis Montesano, Alexandre Bernardino, and José Santos-Victor "Language Bootstrapping: Learning Word Meanings from Perception–Action Association" IEEE Transactions on Systems, Man and Cybernetics, Vol. 42, No. 3, June 2012, pp. 660-671.

[8] George A. Miller "WordNet: A Lexical Database for English". Communications of the ACM Vol. 38, No. 11, June 1995, pp. 39-41.