CONCEPT EXTRACTION FOR MEDICAL DOCUMENTS USING ONTOOGY

Vaibhav Gangarde\(^1\), Amol Bhore\(^2\), Rahul Kamble\(^3\), Kunal Shah\(^4\)

\(^1\) Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, Savitribai Phule Pune University,
\(^2\) Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, Savitribai Phule Pune University,
\(^3\) Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, Savitribai Phule Pune University,
\(^4\) Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, Savitribai Phule Pune University,

Abstract —: In current search engine which performs searching operation based on keywords which results in enormous amount of data available to the user, from which user can not figure out the essential and most important information. Now a days in the biomedical domain large amount of text documents are unstructured information is available in digital text form. Text Mining is the technique to find interesting and useful information from unstructured text. Extension to text mining is Semantic text mining and its to discover the hidden information from unstructured text and making relationships of the terms occurring in them in medical documents. Most of information is stored in the text format, so in this paper we will focus on the role of ontology for semantic text mining. Specifically, we have presented a model for extracting concepts from text documents using linguistic ontology in the domain of medical.

Keywords — Text-Mining, Semantic-Based Approach, NLP, Ontology, Medical-Documents.

I. INTRODUCTION

Now days the large amount of unstructured information is available in digital text form. In biomedical domain, these information available in medical documents for organizing these information text mining plays vital role in text documents which has two approaches. One is traditional where keyword and semantic based extraction used. Limitation of traditional text mining is overcome by semantic text mining and allows for making relations between words. Use of ontology becomes important in medical domain because it can provide semantic patterns from text documents. The main goal of the paper is focus on role of ontology for mining text in biomedical domain.

Figure 1. Extracting Concept cluster from biomedical text

The paper is divided into different sections which are as following. Section 2 discusses semantic text mining and talks about ontology for extracting concept. In section 3 we are identifying and presented algorithm to find concept cluster. With the help of example we explain concept clusters for a document of medical domain. Last section with applications of text mining and further section devoted to future work in the medical domain.
II. ONTOLOGY BASED APPROACH IN SEMANTIC TEXT MINING

The term ontology refers to formal description of knowledge in the field of data mining techniques and application. It is formally describes list of terms to represent important concepts such as classes of objects and relationships between them. To discover common meanings and the solution to collect information at common place called ontology.

A. Concept based Information Retrieval for estimating Concepts in Ontology Model

Using ontological approach, the documents are represented by group of concept cluster. In the document, a number of concept clusters is represented as a weight of terms in a documents and it makes a lexical chain in the documents.

B. Representation of Concepts Cluster

In the proposed model, a concept cluster represented by sequential process. First to estimate the semantically important terms within a document, the representation of concepts can be identified through a graph. Each node of the graph represents a unique term. Each term has an element for giving frequency of the term. Basically, the frequency shows identity relationships between the terms. Further this method facilitates calculation of weight for each term indicating importance of each term. To be more specific, concept based model is represented as follows:

**Definition:** (Concepts Cluster)

Concept cluster look like a lexical chains which represents the concept or subject of a document.

Let Nd = \{n1, n2, ..., ni\} is the set of nouns in the documents,

Let us T= \{(t1,f1), (t2,f2), ..., (tm,fm)\}

Here ti = term and fi = frequency

The set of lexical Relations:

\[ R_l = \{\text{identity, synonyms, hypernym-hyponym, meronym-holonym}\} \]

Let suppose C= \{C1, C2, ..., C m\} are concept cluster for each documents,

Let \( M_f(R_k, tg) \) is the sum of frequency of all the terms. \( tg \in T \) from relation \( R_k \in R_l \)

Let \( W(R_k) \) is the weight for relation \( R_k \).

ScTerm(tg) is the score term tg.

The following equation is defined to find score of the concept cluster:

\[
\delta_{ScTerm}(tg) = (f_{tg} - 1) \times W(R_1) + f_{tg} \times \sum_{k=2}^{m} M_f(R_k, tg) \times W(R_k)
\]

Where 1 ≤ g ≤ m

Relations between the terms = \{R1 = identity; R2 = synonym; R3 = hypernym-hyponym R4 = meronym-holonym\}

III. IDENTIFYING CONCEPT CLUSTERS

In the proposed model, a concept cluster can be constructed by ontology and first to estimate the semantically important terms within a document, the representation of concepts can be identified through a graph. Every concept clusters joined by the component of the graph. We simplified algorithm by concept based IR.

A. Proposed Algorithm

**Input:** Medical Documents

**Output:** Extract Concept Cluster and Ranked list of medical Documents
1. Let’s take Medical text documents
2. W (Rk) is a weight for (Rk ) semantic relation.
3. For every document D, find all nouns in the documents D which represent by T term, use (POS) tagger [16].
4. Repeat loop ( )
   4.1 (a)determining a concept cluster for document D, which shows a graph C (V, E).
   4.2 Let x ∈ T, x term as vertex V for concept C.
   4.3 find frequency fx of x in the Document D
   4.4 find T = T - x
   4.5 Let y ∈ T, y term belongs to any term x ∈ V and relation Rk ..
(b) The frequency fy should assigned for term y in D
(c) Allocate T = T-y. Until T= 0
5. Calculate the weight for the terms in the D by find in concept cluster.
   (a) Obtain concept cluster for the term x.
   (b) Redo all steps within concept cluster.
   (c) Describe the relations for node x, and connect it to edges x.
   (d) Find out the weight of term x by given equation (1).
6. Each concept clusters identified and represented by document. After that weight can be appointed to concept based model.
7. Documents can be compared by the techniques accordingly.

B. Algorithm Based Example
Simple example for extracting concept cluster is as follows:

“There are different types of diseases. It is caused through the viruses such as cold is a common, influenza, dengue fever and AIDS. Most Diseases are typhoid fever, cholera, tuberculosis and anthrax are caused by bacteria. Protozoan microbe is familiar diseases, such as malaria and kalaazar.”

Above example of medical document, first we use nouns to extract the concept cluster. Figure 2 shows the lexical chain of document “Diseases”. In our proposed an algorithm defined (Step 5) concept clusters may extracted in the document “Diseases” are first finding a concept cluster, we joined related terms by lexical chains, after that we assign the weight to concept cluster and it depends on the bases of relations of one word with other word. Each node of the graph represents a unique term. Each term has an element for giving frequency of the term. Basically, the frequency shows identity relationships between the terms. Finding a concept cluster, we joined related terms by lexical chains, after that we assign the weight to concept cluster and it depends on the bases of relations of one word with other word. In this document one can observe that we emphasized on concept “Diseases”. Concept clusters that are being generated after applying this algorithm also depicts that the term “Diseases” is the most prominent one in this document, having maximum weight. In this document, there are only few terms which have no relation and those terms are not mentioned in this figure.

![Figure 2. Cluster of sample Lexical chains of medical text “Diseases”](image-url)
IV. APPLICATIONS OF TEXT MINING

Text mining has its own applications in various areas of research including NLP, Biomedical sciences, telecommunications, Biotechnology etc. In this paper our focus is on medical domain. Classification is done based on similarities between data items. Defining similarity itself is very crucial to clustering. Traditional similarity approaches are based on finding overlap between the keywords within the documents. Again these similarity measures are unable to capture semantic similarity aspects. Ontology can provide various relationships which can be used to find semantic similarity between the Text mining, also referred to as text data mining, roughly equivalent to text analytics refers to the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output.

Some examples of practical applications of text mining techniques include:

1. Spam filtering
2. Creating suggestion and recommendations (like Amazon)
3. Monitoring public opinions (for example in blogs or review sites)
4. Customer service, email support
5. Automatic labeling of documents in business libraries
6. Measuring customer preferences by analyzing qualitative interviews
7. Fraud detection by investigating notification of claims
8. Fighting cyber bullying or cybercrime in IM and IRC chat
9. Text clustering
10. Web Searching

V. CONCLUSION AND FUTURE WORK

In this paper, we discussed the methods of analysis and integration of various methods of semantic text mining in biomedical literature and articles. Semantic text mining differs from traditional text mining. It is able to capture the meaning and relationships between the words present in the document. By the help of ontology it becomes easier to handle the complex semantic relationships among word/terms in the text. By tracking the concept hierarchy ontologies make the text mining approaches recognize the relationship between two terms for measuring the semantic similarity between documents. Semantic text mining techniques can prove to be very beneficial in medical domain also. In this paper, we are using ontologies for concept extraction and we have provided an algorithm to find concept based clusters. After that, we assigned a semantic weight to each terms for each document. The implementation has been shown using ontology which is a linguistic ontology for English language. It has been shown that even this general ontology can be useful for finding semantics of documents in medical domain. We have used POS tagger for finding the nouns and also use RapidMiner for the text processing techniques of text mining. The use of medical based ontologies can further improve the results of our approach. When compared to that of the other model and also MeSH ontology is used for analyzing the concepts in a better way.
VI. REFERENCES


[8] Andreas Nürnberg, Gerhard Paaß, FraunhoferAiS. A Brief Survey of Text Mining


