ONTOLOGY ORIENTED CONCEPT BASED CLUSTERING

Anbarasi .M.S[1], Iswarya.V [2], Sindhuja.M [3] , Yogabindiya.S [4]

[1,2,3,4] *Department of Information Technology, Pondicherry Engineering College, Puducherry, India*

Abstract

Worldwide health centre scientists, physicians and other patients are accessing, analyzing, integrating and storing massive amounts of digital medical data in different database. The potential for retrieval of information is vast and daunting. The objective of our approach is to differentiate relevant information from irrelevant through user friendly and efficient search algorithms. The traditional solution employs keyword based search without the semantic consideration. So the keyword retrieval may return inaccurate and incomplete results. In order to overcome the problem of information retrieval from this huge amount of database, there is a need for concept based clustering method in ontology. In the proposed method, WorldNet is integrated in order to match the synonyms for the identified keywords so as to obtain the accurate information and it presents the concept based clustering developed using k-means algorithm in accordance with the principles of ontology so that the importance of words of a cluster can be identified.

Keywords: *Ontology, Concept based clustering, K-means algorithm and information retrieval.*

--***--

1. INTRODUCTION

The steady and tremendous progress of computer hardware technology has led to large supplies of powerful and affordable computers, data collection equipments and storage media. The technology provides a great boost to database and makes a huge number of databases available for information retrieval and data analysis. Due to increasing and overwhelming amount of information available, there is a need for tools to automate the information search. Ontologies help to disambiguate the information and it helps automatic information processing which provides standard concepts which relate to specific concepts and terms and therefore eliminate the irrelevant information. It contains information about the relationships between the concepts and it is used to express the semantics of the terms.

Keywords contain the most important information which emphasizes the entire material. It is considered as the core operation for processing any text material. Keywords play a crucial role in extracting the correct information as per user requirements[3]. Every year thousands of books and papers are published. It is very hard to manually organise and analyze it. So there is a good information extraction which provides the actual contents of a given material. As such identify and extract the concepts, ontologies can be integrated into the process of clustering as a background knowledge.

The special requirements of good clustering algorithm are: 1) It should better preserve the relationships between words like synonyms since there are different words of same meaning. 2)Associating a meaningful label to each final clusters is essential.In our proposed system, the user will be able to find from what type of disease they are suffering from , by providing symptoms as their inputs and also theywill be able to find what type treatment they have to undergo for their disease. Our proposed system not only effective keywords are necessary as they express the importance of the entire material. With the identified keyword the synonyms are found so that the retrieval of information will be very accurate. Since different users view the same thing in different perspective

Clustering is one of the imperative techniques for organising data in an unsupervised manner. It is particularly used in handling large data. It is the process of grouping the similar items together according to some distant measures. It groups all the data so that the data within the group are more similar than the ones in other group [6,9]. It makes relevant data tend to more closely related to each other than the non relevant data. It is a precise operation used in variety of information needs

Vector space is one of the most commonly used representation schemes [7]. The frequency of every single term is recorded as a vector. The bag of words representation used for clustering methods is often unsatisfactory as it ignores relationships such as synonym between the important terms that do not co-occur literally. In order to overcome the lack of semantic consideration, the concept based text clustering method is proposed. In order

considers the keywords but also the concepts (Different words which shares the same meaning is called synonyms. Set of these different words that have the same meaning is called concept)[6]. So our proposed system works well by matching the synonyms for the given input and makes the search process effective and for efficient access.

2.PROPOSED METHOD

Proposed system considers not only the keywords but also the concepts based on the background domain knowledge. Proposed system is divided into following modules Text pre-processing, Ontology creation and concept based clustering.

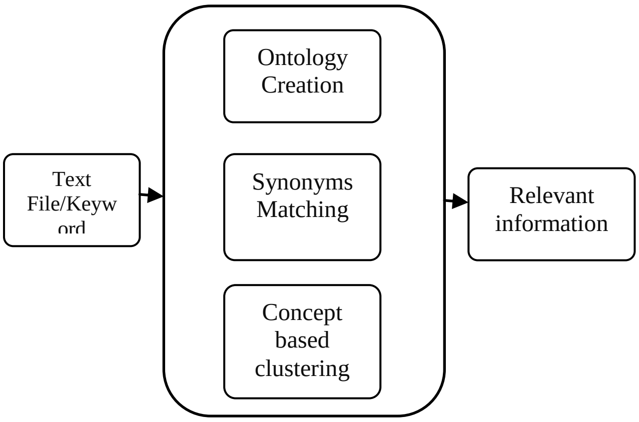


Fig-1:OOCBC High Level System Architecture

In the proposed system, Text file and keywords are given as the input. The keywords are searched in the ontology for its existence. If it exists, the corresponding parent, grandparents, children, and its synonyms will be captured. Based on the keyword, the whole path is retrieved from the ontology.

2.1 Pre-processing Phase

Patient health report collection is the initial stage for this phase. The textual information is stored in many kinds of machine readable form, such as PDF, DOC, HTML, and XML and so on. After the health reports are collected it is transformed into TXT format and maintained in the text files. The system transform the term related into concept represented one. Mainly, punctuation and special characters are removed. Finally keywords are extracted in this phase for further processing[1,2] .

2.2 Medical Ontology Creation

An Ontology is a representation vocabulary, often specialized to some domain or subject matter. It is a representation of a set of concepts within a domain and the relationships between those concepts [4,5].
Its main components are:

- Classes represents concepts which are taken in broad sense.
- Attributes represents properties of each concept.
- Relations represents the type of association between concepts of domain.

Here in our system we represent ontology for medical field. The medical ontology depicts the semantics of the terms used in the medical field. Medical ontologies are valuable and effective methods of representing medical knowledge. In the process of medical diagnosis disease has several symptoms associated with it. Organising diseases in an ontology hierarchy is extremely useful as it forms a pathological classification of diseases for use in medical systems. So here we create ontology for diseases and symptoms. We have identified some diseases and we also have keenly analysed about the symptoms related to each and every disease. As this is related to medical field, the query must be perfectly identified and should retrieve a correct result for it. In any case, a person who constructs the ontology needs to have some knowledge and prior experience in ontology construction and some knowledge of the domain. Usually, domain specialists are discussed to explain the meaning of domain-specific concepts.

Here there is a need for an effective way to store and retrieve knowledge related to human diseases. In this direction ontologies play a crucial role in defining standardized concepts. Beyond defining standards, medical ontologies are much more than biomedical vocabularies. They arrange concepts into ISA and sibling hierarchies, which effectively relate these concepts in a structural way that provides valuable inferences upon retrieval. Ontology hierarchies are valuable methods of knowledge representation. The most common case in building ontology is to groundwork the ontology vocabulary on relevant to medical guidelines. This means that all the relevant data from the guidelines has to be represented in a methodical way using a clergy of concepts and relations. Other sources of medical knowledge include medical articles, other medical ontologies or terminologies and most importantly, proficient knowledge .

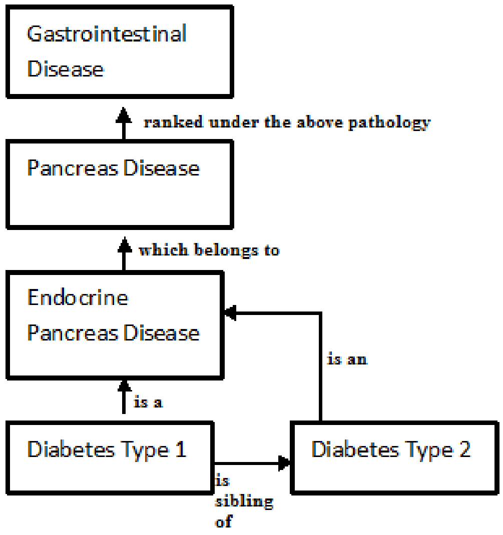
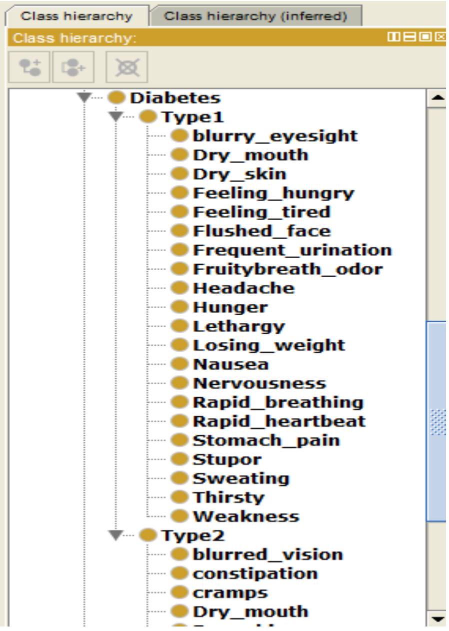


Fig-2:Ontology Hierarchy with ISA and sibling hierarchies

2.3.Medical Ontology Design

There are two standard approaches in designing ontology. First one is that tiny particles of the ontology are built first and then later joined to form the ontology using higher-level abstract classes. This is the bottom-up approach that is not implemented often in medical applications. The other way is to generically design the upper classes and then develop small parts of the hierarchy, so called top-down approach. This is used for large medical ontologies as well as taxonomies. It is advisable to start the process by creating classes first, then add properties or slots and finally conclude with individuals [4,5].

Ontology is made up of classes, properties or slots, relationships between classes and individuals. Individuals are instances or elements of the particular domain. Classes are composition or groups of individuals. Properties or Slots are the connection between classes or individuals. An example of a medical ontology class is «Disease». It is the super-class of all the Disease types. All the diseases comes under the class «Disease». The other types of diseases, example, Diabetes comes under the class «Disease». A class can be more formal or general (upper class) or more particular (subclass). For example, A specific class of «Disease» is «Diabetes». The diabetes types comes under the class «Diabetes». For example Diabetes type 1 and Diabetes type 2. All the symptoms of diabetes comes under the class Diabetes type 1 and type 2 which comes under the class «Diabetes». In our case the «Disease» class acts as the most general class. There is no strict and apparent way in which medical knowledge must be represented. After creating the class hierarchy, Object properties are created. The object properties for our medical domain are hasSymptom and isTreatedBy is specified .



- Diabetes
  - Type1
    - blurry_eyesight
    - Dry_mouth
    - Dry_skin
    - Feeling_hungry
    - Feeling_tired
    - Flushed_face
    - Frequent_urination
    - Fruitybreath_odor
    - Headache
    - Hunger
    - Lethargy
    - Losing_weight
    - Nausea
    - Nervousness
    - Rapid_breathing
    - Rapid_heartbeat
    - Stomach_pain
    - Stupor
    - Sweating
    - Thirsty
    - Weakness
  - Type2
    - blurred_vision
    - constipation
    - cramps
    - Dry_mouth

Fig-3: Class Hierarchy

Table-1: Hypertension Disease Class representing symptoms of hypertension

Disease Class Name	Object Property	Symptoms
Hypertension	has_symptom	Drowsiness
Hypertension	has_symptom	Blurred Vision
Hypertension	has_symptom	Tinnitus
Hypertension	has_symptom	Nosebleed
Hypertension	has_symptom	headache
Hypertension	has_symptom	Nausea
Hypertension	has_symptom	Flushing
Hypertension	has_symptom	Palpitation
Hypertension	has_symptom	Frequent urination
Hypertension	has_symptom	Urgency of urination
Hypertension	has_symptom	nocturia
Hypertension	has_symptom	Dizziness
Hypertension	has_symptom	Breathing Difficulty
Hypertension	has_symptom	Fatigue

Here the object property is created for the disease hypertension with relevant symptoms. Likewise the class hierarchy is created for different types of disease and symptoms relevant to all the diseases.

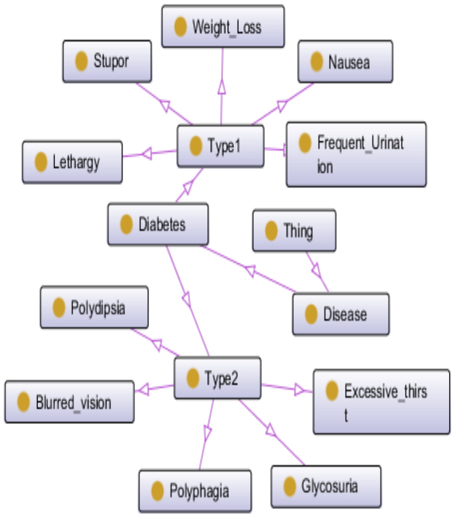


Fig-4: Class hierarchy for diabetes

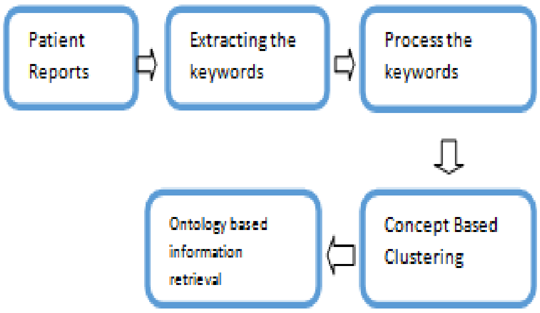


Fig-5: Flow Diagram

Figure 5 represents the flow diagram of our system. First patient health report is fed into the system and the keywords are extracted from the report and it is processed so to obtain the result based on concept based clustering with the principles of ontology. Here the information is retrieved based on ontology.

3.CONCEPT BASED CLUSTERING

3.1 Cosine Similarity

The similarity between the different diseases are measured based on the symptoms. A vector based representative metric is the cosine measure is used which is defined by the cosine of the angle between the two vectors:

$$Sim(V1, V2) = [V1.V2]/|V1||V2| \text{--------}(1)$$

where V1 and V2 be term frequency vectors to be compared for its similarity[7,8]. Initially the cosine similarity values are taken as centroid for the clusters. Each time centroid value will change thus the highest similarity value shows that the symptoms are more related to the cluster.

3.2.K-means Clustering

Concept based clustering is nothing but grouping the symptoms based upon the concept (Different words which shares the same meaning is called synonyms. Set of these different words that have the same meaning is called concept)[6]. Here clustering process is done based upon the symptoms with the identified synonyms.

The k-Means algorithm implemented as a simple procedure that initially selects k random centroids, assigns each example to the cluster whose centroid is closest, and then calculates a new centroid for each cluster. Examples are reassigned to clusters and new centroids are re-calculated repeatedly until there is no change in clusters[1,9].

4.EXPERIMENTAL RESULTS

The patient reports are collected and it is preprocessed in order to obtain the symptoms from the report. The extracted keywords are dynamically mapped with the online WordNet dictionary. The patient health reports consists of patient information, data used for analysis, vitals and their symptoms. In our system, we have to preprocess the patient report in order to extract the keywords. The keywords are nothing but the symptoms. We need to extract the symptoms from the patient report in order to find the disease and to dynamically map the synonyms for each of the symptoms extracted from the patient report. Only the symptoms are needed other details are not needed. Different users may view the same thing in different perspective. So user will not be aware of all the words or all the synonyms with respect to a particular word. Everyone thinks in a very different manner. For example, user may provide the input as tiredness, weariness, tire and all the possible meanings instead of the word fatigue. The result obtained for the user query is represented in tabular form.

Table-2: Symptoms acquired for the relevant disease

DISEASES	SYMPTOMS(represented in numbers)
Diabetes Type-1	2
Heart Stroke	1
Hypertension	12

The numbers in table1 is nothing but the number of symptoms relevant to all the diseases. Here the maximum number of symptoms corresponds to Hypertension where the symptoms are clustered based on the synonyms. So it will be identified that hypertension is the relevant disease for the user's query and the diagnosis procedure for the relevant disease will be shown to them. Finally the system is measured for finding the efficiency of the proposed system. To evaluate the accuracy of our clustering algorithm, we use Recall and Precision performance metrics. The value of precision and recall can be calculated as:

$$Precision \quad = x/x+y \quad \text{-------------}(2)$$

$$Recall \quad =x/x+z \quad \text{-------------}(3)$$

where x is the number of total true positives, i.e. the total number of items clustered together in predefined class and that are indeed found together by the clustering algorithm. y is the total number of false positives, i.e. the number of items not supposed to be found together but are clustered together and z is the number of total false negatives, i.e. the number of items which are expected to be found together but not clustered together by the clustering algorithm. The result of the experiment based on accuracy is shown in the graph where it is matched with the highest similarity.

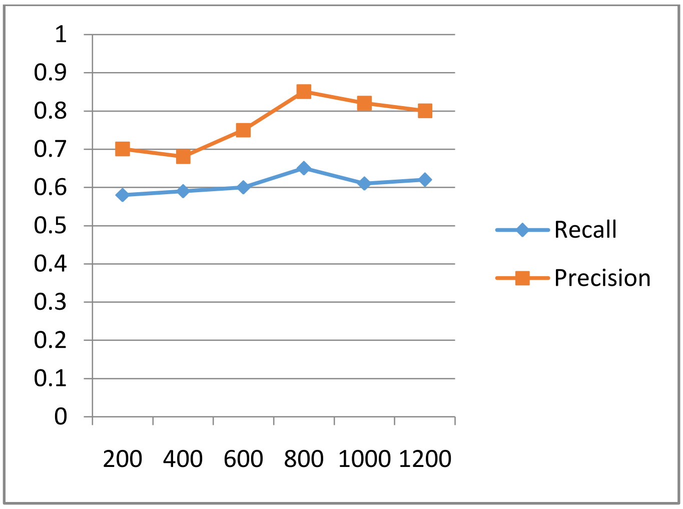


Fig-6: Accuracy of the proposed system

5.CONCLUSION AND FUTURE WORK

The World Wide Web grows and changes rapidly and many researchers are stepping into the era of ontology. There is a highly diverse group of plain text available in free form. The paper articulates the unique requirements of text clustering with the support of specific domain ontology. With the use of ontology, the proposed system is able to categorize the items on the basis of the concept level. The dimensionality of the data gets condensed in this proposed concept based clustering model. It is an efficient method for retrieving the information even from very large databases. Concept based clustering exhibits a better performance than the traditional term based clustering. This system can further be enhanced for video and image retrieval even from huge repository.

REFERENCES

[1] Hmway Hmway Tar , Thi Thi Soe Nyunt,” Ontology-Based Concept Weighting for Text Documents” International Conference on Information Communication and Management.

[2] HmwayHmwayTar ,ThiThiSoeNyaunt, “ Enhancing Traditional Text Documents Clustering based on Ontology”, International Journal of Computer Applications.

[3] Jasmeen Kaur 1, Vishal Gupta, “Effective Approaches For Extraction Of Keywords”, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6.

[4] Matthew Horridge, ―A Practical Guide To Building OWL Ontologies Using Protege 4‖ Edition 1.3. The University of Manchester.

[5] protege.stanford.edu.

[6] Rekha Baghel, Dr. Renu Dhir, “A Frequent Concept Based Clustering Algorithm”, International Journal of Computer Applications.

[7] S.Logeswari, Dr.K.Premalatha, “Biomedical Document Clustering Using Ontology based Concept Weight”, International Conference on Computer Communication and Informatics.

[8] Shehata, Fakhri and Mohamed S.Kamel, “An Efficient Concept Based Mining Model for Enhancing Text Clustering”, journal of IEEE Transactions on Knowledge and Data Engineering.

[9] Steinbach M, Karypis G and kumar V, ” A comparison of document clustering techniques”, KDD Workshop on text Mining.