# Concepts Identification in Large Scale Datasets for Efficient Text Categorization

**[1]Y. Sri Lalitha, [3]Lohitha Bhogapati**
[1] Professor, GRIET, Hyderabad
[2]M.Tech Student, GRIET, Hyderabad

**ABSTARCT**
Concept Based Document removal is an increasing modern Research with the intention of activities in the direction of gather important in sequence as of normal words processing term. It might be there uncertainly eminent because the path of investigative texts toward takes out in sequence with the intention to be realistic happening exacting purposes. In this case, the mining representation capable of detain provisions that identify the concepts of the ruling or document, which tends toward notice the theme of the document. In an vacant job, the concept-based taking out representation be utilized merely intended for usual transcript credentials clustering in accumulation to clustered the transcript parts of the credentials in count to capably discovers important the same concepts between credentials, according toward the semantics sentence. however the negative aspect of the job be with the intention of the accessible job cannot subsist connected toward net credentials clustering along with the transcript categorization intended for the credentials be an undependable lone. Concept-Based drawing out representation used for attractive transcript Clustering.
**Keywords:** Concept-based drawing out form, Concept-based similarity, Text clustering, Document clustering, Hadoop.

## I. INTRODUCTION

Within transcript drawing out techniques, the tenure occurrence of a tenure (word or expression) be computed toward investigate the significance of the phase in the text. Nevertheless, two terms can comprise the identical occurrence in their documents, but one term contributes more toward the significance of its sentences than the other term. It be significant toward message with the intention of extracting the associations among verbs and their opinion within the identical verdict have the possible used for analyzing provisions in a verdict. The in sequence regarding who be responsibility what did you say? Toward whom clarifies the part of every tenure into a verdict toward the significance of the most important issue of so as to verdict.

Here, a narrative concept-based drawing out form is future which captures the semantic configuration of every tenure inside a verdict accumulation for text somewhat than the regularity of the tenure inside a text merely. Here, three procedures intended for analyzing concepts resting on the verdict, text, in adding together to quantity levels are computed. Each verdict is labeled through a semantic position labeler so as to determine the provisions which have a say toward the verdict semantics coupled through their semantic roles in a sentence. Each term that has a semantic job within the verdict, be called a idea. Concepts able to exist also expressions or else phrases in addition to be absolutely reliant resting on the semantic configuration of the verdict.

When a latest text be introduced toward the method, the wished-for drawing out representation in notice a idea equivalent since this content here in the direction of every the earlier processed documents within the data set through scanning the latest document also extracting the identical concepts. This resemblance calculates outperforms other resemblance procedures. So based in resting on tenure investigation models of the text merely.

The resemblance among credentials be based resting on a grouping of sentence-based, document-based, along with

corpus-based idea investigation. usually, transcript text clustering methods effort toward set aside the credentials keen on groups anywhere every cluster represents a number of theme so be there dissimilar than individuals topics that are base taking place such a characteristic vector. Examples contain the cosine quantify by the Jaccard measure. The comparison among the credentials is measured by one of several similarity procedures that are base on such a feature vector. Examples comprise the cosine evaluate in the Jaccard quantify and nearness. Proximity will check the connection among two documents very precisely.

## II. RELATED WORK

Pradhan et al. has cast tagging problem [6]. In his work he research that usual, exact and wide-coverage techniques that can annotate logically happening text among semantic squabble structure can play a key task in NLP application is Information Extraction, and Question answer Summarization. Low semantic parsing the procedure of conveying a plain WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW, etc. Formation to sentence in text is the procedure of produce such a markup. When accessible with a ruling, a parser should, for each predicate in the verdict, identify and label the predicate's semantic opinion. This procedure entail identify groups of words of sentences that signify these semantic advice and conveying specific labels to them.

Gruber and Fillmore proposed thematic roles. Normally, the semantic formation of a sentence may be exemplified in the structure of verb disagreement structure. The study of the roles correlated with verbs is referred to a thematic role or container role study [1]. Thematic roles are sets of grouping that make available a shallow semantic language to describe the verb arguments.

Marcus et al., 1994 planned that predicate argument associations are distinct for part of the verbs. He has completed his job on consequences using PropBank1 (Kingsbury et al., 2002), a 300k-word corpus in which predicate argument dealings are noticeable for part of the verbs in the Wall Street Journal (WSJ) part of the Penn Tree- Bank (Marcus et al., 1994) [6]. The opinion of a verb is labeled ARG0 to ARG5, where ARG0 is the PROTOAGENT (frequently the theme of a transitive verb) ARG1 is the PROTO-PATIENT (frequently its straight object), etc. Prop Bank effort to care for semantically associated verbs constantly. In accumulation to these CORE ARGUMENTS, supplementary ADJUNCTIVE ARGUMENTS, referred to as ARGMs are also noticeable. Some cases are ARGMLOC, for locatives, and ARGM-TMP, for sequential. Fillmore expressed a shallow semantic interpreter [7] depend on semantic roles that are fewer field specific than to airport or joint endeavor company. These roles are defined in intensity of semantic frames (1976), which describe abstract actions or relations, along with their participants.

Gildea moreover Jurafsky was the initial to concern a statistical knowledge method to the FrameNet database [6]. They existing a discriminative model for formative the nearly all likely job for a basic, given the frame, predicator, and extra features. S. Y. Lu we planned a syntactic cluster procedure, in which each formed cluster is described by a outline syntax [8].

The process gives way not just the clustering outcome syntax for every cluster. In classify to do so, a grammar must be conditional when a latest cluster is initiate, and later on it is simplified whenever an input pattern is extra to the alike cluster. Error-correcting parsers are employed to measure the distance among an input pattern and the language generate from the incidental grammars.

The enter pattern is after that classify according on the way to nearby neighbor syntactic recognition rule. The importance of the syntactic clustering process is the utilize of syntax in which the ladder of the formation of outline is portrayed. S. Kaski et al., in his work talk about that lone of the usual methods of searching for texts that equal to a

query is to guide all the words (here after called terms) that have come into view in the document collection [9]. The query itself has a typical file with suitable keywords, is evaluate with the term list of all one document to discover documents that contest the query. Conditions can be joint by Boolean logic in classify to control the breadth of matching.

**SYSTEM ARCHITECTURE**

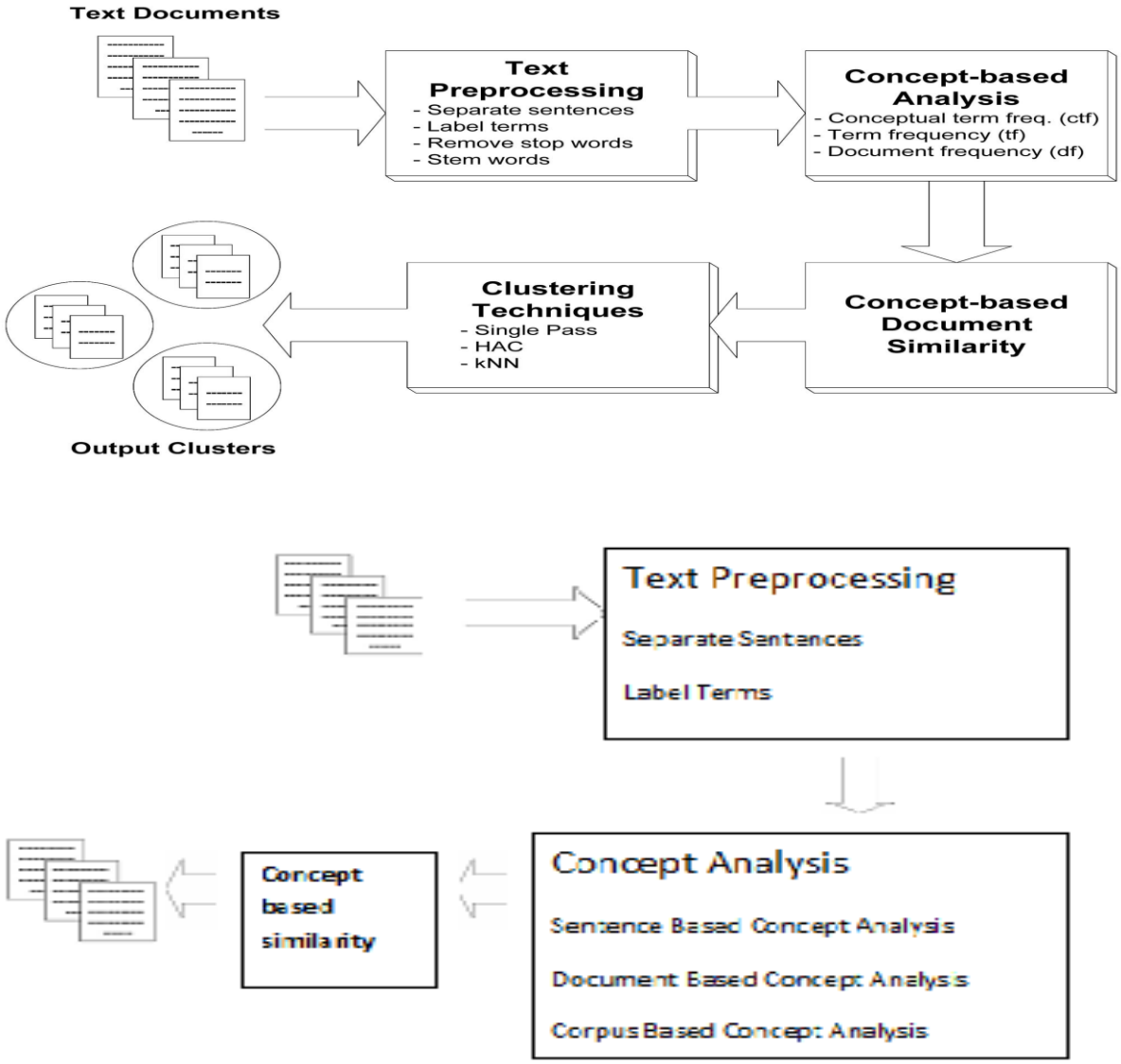


Fig 1: Architecture of Concept Based Model

## III. IMPLEMENTATION

**Concept based Mining Model Process**

The project contains sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, moreover concept-based similarity measure. A raw text document is input for likely model. All

documents have well defined sentence limitations. Apiece sentence in the document is branded mechanically founded on parser. Once operating the semantic role labeler, all sentences in the document may have one or else more labeled verb argument structures. In this form, together the verb and the argument are believed as conditions. Single term may be argument for more than lone verb in the similar sentence.

So that the way of term can have more than single semantic role in the identical sentence. In such cases, this tenure plays significant semantic roles that contribute to the connotation of the sentence. A label conditions both word or phrase is measured as model. The System architecture consists of the follow major modules are Text preprocessing, Concept Analysis and Concept based similarity measure.

### A. CONCEPT BASED CLUSTERING

K-medoid is a classical partitioning method of clustering that clusters the data set of n objects into k number of clusters. This k: the amount of clusters necessary is to be put by addict. This algorithm minimizes the total of variation among every object and its equivalent situation point. It randomly chooses k objects in dataset D as preliminary representative objects known as medoids. A medoid is distinct as the object of a cluster, whose standard difference is minimal to every the things in the cluster i.e. the majority centrally positioned situation in the given data set. It then assigns both objects to the adjacent cluster depending ahead on the object's detachment for the cluster medoid. Following assigning data object to a exacting cluster the fresh medoid is determined.

1) Input k: the figure of clusters. D: a dataset contains n objects.
2) Output A set of k clusters.
3) Algorithm 1. At random decide k objects in D in the primary representative objects; 2. For every object through data set D.

### B. CONCEPT BASED INDEXING

The weight in the individual term i document j is defined as in bellow equation.

$$wji = tfji \times idfji = tfji \times \log2 ( n/dfji)$$

where tfji is the amount of incidence of phrase i in the document j, dfji is the period frequency in the group of documents and n is the whole quantity of documents in the group. Concept based indexing recover the load of the idea since it think not simply the concept word but as well all the terms that are linked to the concept word by way of the semantic relations.

### C. TEXT PREPROCESSING

#### i. Label Terms

A rare text document as input for projected model. All documents have fine distinct result limitations. Every sentence within the file is tag mechanically found on the parser. Once operation the semantic task labeler, every one sentence within the document may have sole or extra labeled verb case structures. The labeled verb dispute structure, the output of the role category of task and are imprison and investigated by the concept-based mining model on sentence, document stages. In form, together the verb, argument are careful as terms. One term preserve be an argument to additional than lone verb in the similar sentence. This word can have extra semantic role in the matching sentence. In such belongings, that word plays significant semantic position that adds to the importance of the sentence. In the concept-based mining model, Labeled expression also word otherwise phrase is consider as idea.

**ii. Removing stop words**

In computing end words are words which are drinkable out prior to, or after, processing of normal words data (text). It is controlled by human input and not automated. There is not one definite list of stop words which all tools use, if even used. Some tools particularly let alone using them to maintain phrase seek out.

**iii. Stem words**

In linguistic morphology, stemming is the path for dropping inflected vocabulary to their stem; base or else root type usually a write word type. The stem require not exist the same toward the morphological root of the word; it is typically enough to connected terms map to the similar stem, still if that stem is not inside a legal root. Algorithms for stemming have been deliberate inside computer knowledge since 1968. A lot of search engines care for words through the similar stem as synonyms as a class of query extension, a procedure call conflation. Stemming courses are normally referred in the direction of stemming algorithms otherwise stemmers.

## IV. MAP REDUCE IMPLEMENTATION

Indoctrination model was projected in 2004 by the Google, which is worn in dealing out and generate big data sets implementation. This framework solves troubles, like data distribution, job scheduling, fault tolerance, machine to machine communication, etc.

Mapper Map function require the user to hold the input of a couple of key value and form a set of intermediary key and value pairs. <key,value> consists of two parts, value meant for the data associated to the job, key meant for the "group number " of the value . MapReduce joins by intermediate ideals by same key and after that launch them to reduce function. Map algorithm procedure is illustrated as follows: Step1: Hadoop and MapReduce framework create a map task for all Input divide, and every Input divide is createated by the Input Format of job. Each <Key,Value> communicate to the map task. Step2: Execute Map task, procedure the input <key, value> to figure a latest <key, value>.

This scheme called as "divide into groups". That is, make the correlated values communicate to the similar key words. Output key value pairs don't required the equivalent type of the input key value pair. A known input value pair contains mapped into 0 or else more output pairs. Step3: Mapper's output is sorted can be owed to apiece Reducer. The figure of blocks and the number of job reduce jobs is the similar.

Users can execute Partitioned interface to manage which key is assign to which Reducer. Reducer Reduce function in addition provided to the user, which handles the intermediate key in pairs and the worth put related to the intermediatekey value. Reduce function mergers these ethics, to find a petite put of values, the procedure k is called as "merge ". But this isn't effortless gathering.

There are multipart operations in the procedure. Reducer makes a collection of transitional values set that connected with the similar key smaller. In MapReduce framework, the programmer does not require to concern regarding the particulars of data communication, so <key, value> is the communiqué interface for the programmer in Map Reduce model. The< key, value> can be there as a "letter", key is the letter's relocation address, value is the letter's content.

With the similar address letters will be sent to similar place. Programmers simply require for setting up correctly<key, value>, Map Reduce framework can mechanically able-bodied exactly cluster the values with the similar key mutually. Reducer algorithm route is describes as follow:

**Step1:** Shuffle, input of Reducer is the output of sorted Map per. In this period, Map Reduce will allocate linked block

for apiece Reducer.
**Step2:** Sort, in this phase, the input of reducer is group according to the key. The two phases of Shuffle then Sort synchronized.
**Step3:** Secondary Sort, if the key grouping rule in the intermediate procedure is special from its rule prior to reduce. Here could able to describe a Comparator.

The comparator is worn to group midway keys meant for the next time. Reduce task is a whole, can't be separated. They supposed to be worn jointly in the agenda. We name a Map Reduce the process like an MR process. In an MR process, Map jobs run in parallel, Reduce jobs run in parallel, Map then Reduce jobs run serially. An MR process run in sequential, synchronization among these operations is definite by the MR system, without programmer's participation.

### A. K-MEANS OVER MAPREDUCE

In this part, initial we explain easy or Direct K-Means clustering algorithm. in that case, I will illustrate Distributed K-means clustering algorithm mutually above Hadoop framework. K-means document clustering come in partition technique of clustering wherever one-level (un-nested) partitioning of the data summit is created. If K is to be preferred number of clusters, after that partition approach typically get every K clusters on one time. K-means is lying on the thought that a middle point can characterize a cluster. In exacting, for K-means I utilize the idea of a centroid, which is the mean or median point for collection of points. Basic k-means algorithm is known.

**Input:** K: amount of cluster, D: Top N documents
**Output:** K clusters of documents Algorithm
**Step1:** Generate K centroids C1, C2 …Ck by at random choose K documents from D duplicate awaiting for modify in cluster between two successive iterations.
**Step2:** For every document di in D for j = 1 to K Sim(Cj, di) = Find cosine relationship among di and Cj finish for consign di to cluster j for which Sim(Cj, di) is most ending for.
**Step3:** Bring up-to-date centroid for all clusters end loop.
**Step4:** End K-Means.

### B. TEXT CLUSTERING

Document clustering has to be investigated in different districts of text mining fit information retrieval. Document clustering has to be considered intensively since of its large application regions such like Web Mining, Search Engine, Information Retrieval. Document clustering is routine organization of documents keen on clusters or else groups, end result that, documents inside a cluster contain towering comparison to one more, but are especially unlike to documents in extra clusters.

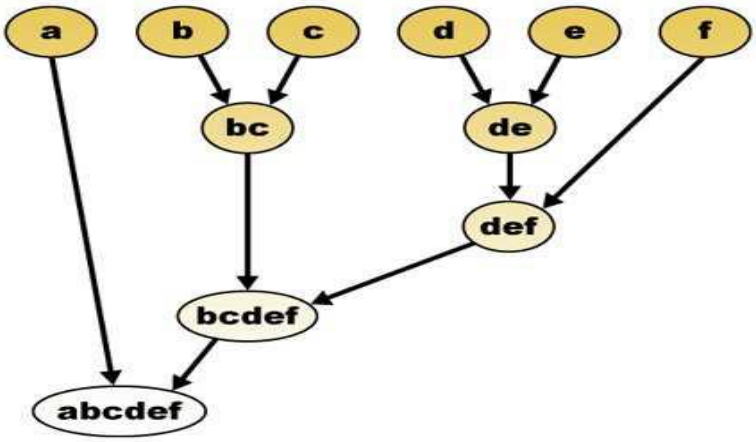


Fig 2: Hierarchical Text clustering

Hierarchical Text clustering defined as additional terms; the grouping is stand on the rule of maximizing intracluster relationship and minimizes intercluster similarity. The major brave of clustering just professionally recognize significant sets that are quickly annotated.

### C. CONCEPT BASED SIMILARITY MEASURE

A concept-based similarity measure, based on matching concepts through sentence, document is devised. The concept-based comparison calculates lying on three vital aspects. First, the analyzed labeled conditions are the concepts to facilitate imprison the semantic structure of all sentence. Second, the frequency of a theory is worn to estimate the part of the model to the significance of the sentence, because the main topics of the document. Last, the amount of documents that contains the analyzed concepts is worn and distinguish amid documents in manipulative the comparison. It classifies to evaluate the similarity Jaccard Distance and Proximity measures are used. Jaccard Distance measure shows the dissimilarity among two items whereas Proximity measure shows the comparison among two things.

## V. EXPERIMENTAL RESULTS

Porter's algorithm was developed in favor of the stemming of English-language texts except the growing significance of information recovery in the 1990s led to a propagation of concern in the growth of conflation techniques that would develop the search of texts printed in other languages. By this point, the Porter algorithm had turn into the regular for stemming English, and it therefore provided a natural model for the processing of further languages.

In various of these latest algorithms the only association to the new use for a incredibly classified suffix dictionary (Porter, 2005), however Porter himself has developed a entire sequence of stemmers so as to depict on his new algorithm alond wrap Romance (French, Italian, Portuguese and Spanish), Germanic (Dutch and German) and Scandinavian languages (Danish, Norwegian and Swedish), like Finnish and Russian (Porter,2006). Porter's algorithm is significant for two reasons.

First, it presents a effortless loom to conflation that seems to job fit in observe and that is appropriate to a choice of languages. following, it have spur concern in stemming as a issue for study in its personal exact, relatively than simply as a low-level component of an information rescue system. The experiment is conducted on the documents that are collected from MEDLINE based on three categories like cancer, virus and eye infection. First of all three documents containing data from Medline is uploaded. Stop word method is applied to eliminate useless or anonymous data from

the document. The outcome of this is that we get formatted document. After that concept search method is applied to all three documents. Particular concept is searched in all the documents. It retrieves the occurrence of searched concept from all the documents. Similarly term method is called, repeating similar procedure as mentioned, again weight is calculated.

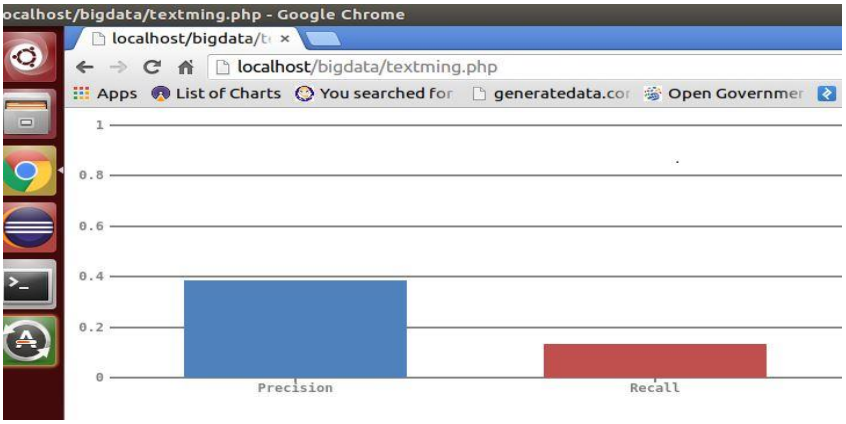


Fig 3: Precision Recall of the Categorized Clusters

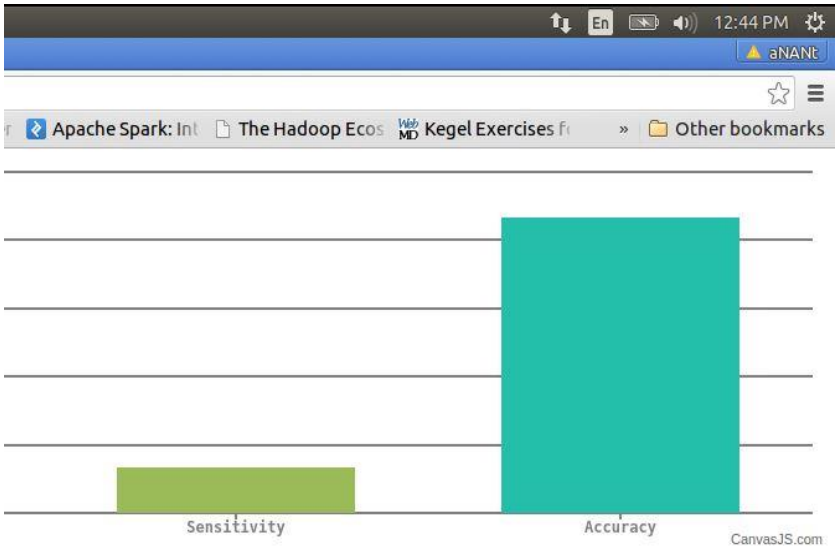


Fig: Accuracy Graph of Text Categorization

## VI. CONCLUSION

Here tried to apply the concept-based approach to text clustering. The projected method exploited fully the semantic makeup by the sentence in the papers in sort to achieve good quality of clustering. To the input document Text pre-processing was initially done where the sentences were separated and labeled with verb argument structures. Further stop words were removed and stemming was done. This was followed by components that performed sentence based, document based, corpus-based and concept-based analysis where the abstract tenure frequency measure (ctf),

concept-based term frequency measure (tf), document term frequency measure (df) and the concept based comparison measure were determined respectively.

**REFERENCES**

[1] Shady Shehata, Fakhri Karray and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering*", IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No.10, pp. 1360 – 1371, October 2010.

[2] B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.

[3] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report 941, Norwegian Computing Center, June 1999.

[4] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Comm. ACM, vol. 18, no. 11, pp. 112-117, 1975.

[5] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[6] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky, "Shallow Semantic Parsing Using Support Vector Machines," Proc. Human Language Technology/North Am. Assoc. for Computational Linguistics (HLT/NAACL), 2004.

[7] C. Fillmore, "The Case for Case," Universals in Linguistic Theory, Holt, Rinehart and Winston, 1968.

[8] S.Y. Lu and K.S. Fu, "A Sentence-to-Sentence Clustering Procedure for Pattern Analysis," IEEE Trans. Systems, Man, and Cybernetics, vol. 8, no. 5, pp. 381-389, May 1978.

[9] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen, "WEBSOM—Self-Organizing Maps of Document Collections," Proc. Workshop Self- Organizing Maps (WSOM '97), 1997.

[10] D. Jurafsky and J.H. Martin, Speech and Language Processing. Prentice Hall, 2000.

[11] U.Y. Nahm and R.J. Mooney, "A Mutually Beneficial Integration of Data Mining and Information Extraction," Proc. 17th Nat'l Conf. Artificial Intelligence (AAAI '00), pp. 627-632, 2000.

[12] L. Talavera and J. Bejar, "Generality-Based Conceptual Clustering with Probabilistic Concepts," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 196-206, Feb. 2001.

[13] H. Jin, M.-L. Wong, and K.S. Leung, "Scalable Model-Based Clustering for Large Databases Based on Data Summarization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1710-1719, Nov. 2005.

[14] T. Hofmann, "The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data," Proc. 16th Int'l Joint Conf. Artificial Intelligence (IJCAI '99), pp. 682-687, 1999.

[15] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," Proc. Knowledge Discovery and Data Mining (KDD) Workshop Text Mining, Aug. 2000.

16. K. Aas and L. Eikvil. Text categorisation: A survey. technical report 941. Technical report, Norwegian Computing Center, June 1999.

17. M. Collins. *Head-Driven Statistical Model for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

18. R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *Proceedings of First International Conference on Knowledge Discovery and Data Mining*, pages 112 - 117, 1995.

19. S. Shehata, F. Karray, and M. Kamel. Enhancing text clustering using conceptbased mining model. In *ICDM*, pages 1043{1048, 2006.

20. W. Francis and H. Kucera. *Manual of information to accompany a standard corpus of present-day edited americanenglish, for use with digital computers*, 1964.

21. T. Joachims. Text categorization with support vector machines: learning with many relevant features. *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137-142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

22. S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology/North American Association for Computational Linguistics (HLT/NAACL)*,
2004.