Automatic Relationship Construction in Domain Ontology Engineering using Semantic and Thematic Graph Generation Process and Convolution Neural Network

Sivaramakrishnan R Guruvayur, R.Suchithra

Abstract— ***In recent studies, Ontology construction plays an important role in translating raw text into useful knowledge. The proposed methodology supports efficient retrieval using multidimensional theory and implements integrated data training techniques before enter the trial process. The proposed approach has used the Semantic and Thematic Graph Generation Process to extract useful knowledge, and uses data mining techniques and web solutions to present knowledge as well as improve search speed and information retrieval accuracy. Established ontology can help clarify what it means for different ideas and relationships. Due to the rise of the ontology repository, the process of matching can take a long time. To avoid this, the method produces a hierarchical structure with in-depth interpretation of the data. A system is designed to remove domain dependencies using a dynamic labeling scheme using basic theorem, and the results show that it is possible to automatically and independently construct an independent domain.***

Index terms: Automatic Ontology Generation, Semantic Web, Semantic Graph Generation, Thematic Graph Generation Process and Convolution Neural Network.

1. INTRODUCTION

Ontology is built primarily as a formal specification of ideas and a link between the properties of these concepts. Punctuation is defined as a partial specification of the conceptual terminology used to form knowledge theories in the field of speech. Ontology is applied in areas such as natural disaster management, medicine, military, enterprise, agriculture, Wikipedia, automobiles, and so on. Autism is also presented as a formal representation of knowledge through a set of ideas in a field and the relationships between these concepts. Pathology has four major components for domain representation. They have:

i. Ideas are a set of elements in a domain.

ii. Relationships determine the interaction between ideas.

iii. The examples show specific examples of ideas in the field.

iv. Axioms mean statements that are always true.

Ontologies find use in the Semantic Web, Artificial Intelligence, Systems Engineering, Biomedical Informatics, Software Engineering, Enterprise Bookmarking, Library Science, and Information Architecture in a type of representation of knowledge about the domain or a part of the domain. There are 4 groups of ontologies: static, dynamic, social and intentional.

A static ontology elaborates items in existence; their attributes and the relationship with them. Dynamic ontology explains the domain as states and their transitions including processes. Social ontology describes social scenarios, permanent structures in an organization or changing networks of independencies and alliances. Intentional ontology includes the domain of agents, things wanted, believed in, proved, disproved, and discussed about. In any domain, the ontology is the core system of representing information for that domain. Without the existence of the ontology, or the conceptualizations which form the basis of knowledge, a vocabulary of knowledge representation cannot exist. A domain specific ontological study gives clarification to knowledge structure.

Creating domain ontology is also important in defining and using a framework of enterprise architecture. Therefore, the first stage of creating an effective system for representation of knowledge and vocabulary is an effective ontological study of the domain. A weak analysis will result in knowledge bases that are incoherent.

For building a language of knowledge representation based on analysis, an association of terms along with ontological concepts and relations, and creation of syntax for knowledge encoding pertaining to the concepts and relations is necessary. This language of knowledge representation can be shared with those having similar requirements for representing knowledge in that domain, hence negating the requirement for repeating the process of knowledge analysis. Ontology sharing can hence create the basis for domain specific information-representing languages. When compared with the earlier generation of such languages (say, KL-1), these languages have rich content besides having a many terms embodying complex constituent of the domain.

Domain specific ontologies are created for the following purposes:

Manuscript published on 30 September 2019
* Correspondence Author

Sivaramakrishnan R Guruvayur*, Department of Computer Science, Jain University, Karnataka, India
R.Suchithra, Department of Computer Science, Jain University, Karnataka, India

1. Sharing common understanding of structures of knowledge among software agents or users.
2. Enabling reuse of domain knowledge for consumption of upstream and downstream AI applications.
3. Making explicit domain assumptions – Connected data.
4. Separating operational and domain knowledge.
5. Analyzing knowledge of the domain.
6. Challenges in Building Domain Specific Ontologies.

Following are the challenges in building domain specific ontologies:

1. Insufficient coverage for limited domains: Typically, the coverage for domains having lesser web presence is lesser than that of domains that are more popular.
2. Relational identification: Generally, the extraction relations are not spelled out in advance. Hence, for domain specific ontologies, identifying relations needs deep domain expertise.
3. Resolving entities: The issue of identification and grouping/linking various manifestations/occurrences of the same real-world item is a difficult task.
4. Disambiguation of Entities: A phrase or word may imply more than one entity. Entity disambiguation involves association of the phrase/word with the most relevant entity.
5. Problem of temporal knowledge base: There are facts which vary with time; hence mapping the phrase/word with the relevant entity can be an issue.
6. Extracting values: Ontologies are generally represented in the form of triples, i.e., <Entity, Relationship, Entity> . The system is required to learn all possible formats of entity/relationship to enrich the ontology.
7. Confidence of facts: There are many disputable facts which depend on the source of information. Hence, it is often difficult to locate a unique entity if there is a conflict.

Unfortunately, building and maintaining a theory is a difficult task. Classical theoretical construction relies on domain experts, but is expensive, time consuming, and controversial [14]. Aside from the lack of standards, the field also lacks a method for acquiring full automation knowledge: theory building is a time-consuming and costly procedure. Although current methods of building theory can achieve partial automated classification, there are limitations such as human labor requirements and limitations. In order to overcome the above problems, this paper proposes a novel method based on Semantic and Thematic Graph Generation Process and Convolution Neural Network approach to automatically construct ontology. And this ontology is domain independent also.

The paper remains are prepared as follows. Section 2 reviews your work. The steps for initial ontology creation and ontology updation and validation process are proposed in Part III. Section IV describes the experimental results of the proposed method. In this section the proposed method is compared with the existing approaches to show the efficiency of the novel approach. The conclusion as well as the future work of this study are presented in Section V.

II. RELATED WORK

The ML method teaches input relations - from examples to interpret new inputs. Therefore, their effectiveness is highly dependent on the choice of presentation data (or functions) that they perform [20]. Different models have been proposed to represent words as continuous vectors to evaluate the representation of subsequent words and generate distributed digital models (DSMs). DSMs derive representations in such a way that words occurring in similar contexts have similar representations, and so the context must be defined. Classical DSMs include latent semantic analysis (LSA) [21], which typically accepts whole documents as context (e.g., text files, examples), and hyperlinks analog to language (HAL) [22] that accepts slides. Random indexing [23] is emerging as a successful option for SL. LSA, HAL, and random indices were stimulated by SDS. An example of a valid DSMs is a valid LSA. While SDSs compare with the use of distance meters in high-dimensional space [26], ECDs, such as SDs or PMSs, measure similarity between terms according to the extent to which they share the same spatial distribution [26]. Most DNAs are costly to calculate and load associated with model building or modification due to the large number of dimensions involved in large-body modeling [26]. This study applies to neural language models, e.g. representative classification of words learned from neural networks (NN). Although neural models are not new to DNA, recent advances in ND have made it possible to extract words from thousands of words, thus increasing interest in in-depth studies and EMM and S-models. Gear-turn [16, 17]. CBOW and Skip-gram gained popularity when they formed the basis of comparative metaphor [27] and served as the basis for comparative practice [28]. CBOW and Skip-gram have already been trained to produce high quality keywords from English Wikipedia [27, 29]. Pyysalo et al. [30] and Minoriro-Gimenez et al. [19] was the first to apply the neural language model to the PubMed Corporation. Pyysalo et al. [30] used Skip g, which contains 22 M PubMed articles and 672 K PubMed Central Open Access articles. . It is necessary to represent the available words (1–5 g) from the reusable literature. Minarro-Gimenez et al. [19] used small datasets from Pubet, as well as from other medical fields (eg, mmometrics, 31mm), MIDV [32], and non-medical sources (e.g., Wikipedia [33]). Many later studies established the embeddedness with CBOW and Skip-gram using the PubMed Corporation. We describe some studies by considering four works focusing on their words, texts, ideas, and relationships. At the end of this section, we include studies that include word generation based on technology. The similarity and coherence of the Working Group et al. [34] is consistent with more recent studies (Hill et al. 2012). Pedersen et al. [34] states: "Similar ideas are immediately considered in relation to their similarity." And Pedersen et al. [34] and Hill et al. [36] Consistent with Reed's view, [37] , the most direct affinities, is a special case of the most significant correlations. Pedersen et al. [34] advocate half-similar measures based on relationships in which hierarchical ideas are directly or indirectly related. Prior to Pedersen et al. [34] Caviedes and Cimino [38] examine indicators of conceptual similarity based on the minimum number of parental links between concepts.

A study by Caviedes and Cimino [38], Pedersen et al. [34] Hill et al. [35] and Parochov et al. [36] provided a dataset of word pairs together with relevant / similar human judgments. [35]. The data set of Hill et al. among the 999 text pairs, the test collection was similar to the S word. S-353 [39] (353 pairs) and MEN [40] pairs (3K words) are common English words, where these datasets can be considered as the gold standard for evaluating semester samples. Muneeb [41] et al. applied Skip-gram and CBOW to 1.25 M articles, PubMed and evaluate the quality of the embeddings using Pedersen et al. 34 pair of words. Muneeb [41] et al. concluded that skip gram is more appropriate than CBOW for similarities and similarities. Chiu et al. [42] was used by Pyysalo et al. [30] .A dataset and an English publication of more than 10 ms PubMed of BioOIC Challenges [43] for internal evaluations of Gregg and CdW embedded with Pokwoo Al. [36] pairs of words. Chiu et al. [42] Concluded that skipping gram generally shows better results for similarity and connectivity than synergy. Task Hill et al. [33] interprets "cohesion" as "association" and the most similar are synonyms. Two well-known sets of semantics assessments are 80 TOEFL queries (English in foreign languages) from [21] and 50 ° C. S. R. (English as a Second Language) . Both studies [21] and [44] have four encyclopedic questions that require knowledge of common English words. It should be noted that the TOEFL dataset was used in the LSS [21].

Sebastiani [15] states that classification of text is "the activity of labeling a natural language text with a subject type of a predefined set". Therefore, submitting a keyword or keyword phrase from MeSH to a title or title, PubMed / MEDLINE plus an abstract, is a type of text classification known as the MeSH index. The 2017 BioSMS Competition has three tasks, one of which is the indexing of the SCH. Request for participants to classify new articles from Pubet before the curator manually provides the MeSH terms with some assistance from the LMS Medical Text Index [57]. MeSHLabeler is a MeSH indexing algorithm (Liu et al ... [58]) that outperforms MTI and wins the BioASQ competition to index MeSH over two and three years of competition. Both MTI and MeSHLabeler [58] use classical transcriptions. Pheng et al [59] used more than 1 mm of amygdala extracts. (Some Downloads from the NHL and some from the 3-year GOS competition) and introducing SimaAA, a CBOW profitable workflow with better results (2% - Small Scale Small F) from MeSHLabeler. It should be noted that MTI, MeSHLabeler and DeepMeSH used the implementation of the nearest-neighbor K algorithm.

There are four basic types of folk theory building methods. The seven-step method [1] was developed by Stanford University, the main method for the development of pathology. This method is applicable for the development of pathology but lacks verification, evaluation or feedback from users. TOVE [2] is a method of modeling science based on pathologic evaluation. Groning and Fox used it to upgrade the engineering sciences of the HSE. The skeleton method [3] is a summary of the experience of EC (corporate ontology) by Uschold and King. The method clearly describes the process and direction of the etiology, but does not specify the purpose of the pathology. METONTOLOGY [4] is designed to develop chemical theories capable of formulating theory in accordance with the knowledge of a single level.

This approach emphasizes the reuse of science but does not reflect the evolution of theory. The methods of classification are classified into two categories: principle-based and statistical-based classification [5, 6]. In terms of principle-based categorization, it must be supported by specific knowledge and principles in this area. Also, many statistical machine learning methods based on statistics are applied to the text classification system. The earliest method of machine learning was Naive Bayes [7, 8]. From now on, almost all major machine learning algorithms have been applied to the field of text classification, such as KN (Neighborhoods), DNA (Vector Support), Quadratic Methods, and Solution tree [9-12].

III. ONTOLOGY MODEL CREATION

The model proposed in this document is based on extracting Jaccord relations from text documents and using conceptual and relational ontological models. The proposed innovation of the scientific model is a combination of the use of two different extraction methods, Semantic Graph and Thematic Graph and validates the results with a third method that analyzes external service descriptors.

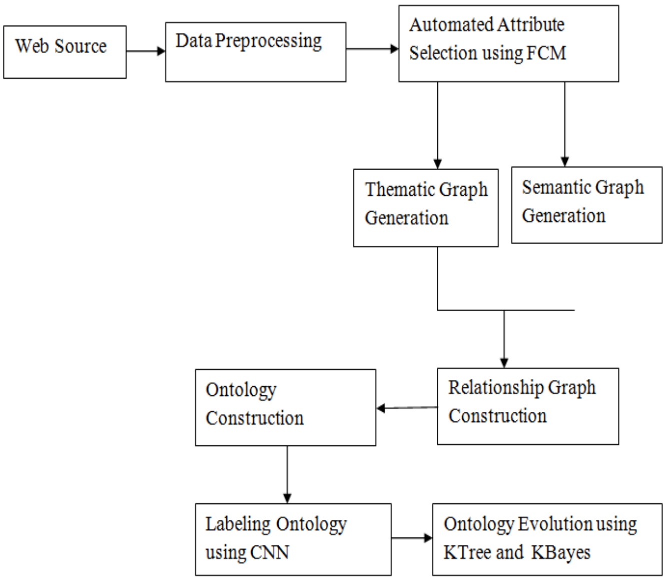

Web Source
Data Preprocessing
Automated Attribute Selection using FCM
Thematic Graph Generation
Semantic Graph Generation
Ontology Construction
Relationship Graph Construction
Labeling Ontology using CNN
Ontology Evolution using KTree and KBayes

Fig. 1. Ontology Designing Process

We used these three methods to demonstrate the feasibility of our model. Other more sophisticated methods, such as machine learning (M) and information retrieval (EIS), can also be used to implement the model.

However, the use of direct methods emphasizes that many methods can be "accounted for" and that the results are attributed to the process of combining and verifying the model. The overall process of pathology is described in Figure 1.

3.1 Preprocessing

This is the first step in the process. At the processing stage, the data is transformed into a data cleaning step. Text is just a word sequence or, more precisely, a character sequence. But when we are usually concerned with language modeling or natural language processing, we are concerned with the whole term rather than concerned with the depth of character of our textual data. One reason for this is that in the language model, the individual characters do not have much context. Characters like 'd', 'r', 'a', 'e' do not have any context, but when rearranged into words, they can form "readings" that may explain some actions that probably did exactly that now.

3.1.1 Vectorization

A vector is just a method of converting words into a long list of numbers that can have some complex structure to be understood by a computer using some kind of machine learning or data extraction algorithm. But even before that, we had to perform a sequence of operations on the text so that our text was "cleared". The process of "clearing" data can vary depending on the data source.

3.1.2 Removing Unwanted Characters

The key step is the process of clearing the text. If we are missing some text from the HTML / XML source, we will have to delete all HTML units, non-alphabetical punctuation, and other character types that are not part of the language. Common methods for such cleanup include regular expressions that can be used to filter unwanted text. There are a number of systems where the main English characters are kept, such as stop signs, question marks, surprising symbols.

3.1.3 Tokenization and Capitalization/De-capitalization

Tokenization is simply the process of dividing a sentence into words.

3.1.4 Removing/Retaining Stopwords

This cleaning step also depends on what you will eventually do with your data after the previous process. The word stop is a frequently used word and is so common that they lose their lofty meaning. Words like "yes this is" are just a few examples of the word stop. In applications such as file search engines and file rankings where keywords are more important than common words, quitting may be a good idea, but if there are programs like search, lyrics or search for specific quotes. Termination can be important. Think of examples like "do or not," "see what made me do it" and more. The word stop in such phrases really plays a role and therefore should not be omitted.

There are two common methods of removing terminations, and both are clear. One way is to count all event words and assign a numeric value for the number and get rid of any words / words that occur more than the specified value. The other way is to have a predefined word break list that can be removed from the list of symbols / symbols. The expressions of some people, such as "Lola Lov, Blues Basset", may also be valuable information when working on systems based on sentimental / sentimental analysis, but for systems that require a more formal application type. This may also be possible so out.

3.1.5 Parts of Speech Tagging

POS marking refers to the assignment of a language category (often called SES) to each word in a document based on syntactic behavior and morphological features. The most common types of POS in English are: nouns, adjectives, adverb pronoun preposition, conjugation and pronunciation. There are other types that arise from different forms of these types, such as the verb may be the basic form or the past tense.

3.2 Attribute Selection and Relationship Graph Construction

Once the process completes, the next step is to select the attribute from the pre-processed data. This attribute selection is used to reduce the dimension of the data. To reduce the amount of data, the speed of cancer formation has been optimized. When selecting the attribute, the word vector is reduced to eliminate unnecessary word order in the operator.

3.2.1 Attribute Selection using Fuzzy C-Means Clustering

Clustering involves separating attributes from the input of preprocessed data into different classes based on weight values. Therefore, this attribute in the same class is as approximate as possible, and the attributes in the class are as different as possible. Several clustering algorithms have been introduced before. Because clusters can be considered as a set of group data sets, classification methods are possible according to whether the subsets are blurred or clear. In general, the efficiency of the fuzzy clamp method is better than other existing clustering methods. Because Fuzzy-C means (FCM) places attribute data in a group of clones, with each attribute in the input data belonging to each cluster to a certain level.

The main idea of FCM is to represent the similarity of a point with a functional cluster (membership function) where the value (membership) is between zero and one. They show the strength of the relationship between this data element and the specific cluster. The FCM algorithm was developed by Dunn [13]. FCM introduces membership at each sampling point in all clusters through a membership function that varies between zero and one. In this way, the point at the edge of the cluster may be smaller than the cluster in the middle of the cluster. The membership count for each sample point must be a number. The FCM clustering algorithm can be summarized as follows.

Let P={ p_1 , p_2 , p_3 , ..., p_q } be a set of given dataset, where each data point p_q (q=1,2,3,...,n) is a vector of size n . U_{vq} is a set of real matrices and v be an integer, $2 \leq v < q$. Then fuzzy C- mean partition space for P is,

$$MF_{FCM} = \{U \in U_{vq} : U_{jk} \in [0,1]\}$$

$$\sum_{j=1}^{c} \mu_{jk} = 1 \text{ where; } 0 \leq \mu_{jk} \leq 1, k = 1, q$$

μ_{jk} of the membership of k^{th} data point in the j^{th} cluster, $j = \{1,2,3, \dots \dots c\}$ (3.1)

The algorithm starts with randomly selecting centers and then in every iteration, determines the Fuzzy membership of each attribute, until there is no change in the cluster centers. In addition, once the cluster centers are established, each attribute is assigned to the group with which it has the highest membership value. It is based on minimization or iterative optimization of the following objective function under the fuzzy constraints defined in (1.1)

$$H_m = \sum_{k=1}^{n} \sum_{j=1}^{c} \mu_{jk}{}^{m} \; d^2(p_k v_j) \qquad (1.2)$$

where

$$\mu_{jk} = \left[\sum_{i=1}^{c} \left[\frac{d(p_k, v_j)}{d(p_k, v_i)} \right]^{\frac{2}{m-1}} \right]^{-1} \qquad (1.3)$$

$$v_j = \frac{\sum_{j=}^{n} \mu_{jk}{}^{m} p_k}{\sum_{j=1}^{n} \mu_{jk}{}^{m}} \qquad (1.4)$$

The equation (1.2) is a least squares function, where the parameter n is the number of data sets and c is the number of classes (partitions) into which one is trying to classify the data sets. $d^2(p_k v_j)$ is the Euclidean distance, v_j is center vector of jth cluster and p_k is vector of kth attribute. The process stops when $|\mu(t+1)-\mu t| \leq \varepsilon$ or a predifined number of iteration is reached, ε is a termination criterion, small positive constant between 0 and 1.Hm is the objective function which is used to assign every attribute onto the corresponding cluster

FCM Algorithm

1. The algorithm is similar to the k-means algorithm:
2. Select the number of clusters.
3. Assign the function to a point so that it is in a cluster.
4. Repeat steps 2 and 3 until the algorithm enters (that is, the coefficient shift between the two alarms does not exceed the given stimulus level).

Calculate the insulin for each cluster using equation (1.4) above.

6. For each point, calculate the value of its member function in the cluster using equation (1.3)

The algorithm minimizes variation in clusters but has the same problem as k-means. The minimum is the local minimum and the result depends on the initial weight selection. Using a mixture of Gaussian along with the maximum expectation algorithm is a more statistical approach that includes some ideas: some class membership. Another algorithm that is closely related to fuselage is the K softness method. Electronic devices are a critical tool for data processing when grouping objects. The mathematician introduced the delay term in the FCM algorithm to improve the accuracy of the noise coherence.

3.2.2 Construct Relationship Graph Model using Semantic and Thematic graph

After reducing the dimension of the data, the next step is to create a model of the correlation graph. In this model of generational relationships, the subject / object is separated and projections (connections) are found in any context. This relational graphic model is constructed using graphical and thematic content.

Semantic graph construction

The first step in preparing a ontology-based classification article is to create a raw graphic based on the text of the document. The purpose of graphical graphs is to shift the focus of the analysis of words, strings, and phrases contained in the document to the human image and the semantic connection between them. We will assume that the step-related word retraction and termination can be applied to the text of the document prior to the subject marking step.

The ontology subject found in the analysis file is defined by the matching clause of the document with the consonants of the subject (used as the subject name) contained in the theory. Such letters are usually represented as values of specific properties associated with the formulation and used as their identities. We assume that these properties determine the name of an entity (commonly known as its label) and can also define the same meaning (nickname) for the entity's name. We determine the unit match weight based on which markers are used in the match. We like real matches with company labels. The topic names can be compared in many places in the document. Important information, similar to the frequency of words used in traditional text classification methods, is important. The randomness of the unit with many phenomena is reflected by the weight gain of the unit. However, to determine the overall unit weight gain, we use the following formula to determine the initial weight of each unit:

$$w = 1 - \frac{1}{1 + \sum_{i=1\dots n} s_i * p_i}$$

In formula, w is the initial weight of the unit and n represents the number of matches for the formulation. The term pi represents the weight of a finite property that associates the corresponding letters (names or nicknames) with elements in the IT and C matches as a measure of the similarity between the literal and the corresponding text phrases, taking into account any differences. One composed of compound words and / or stopped word removal. If the subject identification process does not include a word restriction and a word limit set to 1.

Thematic graph

Analytical documents can cover more than one topic. In addition, many elements can be integrated into Czech graphs during random stages, even if they are not related to or perhaps related to the main topic of the document. In addition, specific clauses in the file may result in multiple entity identifications, but may be the only one that represents the correct match in the context of the document. The steps of this algorithm involve the selection of a graph of the previous semester graph, which is the best explanation of the unit and the relationship that is recognized. We call graphs in such terms.

Retrieval Number C6832098319/2019©BEIESP
DOI: 10.35940/ijrte.C6832.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The choice of thematic graphics is based on the assumption that the topics in the topic are interconnected, forming relevant components in the graph. Graphs are made with the help of people and relationships from theology, so the subject and relationships in this component must fall into one subject (category). A subject in a graph that is unrelated to another element or belongs to another is perhaps a smaller set of related components, most belonging to another subject. If the document focuses on a specific topic (which is an assumption for automatic text classification), the digital graphics of the file should have a single or minimal dominant theme chart that matches the main theme of the document. For further analysis and classification, we select the thematic graphs with the largest number of objects and the largest number of elements. If several thematic charts had similar estimates, they were included for further analysis.

If more than one thematic graphic is selected, it may indicate that the file is focused on more than one subject. Selection of the dominant graph effectively eliminates elements that are not relevant to the main topic of the document, such as incorrectly selected units or ambiguous entities of the same name. In addition, graphite reduction results in the removal of porous (or slotted) sections that have a poor correlation with the dominant graph. This step reduces the amount of low-cost information, reduces the level of disturbed information, and allows analysis to shift to the main topic of the document. In addition, we calculate the mid-points of a subject in a thematic graph to find the central unit, most often the mark of the subject. In our experiments, we used a geographical measure to find most central units. A measure of geographical location is defined as the sum of the sum of the shortest paths between the selected peaks and the other peaks in the composition:

$$\text{Centrality}(X_k) = \frac{1}{\sum_l s(X_k, X_l)}$$

where $s(X_k, X_l)$ is the shortest path distance in between vertices X_k and X_l. The computation of the organ and the measure of the central segment leads to the localization of the underlying structure in the graph. The best bodies and most central organizations are selected as the core of the thematic graph. They are determined to be most relevant to the subject of the document. Note that the best authorities should not be the most central and vice versa.

3.3 Ontology Construction

In this step, the Ontology Model is constructed using Jaccard relation estimator. After constructing algorithm, the next step is to labeling the constructed ontology based on Convolution Neural Network. The CNN based ontology construction is separated into two phases such as offline and online phases. The block diagram of ontology construction based on convolution neural network is shown in fig.2. The amount of data is divided into different categories using labels based on different domains. In the study phase, preprocessing, selection, attribution, and theorem formation with loss function were performed to develop predictive models. First, mark the training data set. Data Sizing Before performing data size changes. Finally, neural network channels are used to automatically generate pathology. The dataset was taken from the net. The network is one of the pre-made models. If you want to train from the first layer, we need to train the whole layer (meaning) to the last layer. Therefore, time consumption is very high. This will affect performance.

To avoid this problem, a model based on brain training was used for the classification step. In the proposed CNN we will only teach the latest Java implementation. We do not want to train all layers. Therefore, the computation time is low, while the productivity is high in the scheme of the proposed ontology. The loss function is calculated using the gradient generation algorithm. Raw image data were compared to classifications using the evaluation function. The quality of a given set is measured by the loss function. It is based on the results obtained with the key labels in the offline data approved. Calculating the loss function is crucial for improving accuracy. If the loss function is very high, the accuracy is low. Similarly, accuracy is high, low loss function. The slope value is calculated as a loss function to calculate the slope algorithm.

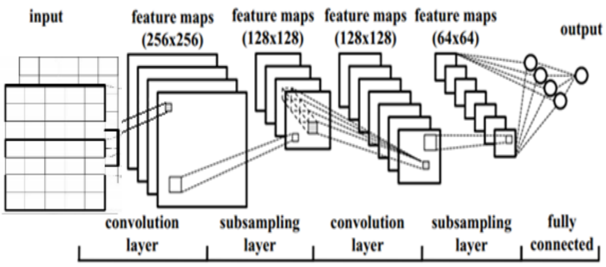


Fig. 2 Ontology construction using CNN

Repeatedly evaluate the gradient value to compute the gradient of loss function. Algorithm for CNN based Classification is shown in Algorithm 2.

Algorithm 2

1. Apply convolution filter in first layer
2. The sensitivity of filter is minimized through smoothing the convolution filter (i.e) subsampling
3. Transferring the signal from one layer to another is controlled by an activation layer
4. Tighten up the training period using the Rectified Linear Unit (RELU)
5. Neurons in the flow layer are associated with each neuron in the next layer
6. During a offline process at the end add a loss layer to give feedback to the neural network

3.4 Ontology Evolution

The development of pathology consists of four steps:

1) To create new ideas,
2) Define the relationship
3) Identification of contact types
4) Restart the configuration process for the next WSDL file.

Creating a new idea is based on improving a defined idea. Reviving an idea in the previous step does not guarantee that it must be integrated with current pathology. Instead, emerging ideas must be analyzed in relation to current pathology.

To evaluate the relationship between concepts, we use the KTree and KBaye algorithms.

4. RESULT AND ANALYSIS

4.1 DATA SET USED

In this paper, the ontology is constructed from various domains. For creating multi domain ontology Agriculture, Cancer, Pizza and Books domain are chosen. Next the crawler searches for the terminologies in the great amount of unstructured text. The training set is a set containing more than 68000 texts collected from agricultural university, cancer instiutes, pizza shops and central libraries. From the collected data, this paper use 47600 marked texts as training set, 13600 of them as verification set and 6800 of them as test set.

4.2 Performance Parameters

To evaluate the performance of the proposed ontology constructing process, several performance metrics are available. This paper uses the Detection Accuracy, Precision Rate, Recall Rate, Sensitivity, Specificity, F-Measure and Error Rate to analyses the performance.

5.1.1. Detection Accuracy

Detection Accuracy is the measurement system, which measure the degree of closeness of measurement between the original labeled texts and the correctly labeled texts

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4.1)$$

Where, TP – True Positive
FN – False Negative
TN – True Negative
FP – False Positive

5.1.2. Error Rate

Error Rate is the measurement system, which measure no of falsely recognised characters form the given input character datas.

$$\text{Error Rate} = \frac{\text{No of Datas of Falsely labeled texts}}{\text{Total No of texts}} \quad (4.2)$$

5.1.3. Precision Rate

The precision is the fraction of retrieved instances that are relevant to the find.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4.3)$$

Where, TP – True Positive
FP – False Positive

5.1.4. Recall Rate

The recall is the fraction of relevant instances that are retrieved according to the input data.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3.3.4)$$

Where, TP – True
FN – False Negative

5.1.5 Sensitivity

Sensitivity also called the true positive rate or the recall rate in some field's measures the proportion of actual positives.

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$

where, TP – True Positive (equivalent with hit)
FN – False Negative (equivalent with miss)

5.1.6. Specificity

Specificity measures the proportion of negatives which are correctly identified such as the percentage.

$$\text{Specificity} = \frac{TN}{(FP+TN)}$$

where, TN – True Negative (equivalent with correct rejection)
FP – False Positive (equivalent with false alarm)

5.1.7 F-Measure

F-measure is the ratio of product of precision and recall to the sum of precision and recall. The f-measure can be calculated as,

$$F_m = (1+\alpha) * \frac{\text{Precision} * \text{Recall}}{\alpha * (\text{Precision} * \text{Recall})}$$

Experiment No 1: Analysis of Attribute Selection Approaches

In this experiment, this work will evaluate the contribution of each attribute selection approaches which are employed in the work. To assess the efficiency of this ontology construction technique, the PPV, DSC, JI and SEN measures are employed. Ideally, a good attribute selection scheme is expected to have a high PPV, DSC, JI and SEN value. Table 1 lists the PPV, DSC, JI and SEN measures of attribute selection approaches.

Table 1: Analysis of PPV, DSC, JI and SEN of Atribute selection Approaches

DataSet				
Attribute selection Approaches				
Metrics	DSC	PPV	JI	SEN
PCA	85.882	92.962	96.962	95.562
KNN	97.692	96.332	98.992	98.752
FCM	98.326	98.157	98.926	99.824

As observed from Table 2, the PPV, DSC, JI and SEN of the FCM in range 96-97, which is superior than that of the other attribute selection scheme. So the FCM is best for the attribute selection scheme. Fig.9 depicted the PPV, DSC, JI and SEN measures of attribute selection approaches.

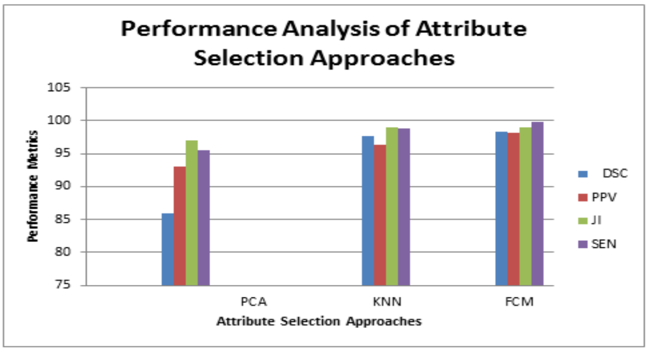


Fig. 3 Analysis of PPV, DSC, JI and SEN of BRATS 2015 for attribute selection Approaches

As observed from Fig.3, the PPV, DSC, JI and SEN of the FCM in range 96-98, which is superior than that of the other attribute selection scheme. So the FCM is best for the attribute selection scheme.

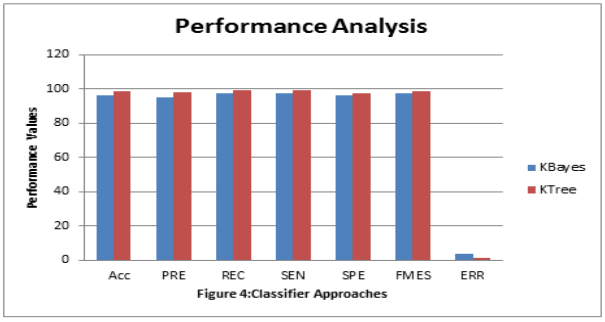
4.3.2 Experiment No #2 : Performance Analysis of Ontology Evaluation

In this experiment, we will evaluate the contribution of each classifier approaches which are employed in the work. To evaluate the performance of this feature retrieval scheme, the Detection Accuracy, Precision Rate, Recall Rate, Sensitivity, Specificity, F-Measure and Error Rate measures are employed. It is shown in equation 4,5,6 and 7 correspondingly. Ideally, a excellent feature retrieval scheme is accepted to have a high Accuracy, Precision Rate, Recall Rate, Sensitivity, Specificity, F-Measure value. Table 1 lists the performance analysis Ontology Generation.

Table 1: Analysis of Detection Accuracy, Precision Rate, Recall Rate, Sensitivity, Specificity, F-Measure and Error Rate

Classifier Metrics	Acc	PRE	REC	SEN	SPE	FMES	ERR
KBayes	96.34	94.98	97.64	97.4	96.139	97.329	3.66
KTree	98.772	97.871	98.972	99	97.263	98.9283	1.228

As observed from Table 1, the Accuracy, Precision Rate, Recall Rate, Sensitivity, Specificity, F-Measure of the KTree in range 97-98, which is superior to KBayes method. So the KTree classifier is best for ontology creation. Fig.8 depicted the Detection Accuracy, Precision Rate, Recall Rate, Sensitivity, Specificity, F-Measure and Error Rate measures of classifier approaches.



As observed from above figure 4, the Accuracy, Precision Rate, Recall Rate, Sensitivity, Specificity, F-Measure of the KTree in range 97-98, which is superior than KBayes method. So the KTree classifier is best for ontology creation.

Experiment No 3: Analysis of Relationship Graph Generation Approaches

In this experiment, this work will evaluate the contribution of each Relationship Graph Generation approaches which are employed in the work. To assess the efficiency of this Relationship Graph Generation technique, the PPV, DSC, JI and SEN measures are employed. Ideally, a good Relationship Graph Generation scheme is expected to have a high PPV, DSC, JI and SEN value. Table 2 lists the PPV, DSC, JI and SEN measures of Relationship Graph Generation approaches.

Table 2: Analysis of PPV, DSC, JI and SEN of Relationship Graph Generation Approaches

DataSet				
Relationship Graph Generation Approaches				
Metrics	DSC	PPV	JI	SEN
Semantic	84.648	91.728	95.728	94.328
Thematic	96.458	95.098	97.758	97.518
Semantic + Thematic	97.092	96.923	97.692	98.59

As observed from Table 2, the PPV, DSC, JI and SEN of the Semantic + Thematic in range 96-97, which is superior than the other Relationship Graph Generation scheme. So the Semantic + Thematic is best for the Relationship Graph Generation scheme.

Fig.9 depicted the PPV, DSC, JI and SEN measures of Relationship Graph Generation approaches.

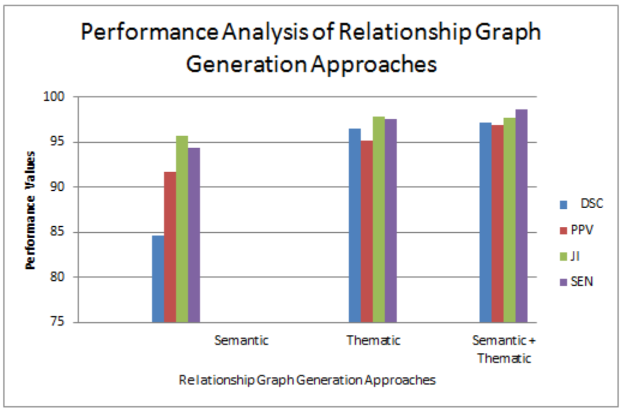


Fig. 5 Analysis of PPV, DSC, JI and SEN of BRATS 2015 for Relationship Graph Generation Approaches

As observed from Fig.5, the PPV, DSC, JI and SEN of the Semantic + Thematic in range 96-98, which is superior than that of the other Relationship Graph Generation scheme. So the Semantic + Thematic is best for the Relationship Graph Generation scheme.

CONCLUSION

Ontology helps us in making the process of acquiring and extracting knowledge in a much easier way. Existing methods of extracting existing information and templates do not allow us to retrieve information properly, so different methods have been developed to solve problems in different fields. In this paper, we discuss a new method to obtain optimal results for selected problems in multidisciplinary fields using effective theoretical theory, data extraction and neural networks. It shows better performance than other methods and acts as a new multi-domain framework. The established methodology can illustrate its importance to different ideas and relationships. The restriction on retrieval of keywords or terms was abolished by applying a query conversion technique that forms the internal query value pair. It comments and lists previous user queries, reduces computation time and costs, and produces improved productivity.

REFERENCES

1. Uschold, M., Gruninger, M.: Ontologies: principles, methods and applications. Knowl. Eng. Rev. 11(02), 93–136 (1996)
2. Grüninger, M., Fox, M.S.: Methodology for the design and evaluation of ontologies (1995) 3. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering (1997)
4. Noy, N.F., McGuinness, D.L.: Ontology development 101: a guide to creating your first ontology (2001)
5. Rinaldi, A.M.: A content-based approach for document representation and retrieval. In: Proceedings of the Eighth ACM Symposium on Document Engineering, pp. 106–109. ACM (2008)
6. Baykan, E., Henzinger, M., Marian, L., et al.: A comprehensive study of features and algorithms for URL-based topic classification. ACM Trans. Web (TWEB) 5(3), 15 (2011)
7. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: AAAI, vol. 90, pp. 223–228 (1992)
8. McCallum, A., Nigam, K.: A comparison of event models for naïve Bayes text classification. In: AAAI 1998 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48 (1998)
9. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49. ACM (1999)
10. Godbole, S., Sarawagi, S., Chakrabarti, S.: Scaling multi-class support vector machines using inter-class confusion. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 513–518. ACM (2002)
11. Lam, S.L.Y., Lee, D.L.: Feature reduction for neural network based text categorization. In: Proceedings of the 6th International Conference on Database Systems for Advanced Applications, pp. 195–202. IEEE (1999)
12. Ruiz, M.E., Srinivasan, P.: Hierarchical neural networks for text categorization (poster abstract). In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 281–282. ACM (1999).
13. Dunn (1973) J. C. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57
14. Navigli, R., Velardi, P., Gangemi, A., “Ontology Learning and Its Application to Automated Terminology
15. Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR). 2002;34(1):1–47.
16. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119; 2013.
17. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the international conference on learning representations (ICLR); 2013.
20. Bengio Y, Lee H. Editorial introduction to the neural networks special issue on deep learning of representations. Neural Netw. 2015;64:1-3. https://doi.org/10. 1016/j.neunet.2014.12.006. Epub 2014 Dec 15. (PMID:25595998)
21. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev. 1997;104(2):211.
22. Lund K, Burgess C. Producing high-dimensional seman-tic spaces from lexical co-occurrence. Behav Res Methods Instrum Comput. 1996;28(2):203–8.
23. Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis. In proc. of the cog-nitive science society (Vol. 1036). Erlbaum: Mahwah, NJ; 2000.
24. Hofmann T. Probabilistic latent semantic indexing. In: Proc. of ACM SIGIR conference on research and development in in-formation retrieval. ACM. Pp. 50–57; 1999.
25. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
26. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform. 2009;42(2):390–405.
27. Neelakantan A, Shankar J, Passos A, McCallum A. Efficient non-parametric estimation of multiple embeddings per word in vector space. EMNLP. 2014; 2014:1059–69.
28. Hu B, Tang B, Chen Q, Kang L. A novel word em-bedding learning model using the dissociation between nouns and verbs. Neurocomputing. 2016; 171:1108–17.
29. Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics. 2015;3:211–25.
30. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text pro-cessing. In: Proc. of languages in biology and medicine; 2013.
31. Merck Manuals, https://www.msdmanuals.com/en-gb/. Accessed 2 Aug 2017.
32. Medscape, http://www.medscape.com/. Accessed 2 Aug 2017.
33. Wikipedia, http://www.wikipedia.org/. Accessed 2 Aug 2017.
34. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform. 2007; 40(3):288–99.
35. Hill, F., Reichart, R. and Korhonen, A., . Simlex-999: evaluating semantic models with (genuine) similarity estimation. Computational Linguistics 2016.
36. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. In: AMIA annual symposium proceedings (Vol. 2010, p. 572): American Medical Informatics Association; 2010.
37. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence; 1995. p. 448–53.
38. Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform. 2004;37(2):77–85.
39. Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, Wolfman G, Ruppin E. Placing search in context: the concept revisited. ACM Trans Inf Syst. 2002; 20(1):116–31.
40. Bruni E, Tran NK, Baroni M. Multimodal Distributional Semantics. J Artif Intell Res. 2014;49(2014):1–47.
41. Muneeb TH, Sahu SK, Anand A. Evaluating distributed word representations for capturing semantics of biomedical concepts: Proceedings of ACL-IJCNLP; 2015. p. 158.
42. Chiu B, Crichton G, Korhonen A, Pyysalo S. How to train good word embeddings for biomedical NLP. ACL. 2016;2016:166.
43. BioASQ challenge, http://bioasq.org/. Accessed 2 August 2017.
44. Turney PD. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. Freiburg, Germany: Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001); 2001. p. 491–502.
57. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Medinfo. 2004;11(Pt 1):268–72.
58. Liu K, Peng S, Wu J, Zhai C, Mamitsuka H, Zhu S. MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. Bioinformatics. 2015;31(12):i339–47.
59. Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics. 2016;32(12):i70–9.