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

A Semantic Model for Concept Based Clustering

S.Saranya[1], S.Logeswari[2]

PG Scholar, Dept. of CSE, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India[1]

Associate Professor, Dept. of CSE, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India[2]

ABSTRACT— Text mining is the data extraction from textual databases or documents. Each word in the document is a dimension giving a structure to the data and reducing the dimensions. In text mining techniques the basic measures like term frequency of a term is computed to work out the weight of the term in the document. Although with the statistical analysis, the original meaning of the term may not take the precise meaning of the term. The proposed system relies on concept based model. In this concept based approach, the concepts are extracted from the documents, and a semantic based weight is computed for effective indexing and clustering. It uses MeSH ontology for concept extraction and concept weight calculation based on the identity and synonymy relations. K-means algorithm is used for clustering the documents depending on the semantic similarity. Experiments are conducted and the results are analyzed.

KEYWORDS— MeSH Ontology, Concept based model, Document clustering, Concept extraction, semantic similarity.

I. INTRODUCTION

As usage of internet becomes more and more analysis of text document also becomes more. Text mining is a defined as the extraction of structure content from an unstructured content. It is also known as text data mining. High importance in text mining usually deals with various combinations of significance, originality, and interestingness. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them.

The internet to vital requirement for clustering method based on ontology. Ontology .A cluster is a collection of data objects that are similar to one another. A group of data objects can be treated together in concert group and so may be measured as a form of data compression. Clustering is also known as data segmentation in several applications as clustering partitions vast data sets into groups based on their similarity.

Text clustering is an involuntary document organization, topic mining and rapid information retrieval. It is closely associated with data clustering. Document clustering use of descriptors and descriptor mining. Descriptors are sets of terms that define the contents in the cluster. Document clustering is commonly measured to be a centralized method.

Many of the existing document clustering techniques uses the model to represent the terms of a document to work out the weight of the term in the document. Although with the analysis of statistical, the original semantics of the term may not take the precise meaning of the term.

In proposed system a concept based mining model along with the semantic smoothing model is developed and involves the concepts are extracted from the documents, and a semantic based weight is computed for effective indexing and clustering. The goal of this concept based model is to reduce general words and give importance to the core words so that the cluster quality can be improved effectively.

This paper is organized as follows. Section 2 describes related work on semantic smoothing and document clustering. Section 3 defines the concept based model combined with semantic smoothing. Section 4 involves the experimental results. Section 5 defines the conclusion.

II. RELATED WORK

Kamaljeet et al [1] presented a novel approach to context sensitive semantic smoothing by making use of an intermediate, semantic representation for sentences, called Semantically Relatable Sequences (SRS) where a sentence are a tuple of words shows in the semantic graph of the sentence as associated nodes describing dependency relations.

Hmway et al [2] presented a technique for clustering the text documents at the concept level based on the weights given to the concepts. The concepts are extracted from the documents based on the domain ontology. The importances of the words are indicated by their weights. Documents are classified depends on their weight which gives better accuracy and performance of the text document clustering method.

Khare et al [3] proposed a new concept-based mining model to improve the text clustering quality. The similarity between documents is calculated based on a new concept-based similarity measure. The anticipated similarity measure takes full gain of using the concept analysis ensures on the sentence based, document and with corpus levels in calculating the similarity between documents. By exploiting the semantic structure of the sentences in documents, a superior text clustering result is achieved.

Shehata et al [4] proposed mining model which consists of concept analysis depending on sentence, document and also based on corpus. The term which contributes to the sentence semantics is analyzed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. It efficiently finds significant matching concepts between documents, based to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure.

Xiaohua Zhou et al [5], proposed a novel context-sensitive semantic smoothing model involuntarily finds the multiword phrases in a document and that statistically records the phrases to the separate terms in the document. When estimating on the new model-based similarity measure it significantly improves the clustering quality over the usual vector cosine measure.

Yubao Liu et al [6], had proposed an improved semantic smoothing method which is suitable for both agglomerative and partitional clustering. It is based on the principle of TF*IDF schema. The outcome of the method shows that the method is more successful than the several other methods in terms of cluster quality.

JiarongCai et al [7], proposed an efficient solution for the improvement of document cluster quality. Many existing semantic smoothing model is not effective for partitional clustering to enhance the document clustering quality. The experimental result confirms that the semantic smoothing model is valuable than other methods with better cluster quality.

Xiaodan et al [8] presented a comparison based on the impact of the semantic similarity measures like path-based, information-content based and feature-based measures on medical document clustering. Medical Subject Headings (MeSH) is used as the domain reference and the documents are indexed based on it. Terms are assigned with dynamic weights based on the re-weighting method proposed by them.

Lei Zhang et al [9] had proposed an ontology-based clustering algorithm with feature weights to reflect the importance of different features. Feature weight in the ontology tree is calculated according to the feature's overall relevancy that shows that ontology-based clustering algorithm with feature weights for more accurate result.

III. BUILDING SEMANTIC SMOOTHING MODEL FOR CONCEPT

A. MeSH Ontology

MeSH ontology is abbreviated as Medical Subject Headings which consists of the controlled vocabulary and a MeSH Tree. It is published by National Library of Medicine. The controlled vocabulary includes the numerous different types of terms like descriptor, qualifiers, publication types, geographic, and entry terms. Descriptors and the entry terms are used in this process. Descriptor terms are the main concepts or main headings in the ontology. Entry terms are the synonyms or the interrelated terms for descriptors. MeSH descriptors are located in MeSH Tree and observed as a MeSH Concept Hierarchy. The MeSH tree contains 15 categories for example category A involves anatomic terms and each category is further partitioned into subcategories and in each subcategory the descriptors are hierarchically ordered from most general to most specific. Descriptors in general shown in more than one site in the tree which are represented in a graph rather than a tree.

B. Term Based Model

Term based model considers the frequency of the terms in the document. Term Frequency–Inverse Document Frequency (TF-IDF) defines the important of the term in the document in a group or corpus. It is frequently used as a weighting feature in information retrieval and in text mining. The TF-IDF enhances to the number of times a term emerges in the document. The term frequency for a term in a document is calculated by the formula (1)

$$\frac{\quad}{\quad} \qquad (1)$$

C. Concept based model

The concept based model considers concept weight for selecting the trait of the documents with the support of ontology. Concept based weighting scheme computes the importance of the underlying text by converting the documents into a bag of concepts. Document clustering mainly used for extraction of better document and text mining. The goal of concept based approach is to mine texts through the analysis of higher level characteristics, minimizing the vocabulary problem and the effort necessary to extract useful information.

The weight of the abstract is calculated by using the frequency of the term and their weight. The semantic weight of the concept for an abstract is calculated using the equation (2)

$$\frac{\quad}{\quad} \qquad (2)$$

where, $w(\text{concept}_i)$ is the weight of the particular concept, Freq_j is the frequency of the particular relation , Weight_j is the semantic weight assumed for the particular relation and N is the number of occurrences of all concepts in the abstract. The document with highest concept weight denotes the presence of the concept in the document. For example the path retrieved for the keyword merkel cell carcinoma from the disease cancer,

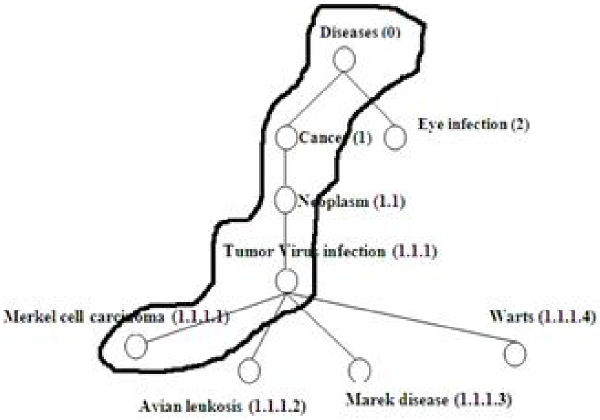

Figure 1 choosing the path for disease

D. Semantic Smoothing Model

In concept based approach with semantic smoothing model, the concept based weighting scheme is used to identify the importance of concepts along with the synonyms based on the concept hierarchy. Semantic smoothing model is the mixture of a simple language model and topic signature translation model.

Semantic smoothing model is the combination of simple language model and topic signature translation model. The translation coefficient (λ) used to direct the influence of two models in the combined model. The translation coefficient is optimized by exploiting the certain clustering quality with training data. The semantic smoothing model is calculated by using the following equation (3) which is the estimation of the semantic smoothing model for concept based approach.

(3)

$p_{bt}(\text{Concept} \mid d)$ denotes the semantic model for concept based approach. $p_b(\text{Concept} \mid d)$ is the simple language model of semantic smoothing. $p_t(\text{Concept} \mid d)$ is the topic signature model and is the translation coefficient. The first component called the simple language model of semantic smoothing determines whether a concept appear in a relevant document often or not. Simple language model is calculated using the equation (4)

(4)

$p_{ml}(\text{Concept} \mid d)$ is the probability of maximum likelihood that document belongs to the particular concept and the Freq_r represents the frequency of the term in the document and the Weight_r signifies the weight assigned to the concept relation and $p(\text{Concept} \mid \text{Corpus})$ is the probability of the entire corpus which belongs to that particular concept.

The second component of document model is the topic signature model. It is also called as multi phrase translation model. The topic signature model determines the probability of translating the phrase to the concept in the training process. Topic signature translation model is calculated by using the following equation (5)

(5)

$p_{ml}(t_k \mid d)$ is the maximum likelihood of the phrase presenting in the document and the $p(\text{Concept} \mid t_k)$ is the probability of translating the phrase into a specific concept after comparing it with the ontology.

IV EXPERIMENTAL RESULTS

The experiments are conducted using the abstract dataset collected from PubMed via MEDLINE. Documents that are collected in the domains based on four categories such as cancer, eye infection, viral disease and respiratory. In each category 100 documents are considered and totally four hundred documents are considered for the experiment. Initially Pre-processing (i.e.) tokenization and stopword elimination is performed, after pre-processing the dimensionality get reduced and terms that are contributing in the identification of the concept are extracted from the documents. The terms are mapped with the Mesh ontology and for its existence. Probability of each term in the document is identified by the term based model. For the concept based model, the concept weight of an individual document is calculated based on the frequency and the dynamic weight assignment of the terms in a document using the semantic relations that are derived from the domain ontology. Identity, synonymy, hyponymy and metonymy are the semantic relations used for the assignment of dynamic weight for the concept.

TABLE I. COMPARSION OF TERM , CONCEPT AND SMOOTHING MODEL

DOMAIN	DOCUMENT	QUERY	TERM	CONCEPT	SMOOTHING
			MODEL	MODEL	MODEL
Cancer	C1	Tumor	0.00735	0.084345	0.11580
	C21		0.00934	0.064766	0.04862
	C36		0.00724	0.065797	0.17682
Eye infection	E1	Conjunctivitis	0	0.06830	0.072624
	E21		0.038709	0.0965	0.13565
	E36		0.029354	0.05405	0.07688

V CONCLUSION

The concept based approach with semantic smoothing model achieves the better result when compared to that of the other model and also MeSH ontology is used for analyzing the concepts in a better way. When usage of concept approach with semantic smoothing model reduces the general words and give importance to the core words so that the cluster quality also improved effectively.

REFERENCES

[1] Kamaljeet S Verma and Pushpak Bhattacharyya, (2009) "Context-Sensitive Semantic Smoothing using Semantically Relatable Sequences", Proceedings of International Jont Conference on Artifical Intelligence (IJCAI'09), pp.1580-1585.

[2] Hmway Tar, Thi Soe Nyaunt, (2011)"Enhancing Traditional Text Documents Clustering based on Ontology", International Journal of Computer Applications (IJCA'11), Vol.33, no.10, pp. 38-42.

[3] Akhil khare, amol n. Jadhav (2011)"An efficient concept-based mining model for enhancing text clustering "International Journal of advanced engineering technology, vol.II, pp 196-201.

[4] Shehata, Fakhri and Mohamed S.Kamel,(2010) "An Efficient Concept Based Mining Model for Enhancing Text Clustering", Journal of IEEE Transactions on Knowledge and Data Engineering, Vol.22, pp. 1360-1371.

[5] Xiaohua Zhou, Xiaodan Zhang, Xiaohua Hu, (2007) "Semantic Smoothing of Document Models for Agglomerative Clustering ",Proceedings of the 20th international joint conference on Artificial intelligence (IJCAI), pp. 2922-2927.

[6] Yubao Liu, JiarongCai, Jian Yin, Zhilan Huang,(2007) "Document Clustering Based on Semantic Smoothing Approach", Proceedings of the 5th Atlantic Web Intelligence Conference (AWIC), Advances in Intelligent Web Mastering, Advances in Soft Computing Vol.43, pp.217-222.

[7] Jiarong Cai,Yubao Liu & Jian Yin,(2007) "An Improved Semantic Smoothing Model for Model-Based Document Clustering ", Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Vol.3, pp.670-675

[8] Xiaodan Zhang, Liping Jing, Xiaohua Hu, Michael Ng and Jiali Xia,(2008) " Medical Document Clustering Using Ontology – Based Term Similarity Measures", International Journal of Data Warehousing and Mining, Vol.4, no.1, pp. 62-73.

[9] Lei Zhang , Zhichao Wang, (2010)"Ontology-based clustering algorithm with feature weights", Journal of Computational Information Systems 6:9,2959-2966.

[10] Xiaodan, Z., Xiaohua Zhou, XiaohuaHu.,(2006)"Semantic Smoothing for Model-based Document Clustering."In:Proc. IEEE ICDM, pp.1193-1198.

[11] HmwayTar ,ThiThi Soe Nyaunt, (2011) " Enhancing Traditional Text Documents Clustering based on Ontology", International Journal of Computer Applications, vol.33, no.10, pp. 38-42.

[12] Zhou, X., Hu, X., Zhang, X., Lin, X., and Song, I.-Y.(2006) "Context-Sensitive Semantic Smoothing for the Language Modeling Approach to Genomic IR", Proceedings of ACMSIGIR, pp.170-177.