Text Clustering using a WordNet-based Knowledge-Base and the Lesk Algorithm

Jyotirmayee Choudhury
Department of CSE
IIIT Bhubaneswar
Bhubaneswar, India

Deepesh Kumar Kimtani
Department of CSE
IIIT Bhubaneswar
Bhubaneswar, India

Alok Chakrabarty
Department of CSE
IIIT Bhubaneswar
Bhubaneswar, India

ABSTRACT
In this paper we are proposing a text clustering method based on a well-known Word Sense Disambiguation (WSD) algorithm, the Lesk algorithm, to classify textual data by doing highly accurate Word Sense Disambiguation. The clustering of text data is thus primarily based on the context or meaning of the words used for clustering. The Lesk algorithm is used to return the sense identifiers for the words used to classify the text files by looking up the senses of a word in a Knowledge-Base similar to the English WordNet (enriched with more informative columns or fields for each synset [synonym set] of the English WordNet database), so as to greatly increase the chances of contextual overlap, thereby resulting in high accuracy of proper sense or context identification of the words. The proposed scheme has been tested on a number of heterogeneous text document datasets. The clustering results and accuracies, obtained using the proposed scheme, have been compared with the results obtained using the K-means clustering algorithm on the Vector Space Models generated for all the heterogeneous textual datasets. Experimental results show that our algorithm performs much better than the Vector Space Model (VSM) and K-means based approach. The technique will thus help the users much better in searching for meaningful contextual information from a highly diversified collection of textual information, which is a key task of the information overload problem.

General Terms
Natural Language Processing, Text Clustering, Word Sense Disambiguation, Information Retrieval

Keywords
lesk, synset, WSD, knowledge base, k-means, vector space model, context, wordnet

1. INTRODUCTION
Text mining is being developed as an emerging technology to handle the increasing text data. The amount of information available on the Web is increasing day by day with globalization and rapid development of Internet. So the need to develop applications to manage this massive amount of varied information is also growing.

Text clustering is one of the fundamental functions in text mining. Clustering means to divide a set of text documents into different categorical groups so that documents in the same categorical group describe the same topic. There are many uses of clustering in real applications, for example, grouping the Web search results and categorizing digital documents [1].

For instance, if two documents use different collections of keywords to represent the same topic, they may be falsely assigned to different clusters. This problem arises due to the lack of shared keywords, although the keywords they use are probably synonyms or semantically associated in other forms[2]. Word ambiguity is a severe problem in the keywords-based methods. For example, if 'bat' occurs several times in a document, should the file be classified to "Sport" or "Mammal" [3]?

Clustering based on only cosine distance similarity between words may cause error in clustering of datasets. So in this paper, we propose a text clustering method based on sense disambiguation.

2. APPROACHES
2.1 Knowledge-based
In this approach, disambiguation is carried out by using information from an explicit lexicon or knowledge-base. The lexicon may be a machine readable dictionary, thesaurus or a WordNet [3].

2.2 Corpus-based
This approach challenges to disambiguate words by using information gained from trained data, rather than taking it directly from an explicit knowledge-base. Training can be carried out either on a disambiguated corpus or a raw corpus. In a disambiguated corpus, the semantics of each polysemous lexical item has been marked, while in a raw corpus, the semantics has not been marked yet [3].

2.3 Hybrid Approaches
A good example of this kind of approach is the Luk's system which uses the textual definitions of senses from a machine readable dictionary to identify relations between senses. It then uses a corpus to calculate mutual information scores between the related senses in order to discover the most useful information. In this way, the amount of text needed in the training corpus is reduced [3].

3. DOCUMENT CLUSTERING
Document clustering is a knowledge discovery technique which categorizes the document set into meaningful groups. Document clustering has been investigated for use in a number of different areas of text mining and Information Retrieval (IR) [4]. Initially, document clustering was investigated for improving the precision or recall in IR systems and as an efficient way of finding the nearest neighbors of a document. More recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to a user's query. Document clustering has also been used to automatically generate hierarchical clusters of documents.

Document clustering is the process of partitioning a set of n unlabeled documents into clusters such that documents in each cluster have high similarity in comparison to one another and share some common concepts but are dissimilar to documents in other clusters. Each concept is conveniently represented by some key terms. Various clustering approaches (such as K-means, Fuzzy c-means, agglomerative, hierarchical etc.) are available to cluster documents. Here for the comparison of our approach with already existing clustering approaches we chose K-means algorithm for the clusters formation as it combines the strengths of partitioned and hierarchical clustering methods by iteratively splitting the biggest cluster using the basic K-means algorithm.

3.1 K-means clustering

K-means clustering is one of the simplest and most popular unsupervised learning algorithms to solve the clustering problem. K-means clustering generates a specific number of disjoint, flat clusters. That is, the K-means function partitions the observations extracted from the data into K mutually exclusive clusters, and returns a vector of indices, indicating to which of the K clusters each feature set has been assigned. This method is more efficient than hierarchical clustering, especially for large data sets and high-dimensional data sets.

The basic K-means Algorithm for finding K clusters is as follows [4]:

1. Select K points as the initial centroids.

2. Assign all points to the closest centroid.

3. Re-compute the centroid of each cluster.

4. Repeat steps 2 and 3 until the centroids don't change.

3.2 Vector Space Model

It is a statistical model for representing text documents as vectors of identifiers, such as, for example, index terms. Correlation between term vectors implies similarity between documents. Each document is represented as a vector in the Vector Space, a matrix. Position 1 corresponds to term 1, position 2 to term 2, position *t* to term *t*.

$d_j = W_{1j}, W_{2j}, W_{3j,.....}, W_{tj}$, where each W_{ij} is the weight of the term *i* in document *j*

$V_{d_j} = [W_{1,j}, W_{2,j}, W_{3,j,.....}, W_{t,j}]$, the VSM entry for document d_j, where,

Term Weight, $W_{i,j} = tf_i * \log(D / df_i)$

$tf_i =$ Term frequency (term counts) or number of times term *i* occurs in document *j*.

$df_i =$ Document frequency or number of documents containing term *i*

$D =$ Total number of documents in a dataset

$idf_i = \log(D / df_i)$, inverse document frequency

The measure of similarity between two *t* dimensional document vectors is obtained by finding the cosine of the angle between them.

Similarity

$$sim(d_j, d_k) = \frac{d_j * d_k}{\|d_j\| \|d_k\|} = \frac{\sum_{i=1}^{t} W_{ij} * W_{ik}}{\sqrt{\sum_{i=1}^{t} (W_{ij})^2} * \sqrt{\sum_{i=1}^{t} (W_{ik})^2}}$$

4. WORD SENSE DISAMBIGUATION IN TEXT CLUSTERING

In Information Retrieval problems, WSD does the task of selecting the most appropriate meaning for any given word with respect to its context. It is used with the aim of improving retrieval effectiveness. As sense repository we used a knowledge-base developed on the basic database structure of the English WordNet. WordNet is a lexical database containing nouns, verbs, adjectives and adverbs, organized in synonym sets (synsets). Many words in the natural language have more than one sense (or lexical meaning), e.g., if a query is submitted which contains a polysemous word "present", but is intended by the user to be interpreted with only one of its senses say in the sense of period of time, then there are many chances of false interpretation, corresponding to the word's occurrences with a different sense say "present" in the sense to give a present. Our sub-requirement in the proposed text clustering approach is to choose the appropriate sense from sense repository, in a reasonably fast, simple yet efficient way, so we have chosen the simple and classical Lesk algorithm out of many other WSD algorithms for document clustering [5] [6].

5. PROPOSED APPROACH

In this section we shall briefly describe the relevant work done in embedding WSD in the document clustering scheme. This method requires WSD, because it is needed to identify the correct sense for each word in a document. This section provides the details of every phase involved in the proposed approach for clustering documents**.**

Pool of English text documents are taken for clustering. The document set (dataset) is input to the clustering system and the system is provided with the word(s) based on which clustering is to be done. The system then first makes two groups of documents, one group for the documents in which the word(s) is/are there and another group for the documents in which the word(s) is/are not found. The system then applies the Lesk algorithm on the text documents in the first group to find the appropriate senses, in the form of synset identifiers or IDs from the WordNet-based knowledge-base, of the given word(s) in all the sentences of all the text documents of the group. After this, categorical grouping of the text documents is done based on the equality of sense identifiers of the word(s), so that each group contains documents of same theme or context.

Algorithm

The scheme presented above can be algorithmically presented as follows:

Step 1: Provide the word(s) for WSD.

Step 2: Search in all documents of the dataset for the presence or absence of the word(s).

Step 3: Group the text documents where the word(s) were not found as a separate cluster.

Step 4: Apply Lesk algorithm on the text documents of the other group by extracting all the individual sentences containing the word(s) to obtain the relevant sense identifiers or IDs.

Step 5: Group text documents in same or different cluster based on the maximum equality of sense IDs for the word(s).

Step 6: Increasingly assign cluster number to all the text documents' groups formed.

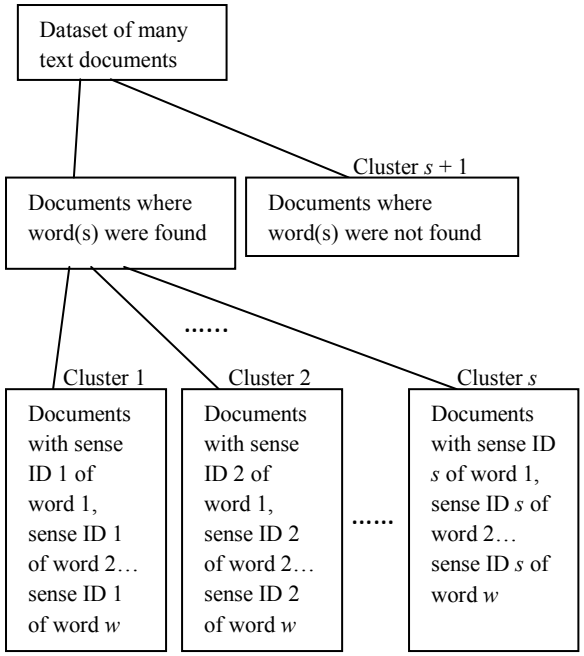

Fig. 1 Visualization of clusters and relationships between clusters based on our approach

6. EXPERIMENTAL EVALUATION AND COMPARISON

We conducted experiments using various datasets of different themes to achieve a systematic improvement in clustering results. The results showcase the effectiveness of the proposed approach.

In Table 1, we present one sample clustering result on a dataset to compare VSM and K-means clustering based approach with our proposed WSD based clustering approach.

Fig. 2 presents the graphical representation of comparison between the two approaches. Out of the 15 text documents of the dataset, 5 documents got misclassified using the K-means approach whereas only 1 document was misclassified using the proposed WSD approach.

Table 2 presents the comparison with clustering results based on human intelligence (i.e., human classified datasets). It lists out the cluster number for the documents based on VSM and K-means approach, the proposed approach and human intelligence. If someone opens the documents and tries to do clustering just by judging the documents then the cluster number given by them is Human Classified Cluster number. A comparison with the Human Classified Cluster numbers, for the dataset, with the cluster number results obtained using VSM and K-means clustering approach as well as with those of the proposed WSD approach clearly shows the error rate values as 33.33% and 6.66 % respectively. Thus with the proposed approach the error rate decreases very significantly.

7. CONLUSION

In this paper, we proposed a text clustering method based on word sense disambiguation. The sense-based text clustering algorithm is an automatic technique to disambiguate word senses and then classify text documents. If this automatic technique is applied in real applications, the clustering quality and search results of e-documents will dramatically improve. It will be a great contribution to the management system of Web pages, e-books, digital libraries, etc [7].

This method requires WSD, because it is needed to identify the correct sense (by finding the synonym set identifier) for each word in a document.

In particular, we compared two approaches of document clustering; the proposed WSD based clustering approach and VSM and K-means clustering based approach with human intelligence. For K-means approach we used the standard K-means algorithm and for WSD based clustering we used the popular Lesk algorithm. Our results indicate that the WSD based clustering approach is better. More specifically, the WSD based clustering approach produces significantly better clustering results. The idea can overcome many language variation problems.

8. FUTURE WORK

For future research, we will focus on the following aspect: to consider semantically related words, for example, to treat 'beautiful', 'gorgeous' and 'pretty' as altogether similar terms, as these words have very similar meaning.

9. REFERENCES

[1] Jing, L., Ng, M. K., Yang, X., and Huang, J. Z., 2006. A Text Clustering System based on k-means Type Subspace Clustering and Ontology. International Journal of Intelligent Technology. 1(2), 91-103.

[2] Kumar, N. K., Santosh, K. G. S., and Varma, V. 2011. Multilingual document clustering using wikipedia as external knowledge. In Proceedings of the second International Conference on Multidisciplinary Information Retrieval Facility (IRFC' 11). Allan, H.; Rauber, A.; Vries, A. P. D. (Eds.). Springer-Verlag, Berlin, Heidelberg. 108-117.

[3] Liu, Y., Scheuermann, P., Li, X., and Zhu, X. 2007. Using WordNet to Disambiguate Word Senses for Text Classification. In Workshop on Text Data Mining in conjunction with 7th International Conference on Computational Science.

[4] Vijayalakshmi, S., Manimegalai, D. 2006. Query based Text Document Clustering using its Hyponymy Relation. International Journal of Computer Applications. 23(1)(June 2011), 13-16.

[5] Mavroeidis, D., Tsatsaronis, G., Vazirgiannis, M., Theobald, M., and Weikum, G. 2005. Word Sense Disambiguation for Exploiting Hierarchical Thesauri in Text classification. In Proceedings of the Ninth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Porto, Portugal, Springer. vol. 3721 of Lecture Notes in Computer Science. 181-192.

[6] Hidalgo, J. M. G., Rodriguez, M. D. B., and Perez, J. C. C. 2005. The Role of Word Sense Disambiguation in Automated Text Categorization. Montoyo, A.; Muñoz, R.; Métais, E. (Eds.), Natural Language Processing and Information Systems: 10th International Conference on

Applications of Natural Language to Information Systems, (NLDB' 05), Alicante, Spain, June 15-17. Proceedings, Lecture Notes in Computer Science, vol. 3513, Springer. 298-309.

[7] Steinbach, M., Karypis, G., and Kumar, V. 2000. A Comparison of Document Clustering Techniques. Proc. KDD-2000 Workshop on Text Mining, Aug 2000.

Table 1. VSM and K-means Clustering Approach Compared with Proposed Approach

Serial number	Document name	Word	Meaning or Sense	Extracted sentence	Sense ID	Cluster no. for the proposed approach	Cluster no. for the VSM and K-Means based approach
1	Doc1.txt	present	Give an exhibition	Students to present...	741	1	1
2	Doc2.txt	present	Give an exhibition	The National Underground...	741	1	4
3	Doc3.txt	present	Give an exhibition	For any presentation people ...	741	1	1
4	Doc4.txt	present	Give an exhibition	How to present …	741	1	4
5	Doc5.txt	present	Give an exhibition	At various happenings …..	741	1	1
6	Doc6.txt	present	Give as a present	What to give…	742	2	3
7	Doc7.txt	present	Give as a present	What is the best …	742	2	3
8	Doc8.txt	present	Give as a present	A gift or a present...	742	2	3
9	Doc9.txt	present	The period of time	The present is …	743	3	5
10	Doc10.txt	present	The period of time	The present tense...	743	3	5
11	Doc11.txt	present	The period of time	"had" is past …	743	3	5
12	Doc12.txt	present	The period of time	Convert a present...	743	3	4
13	Doc13.txt	present	Existing in a specified place	Only the present...	744	4	2
14	Doc14.txt	present	NF	NF	NF	5	2
15	Doc15.txt	present	NF	NF	NF	5	2

Table 2. Comparison with Clustering based on Human Intelligence [grey color points out misclassification]

Document name	Cluster no. for the VSM and K-Means based approach	Cluster no. for the proposed approach	Human Classified Cluster number	Error rate (with VSM and K-Means based approach) in %	Error rate (with proposed approach) in %
Doc1.txt	1	1	1	33.33 (5 misclassifications for 15 documents)	6.66 (1 misclassification only for 15 documents)
Doc2.txt	4	1	1		
Doc3.txt	1	1	1		
Doc4.txt	4	1	1		
Doc5.txt	1	1	1		
Doc6.txt	3	2	2		
Doc7.txt	3	2	2		
Doc8.txt	3	2	2		
Doc9.txt	5	3	3		
Doc10.txt	5	3	3		
Doc11.txt	5	3	3		
Doc12.txt	4	3	3		
Doc13.txt	2	4	4		
Doc14.txt	2	5	5		
Doc15.txt	2	5	6		

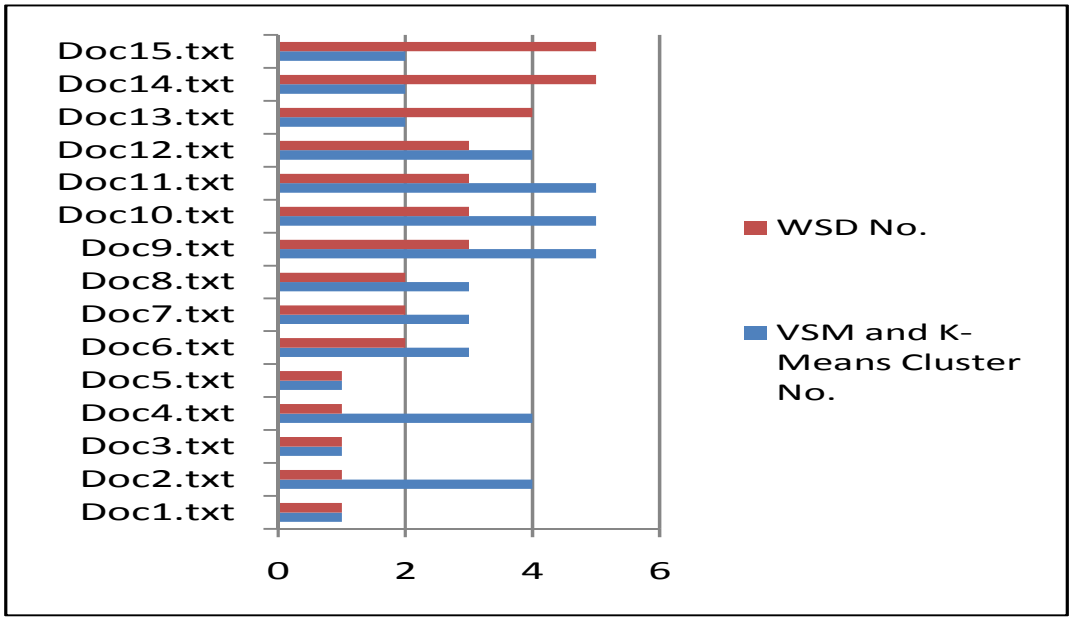

Fig. 2 Graphical representation of comparison between the two approaches compared