Enhanced Semantic Preserved Concept Based Mining Model for Enhancing Document Clustering

Resmi Ramachandran Pillai

Abstract: The project "Enhanced semantic preserved concept based mining model for enhancing document clustering " proposes the enhancement of data mining model for efficient informaion retreival . Concept based mining model is a challenging and a red hot field in the current scenario and has great importance in text categorization applications. A lot of research work has been done in this field but there is a need to categorize a collection of text documents into mutually exclusive categories by extracting the concepts or features using supervised learning paradigm and different classification algorithms. This project aims to Develop a concept based mining model for preserving the meaning of sentence using semantic net & synonym dictionary. The new concept definition can be expressed in the form of a triplet <subject, verb, object> .This triplet is the basic unit for the processing and preprocessing tasks. For increasing the performance, SVD (Singular Value Decomposition) is used.

Keywords: - SVD, Concept, Categories, algorithms, clustering.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectation.Semantic preserved Concept based mining model is used to avoid the problems of polysemy and synonymy in the text mining applications.It is a challenging issue to find accurate and relevant knwolegde in the text documents to help users to find what they actually want.The main advantage of term based mining is that it has highest computational performance compared to concept based mining.A lot of research work has been done in this field but there is a need to categorize a collection of text documents into mutually exclusive categories by extracting the concepts or features using supervised learning paradigm and different classification algorithms.

II. EXISTING SYSTEM

The existing system is based on keywords and its frequency.When we are submitting a querry it counts the frequency of words and seraches based on the frequency. The main disadvantages are it is manual one,costly,imposes waste of time for manual operations, lack of semantic consideration,inefficient term matching,do not consider synonyms and term dependencies,it does not provide partial matching and poor retreival performance.

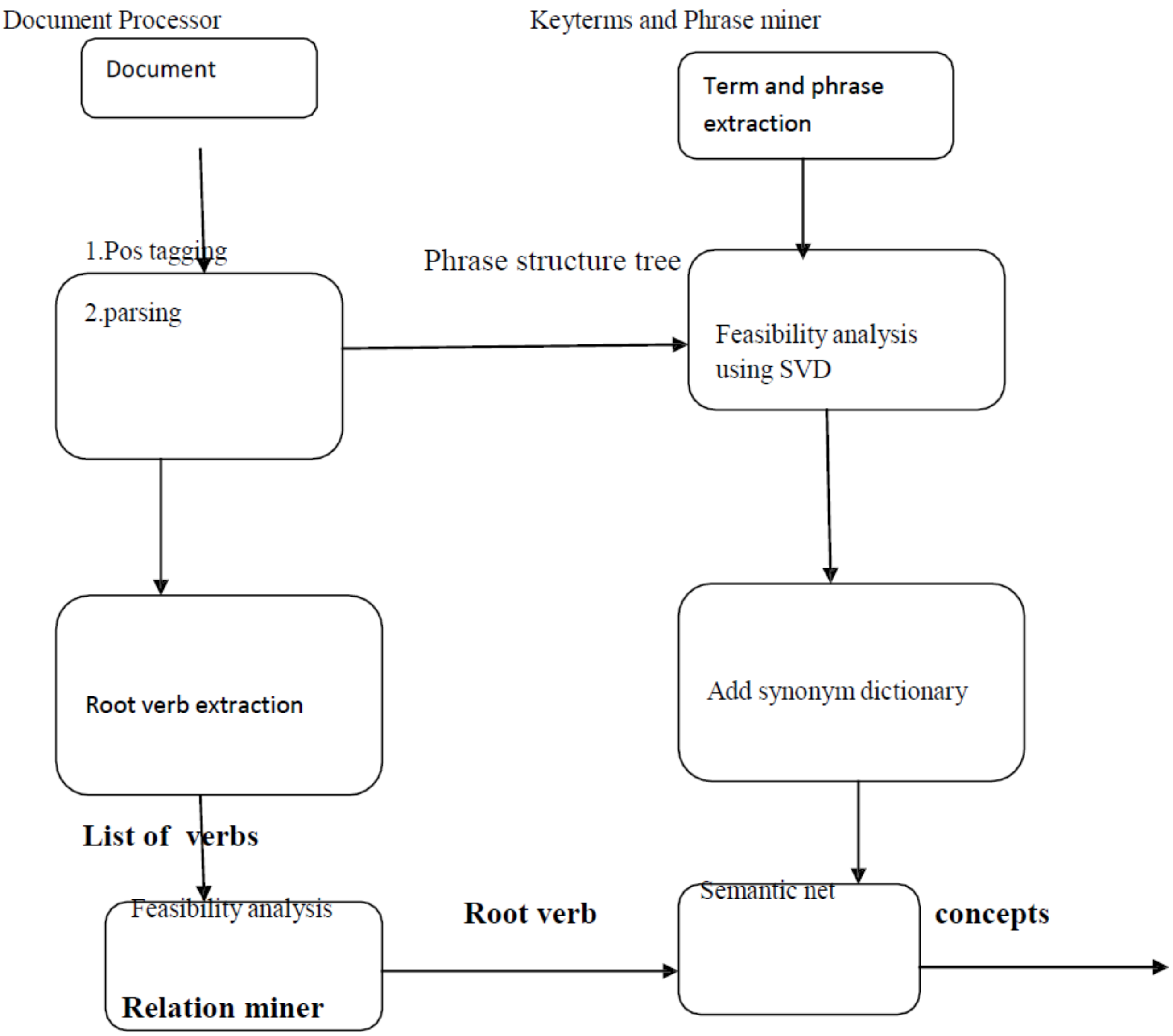
III. PROPOSED SYSTEM

In the proposed system the concepts of the passage is considered and clustered on the basis of the semanticsThe proposed model can efficiently find significant matching concepts between documents,according to the semantics of their sentences. The similarity between documents is calculated based on a new concept - based similarity measure.A raw text document is the input to the proposed model. Each document has well definedsentence boundaries. Each sentence in the document is labeled automatically based on the Prop Bank notations. The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based model on the sentence and document levels. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence..The main advantages are it can be used to create a platform that is capable of identifying & classifying medical care related information from patients,it consider the semantic meaning of the entered texts,efficient term matching,considers synonyms and term dependencies, provide partial matching and it is accurate method.

Manuscript Received on December, 2014

Resmi Ramachandran Pillai, Computer Science and Engineering,Anna University,Tamil Nadu, India.

IV. SOLUTION APPROACH



V. MODULES

1. Preprocessing and parsing
2. Concept Extraction
3. Concept Representation
4. Clustering
5. Updation of Database

VI. MODULE DESCRIPTIONS

6.1 preprocessing and parsing

In computer science, a preprocessor is a program that processes its input data to produce output that is used as input to another program. The output is said to be or may be promoted as being general purpose, meaning that it is not aimed at a specific usage or programming language, and is intended to be used for a wide variety of text processing tasks.The Major Tasks in Data Preprocessing are Data cleaning,Data integration,Data transformation,Data reduction,Datadiscretization,Providewell-accepted,multidimensional,view,Accuracy,Completeness,Consistency,Timeliness.

6.2 concept extraction

6.2.1 Term & phrase extraction

Step1:Consider only those internal NP (Noun Phrase) nodes whose child nodes appear as leaf in the phrase structure tree. Step2:If a node NP has a single child node tagged as noun, it is extracted as a term. Step3:If a node NP has more than one noun or adjective child nodes, then the string concatenation function is applied to club these child nodes, and they are identified as a phrase.

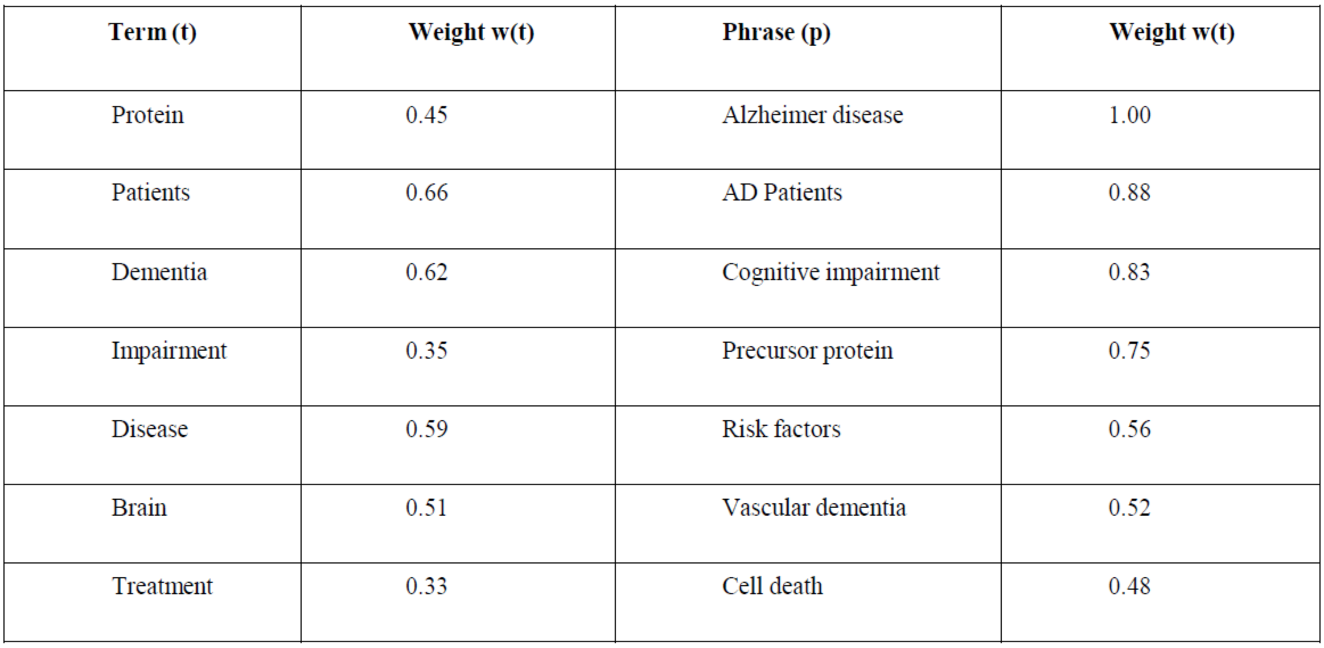
6.2.2 Weight Matrix Calculation

$$: \qquad W(Pi,j)=tf(Pi,j)*idf(Pi) \qquad (1)$$

$$: \qquad Idf(Pi)= \log \frac{(|D|)}{|\{dj:Pi \text{ €}dj\}|} \qquad (2)$$

Where ti: ith term, Pj: jth phrase,tf :term frequency ,idf: inverse document frequency,tf(pi,j) :Number of times Pi occurs in jth document,|D|:Total number of documents in the corpus,|{dj : pi Є dj}| :Number of documents where Pi appears

Table 6.2.2 A partial list of terms, phrases and their normalized weights extracted from a PubMed Abstracts on "Alzheimer disease".



Term (t)	Weight w(t)	Phrase (p)	Weight w(t)
Protein	0.45	Alzheimer disease	1.00
Patients	0.66	AD Patients	0.88
Dementia	0.62	Cognitive impairment	0.83
Impairment	0.35	Precursor protein	0.75
Disease	0.59	Risk factors	0.56
Brain	0.51	Vascular dementia	0.52
Treatment	0.33	Cell death	0.48

6.2.3 Concept-based Extractor Algorithm

The concept extractor algorithm describes the process of combining the weightstat(computed by the concept-based statistical analyzer) and the weightCOG(computed by the COG representation) into one new combined weight called weightcomb. The concept extractor selects the top concepts that have the maximum weightcombvalue. The proposed weightcombis calculated by: weightcombi= weightstati * weightCOGi, The procedure begins with processing a new document (at line 1) which has well defined sentence boundaries. Each sentence is semantically labeled according to . For each labeled sentence (in the for loop at line 3), concepts of the verb argument structures which represent the semantic structures of the sentence are extracted to construct the COG representation (at line 4). The concepts list L is sorted escendingly based on the weightcombvalues. The maximum weighted concepts are chosen as top concepts from the concepts list L. (at line 15 and 16) The concept extractor algorithm is capable of extracting the top concepts in a document (d) in O(m) time, where m is the number of concepts.

6.2.3 Feasibility analysis using SVD

It boost the precision of key terms and phrase extraction process.Each document d is represented as a feature vector, where m is the number of terms, and wti is the weight of term.

d=(wt1,wt2,……wtm)

To generate „term-by-document' matrix (A) by composing feature vectors of all the documents in the corpus. ow vector represents the terms like „„„disease", „„„Alzheimer", etc. Column vector represents the documents, D1, D2, .. Dn.

Singular value decomposition (SVD) can be looked at from three mutually compatible points of view. On the one hand, we can see it as a method for transforming correlated variables into a set of uncorrelated ones that better expose the various relationships among the original data items. At the same time, SVD is a method for identifying and ordering the dimensions along which data points exhibit the most variation. This ties in to the third way of viewing SVD, which is that once we have identified where the most variation is, it's possible to find the best approximation of the original data points using fewer dimensions.

6.2.4 Root verb extraction

For extracting the root verbs, the phrase structure tree need to be traversed for analysing the left entity, right entity and their linguistic dependencies.The important task of the relation miner is to identify the correct root verb along with the correct pair of terms and phrases within which it occurs.The objective in this step is to generate the possible roots of a given derived Arabic word. The analysis takes place in two stages: a. Segmenting the word: the system begins by determining all possible segmentations of the word.

6.3 CONCEPT REPRESENTATION

6.3.1 Add Synonym Dictionary

It is a dictionary which contains all possible synonyms of wors in the querry string.It can be used to serach semantically the queries based on the synonyms also.

6.3.2 Semantic Net

It is a Directed graph used for knowledge representation.The Concept can be expressed in the form of a Triplet <subject, verb, object>.Node represents Entities (Subject, Object), linkRelation (Verb).

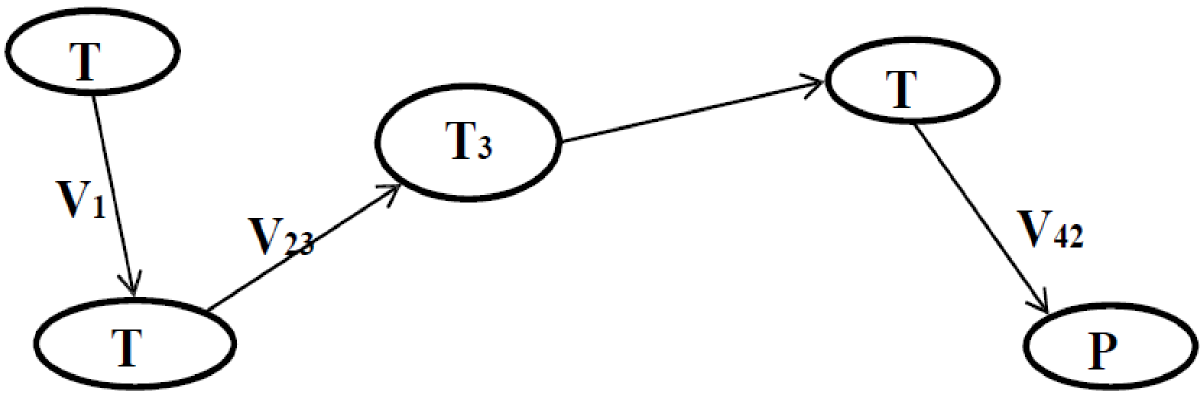


Fig 6.3.2 A semantic net representation

For preserving the sentence meaning the inheritance property of semantic net has been utilized.The transitive property (If A□B and B□C, then A□C) is used for making connection between the related sentences. In some sentences subject or noun phrases can be expressed using the keywords like it, this, that, etc. In these cases the subject of the previous sentence is taken, for making connections between adjacent or related sentences.

6.4 CLUSTERING

The documents are clusterd according to the concepts.A hierarchial agglomerative clustering is performed.

6.4 UPDATION OF DATABASE

Semantic searching results a very efficient retrieval characteristics.So the devised knowlegde can be used to update the database.

VII. PERFORMANCE MEASURES

True Positive (TP): Number of correct components the system identifies as correct

False Positive (FP): Number of incorrect components the system falsely identifies as correct False Negatives (FN): Number of correct components the system fails to identify as correct Precision (P): The ratio of true positives among all retrieved instances

: P=tp/(tp+fp)

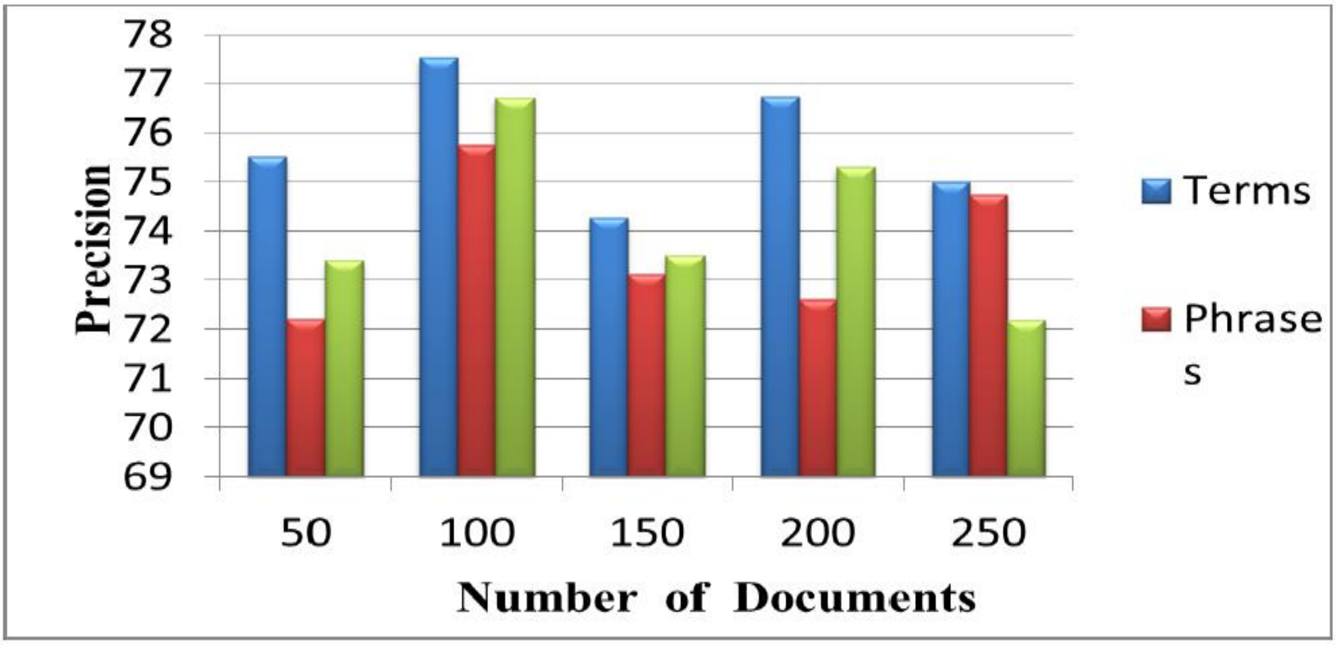
Recall (R): The ratio of true positives among all positive instances

:R=tp(tp+fn)

F-measure (F): The harmonic mean of recall and precision

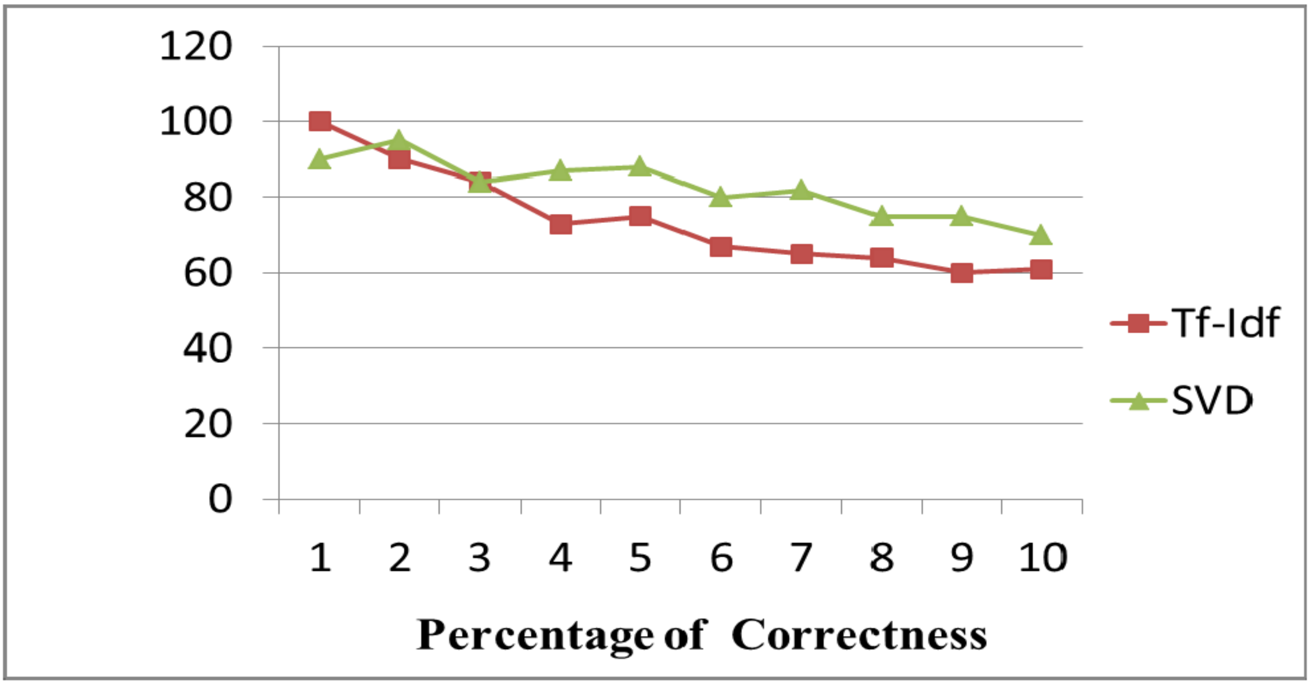
F=2pr/(p+r)

7.1.Precision Graph



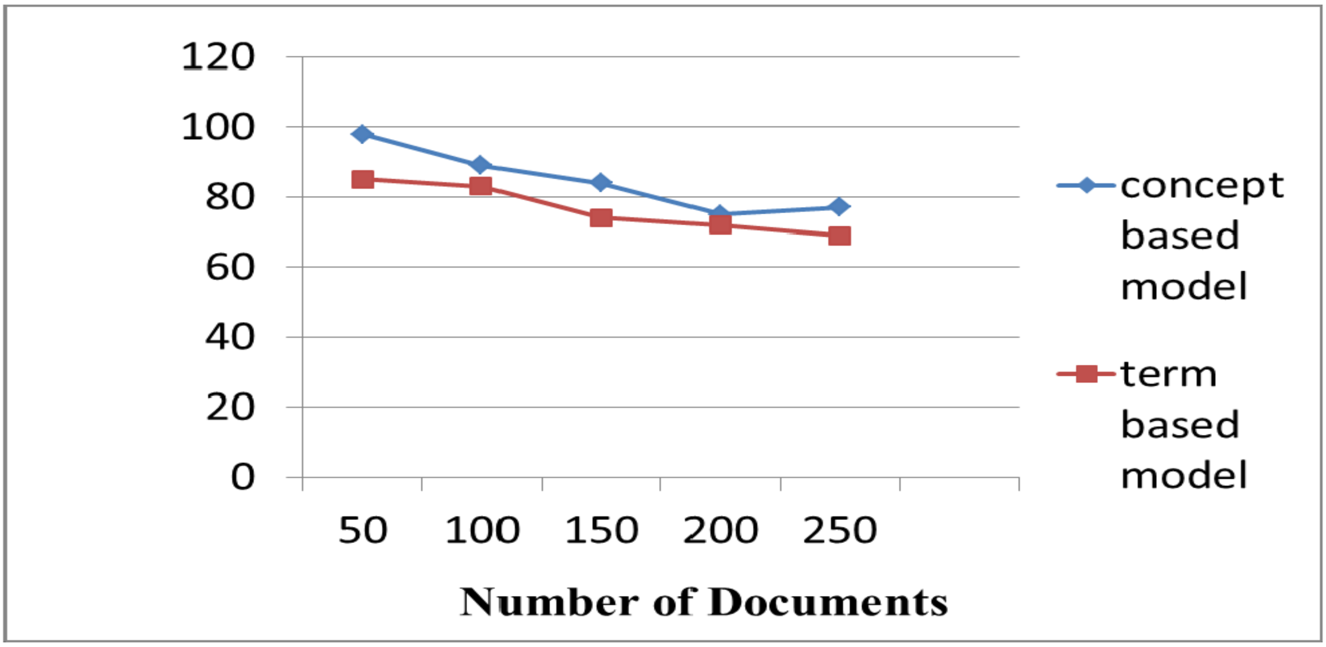
The graph above shows the precision values of documents on the basis of terms, phrases and verbs.

7.2 Performance comparison graph between SVD and TF-IDF



From the performance comparison graph, it can be inferred that SVD performs efficiently when compared to TF-IDF.

7.3Performance comparison graph of Concept based model



Graph shows that concept based model performs efficiently when compared to term based model.

VIII. CONCLUSION

The concept based mining model preserves the semantic structure of the sentences with low error rate. The Inheritance property of the Semantic net provide major contribution for preserving semantics.Our model is compared with term based model(Vector Space Model), results shows that the precision and recall values of concept mining model are higher as compared to the term-based techniques. At the same time, accuracy and relevancy is high. Experimental result shows that the technique used in the proposed work minimizes the time and the work load of the doctors in analyzing information about certain disease and treatment in order to make decision about patient monitoring and treatment. This concept mined document can be used in medical health care domain where a doctor can analyze various kinds of treatment that can be given to patient with particular medical disorder. The doctor can update the knowledge related to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease.

The doctor can gain idea about particular medicine that are effective for some patient but causes side effect to patient with some additional medical disorder. The patient can also use this extracted document to get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies.

FUTURE ENHANCEMENT

In the future, such model can be further extended to include the non-segmented text documents. It can
also be extended to categorize the images, audio and video - related data.

REFERENCES

1. Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger ,“Tackling The POOR Assumption Of Naïve Bayes Text Classifier”, Proceedings Of The Twentieth International Conference On Machine Learning (ICML-2003), Washington DC, 2003.
2. T.Mouratis, S.Kotsiantis, “Increasing The Accuracy Of Discriminative Of Multinominal Bayesian Classifier In Text Classification”, ICCIT‟09 Proceedings Of The 2009 Fourth International Conference On Computer Science And Convergence Information Technology. [3] B.Rosario And M.A.Hearst, ”Semantic Relation In Bioscience Text”, Proc. 42nd Ann. Meeting On Assoc For Computational Linguistics, Vol.430,2004.
3. M.Craven, ”Learning To Extract Relations From Medline”, Proc. Assoc. For The Advancement Of Artificial Intelligence.
4. Oana Frunza.et.al, “A Machine Learning Approach For Identifying Disease-TreatmentRelations In Short Texts”, May 2011
5. L. Hunter And K.B. Cohen, “Biomedical Language Processing:What‟s Beyond Pubmed?” Molecular Cell, Vol. 21-5, Pp. 589-594,2006.
6. Jeff Pasternack, Don Roth “Extracting Article Text From Webb With Maximum Subsequence Segmentation”, WWW 2009 MADRID.
7. Abdur Rehman, Haroon.A.Babri, Mehreen saeed,” Feature Extraction Algorithm For Classification Of Text Document”, ICCIT 2012.
8. Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl.,”Efficient Text Extraction Aalgorithm
9. Using Color Clustering For Language Translation In Mobile Phone” , May 2012.
   U.Y. Nahm and R.J. Mooney, “A Mutually Beneficial Integration of Data Mining and
10. Information Extraction,” Proc. 17th Nat‟l Conf. Artificial Intelligence (AAAI ‟00), pp. 627-
    632, 2000.
11. B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.