Domain Based Punjabi Text Document Clustering

Saurabh Sharma, Vishal Gupta
University Institute of Engineering & Technology, Panjab University, Chandigarh
saurabhsharma381@gmail.com, vishal@pu.ac.in

ABSTRACT

Text Clustering is a text mining technique which is used to group similar documents into single cluster by using some sort of similarity measure & separating the dissimilar documents. Popular clustering algorithms available for text clustering treats document as conglomeration of words. The syntactic or semantic relations between words are not given any consideration. Many different algorithms were propagated to study and find connection among different words in a sentence by using different concepts. In this paper, a hybrid algorithm for clustering of Punjabi text document that uses semantic relations among words in a sentence for extracting phrases has been developed. Phrases extracted create a feature vector of the document which is used for finding similarity among all documents. Experimental results reveal that hybrid algorithm performs better with real time data sets.

KEYWORDS: Natural Language Processing, Text Mining, Text Document Clustering, Punjabi Language, Karaka Theory.

Proceedings of COLING 2012: Demonstration Papers, pages 393–400,
COLING 2012, Mumbai, December 2012.

1. Introduction

The current study was undertaken specifically for clustering of text documents in Punjabi Language as no prior work has been done in this language as per review of literature done to carry out this study. It is an attempt in this direction to provide a solution for text clustering in Punjabi Language, by developing a new hybrid approach of text clustering, which will be immensely useful to the researchers who wish to undertake study and research in vernacular languages. The study proposed and implemented a new algorithm for clustering of Punjabi text documents by combining best features of the text clustering algorithms i.e. Clustering with frequent Item Sets, Clustering with Frequent Word Sequences, keeping in view the semantics of Punjabi language. Efficiency of the three algorithms for Punjabi Text Document Clustering was compared using Precision, Recall & F-Measure.

2. Proposed approach for Punjabi text clustering

Positional languages, which come in the category of Context Free Grammars (CFGs) have used popular approaches for text clustering. A context-free grammar is a formal system that describes a language by specifying how any legal text can be derived from a distinguished symbol called the axiom, or sentence symbol. It consists of a set of productions, each of which states that a given symbol can be replaced by a given sequence of symbols. The sentence structure of Punjabi is different as it belongs to the category of Free order language, unlike in English. Hence, features of free order languages were to be taken into consideration for clustering of Punjabi text.

Paninian framework, a technique for formalism, has been used for extraction of phrases for Indian languages. Karaka relation between verbs and nouns in a sentence is used to analyse the sentence. Sudhir K Mishra [2007] whose work focused on the theory of Karaka, (Panini : Adhikara sutra [Bharati, A. and Sangal, R. 1990]), for analyzing the structure of a sentence in Sanskrit Language, did the prominent work in this category.

Any factor that contributes to the accomplishment of any action is defined as karaka. Punjabi language identifies eight sub types like Hindi and Sanskrit [Bharti, A. and Sangal, R. 1993]. The karaka relations are syntactico-semantic (or semantico-syntactic) relations between the verbal and other related constituents in a sentence. They by themselves do not give the semantics. Instead they specify relations which mediate between vibhakti of nominals and verb forms on one hand and semantic relations on the other [Kiparsky 1982; Cardona 1976; Cardona 1988].

2.1 Pre-processing Phase

In text clustering, some techniques used in pre-processing are removal of punctuation marks, removal of stop words, stemming of words, normalization (where the same word exists in different spellings in case of multilingual words).

For pre-processing, the Algorithm takes Punjabi text documents as input. The first step in pre-processing comprises of removal of punctuation marks. Stop words are not removed, since Karaka theory [Bharati, A. and Sangal, R. 1990] is being used for generating phrases. Karaka theory works only on complete sentences, which necessarily includes, stop words. This does away with the requirement of removal of stop words.

Next step is normalization of those words which are used with different spellings. Purpose of normalization is to maintain uniformity of spelling in all documents which contain that word. This helps in better clustering results. Otherwise some documents may not be identified just because of the difference in spellings.

2.2 Algorithm Details

After the completion of pre-processing step, phrases are extracted from sentences with the help of karaka list. Karaka List is the collection of words which are used to specify role of words as nouns, verbs, objects and gives information about semantics of the sentence.

To overcome the drawback of Frequent Item Sets [Fung et. al. 2003] and Frequent Word Sequences [Li, Y. et. al. 2008] that generated long Sequences by trying all combinations of 2-word sequences using Apriori algorithm [Agrawal R. and Srikant, R. 1994], *Karaka* list is used in proposed approach.

Extraction of phrases from the document with the help of Karaka list generates a document vector containing phrases of various lengths in the same order in which they were originally in input document. This dissuades the computation of k-length sequences in number of steps by trying all possible combinations of (k-1)-length sequences.

2.2.1 Calculation of Term frequency of Phrases

Term Frequency is a numerical statistic which reflects how important a word is to a document in a collection. It is often used as a weighting factor in information retrieval and text mining. The value of Term Frequency increases proportionally to the number of times a word appears in the document which helps to control the fact that some words are generally more common than others. For each phrase, we calculate the Term Frequency, by counting the total number of occurrence in the document.

2.2.2 Finding top k Frequent Phrases

Sort all phrases by Term Frequency in descending order. Then declare top k phrases as Key phrases. These key phrases will be used for finding similarity among all other documents. The value of k is a very important factor for better clustering results. The valid value of k ranges from 1 to n, where n is number of phrases in a document. For experimental results, 20% of phrases are used as value of k.

2.2.3 Finding Similar Documents and Creating Initial Clusters

In this step, initial clusters are created by matching key phrases of documents with each other. If a phrase is found common between two documents, then it is assumed that these documents may belong to the same cluster. All matched documents will be searched for common Cluster Title by using Cluster Titles list.

The main idea of using Cluster Title List is to avoid overlapping clusters, meaningless or ambiguous titles of Clusters. To avoid this major drawback, manually created list of Cluster Titles for specific domain have been used. Text data has been taken for Sports domain. List of Cluster Titles specific to sports have been created manually as no such list is available in Punjabi language.

Documents with same Cluster Title are placed into same cluster. If two documents contain matching phrase but do not contain same Cluster Title, then it is assumed that both documents do not belong to same cluster.

2.2.4 Calculate term frequency of each term for each document and sort them to find top k frequent terms

After creating initial clusters, all those documents which are not placed in any cluster, are placed in a cluster named "Unrecognized". Since, some documents may contain cluster titles but did not appear in top k Frequent Phrases, for those unrecognized documents, VSM model is used [Salton et. al. 1975] i.e. now document is represented as a collection of single word terms obtained by splitting all phrases. The difference between Term and Phrase is that by splitting a single phrase of N word length, N different terms have been obtained. For each unrecognized document, Term Frequency for each term in the document is calculated. Then, all terms are sorted based on their Term Frequency in document, to find top k frequent terms of the document. The value of k can be varied as per the users' discretion from 5%, 10%, 20% and so on. Higher the value of k, more terms will be considered for finding cluster for unrecognized document.

2.2.5 Find cluster frequent terms for new clusters

After calculating top k frequent terms for each unrecognized document. Now, top k Cluster Frequent Terms for each cluster will be identified. Term Frequency of each term of the conceptual document is calculated.

2.2.6 For each unrecognized document assign a cluster

Cluster for unrecognized document, by matching top k Frequent Terms of document with top k Cluster Frequent Terms of each document, is identified.

2.2.7 Final Clusters

After processing of unrecognized documents, final clusters containing documents from initial cluster and documents from unrecognized documents are created.

3. Experimental Evaluation

Natural Classes	F-Measure
ਹਾਕੀ (Hockey)	0.99
ਕ੍ਰਿਕਟ (Cricket)	0.93
ਟੈਨਿਸ (Tennis)	0.87
ਫੁਟਬਾਲ (Football)	0.93
ਬੈਡਮਿੰਟਨ (Badminton)	1.00
ਮੁੱਕੇਬਾਜ਼ੀ (Boxing)	0.89

Table 1. Natural Classes for Hybrid Approach

3.1 Data Set

The text documents are denoted as unstructured data. It is very complex to group text documents. The document clustering requires a pre-processing task to convert the unstructured data values into a structured one. The documents are data elements with large dimensions. The system was tested for 221 text documents collected from various Punjabi News websites which comprised of news articles on sports. This dataset was categorized into 7 Natural classes (see table 1), which were used for the evaluation of all three algorithms.

3.2 Experimental Results and Discussion

To evaluate the accuracy of the clustering results generated by clustering algorithms, F-measure is employed. Let us assume that each cluster is treated as if it were the result of a query and each natural class is treated as if it were the relevant set of documents for a query. The recall, precision, and F-measure for natural class K_i and cluster C_j are calculated as follows:

$$\text{Precision}(K_i, C_j) = n_{ij} / |C_j| \quad (2)$$

$$\text{Recall}(K_i, C_j) = n_{ij} / |K_j| \quad (3)$$

$$\text{F-Measure}(K_i, C_j) = \frac{2 * [\text{Precision}(K_i, C_j) * \text{Recall}(K_i, C_j)]}{[\text{Precision}(K_i, C_j) + \text{Recall}(K_i, C_j)]} \quad (4)$$

where n_{ij} is the number of members of natural class K_i in cluster C_j . Intuitively, $F(K_i;C_j)$ measures the quality of cluster C_j in describing the natural class K_i, by the harmonic mean of Recall and Precision for the "query results" C_j with respect to the "relevant documents" K_i.

In fig.1 the graph plotted for Precision, Recall and F-Measure for all the three algorithms that were studied for clustering of Punjabi text documents, the two algorithms namely, Frequent Itemset and Frequent word sequence, shows a good precision but a very poor recall value. This leads to a very low value of F-Measure which is indicative of its overall poor performance. On the other hand, Hybrid algorithm that shows good Precision, Recall and F-Measure, outperform other two algorithms and hence generate best clustering results.

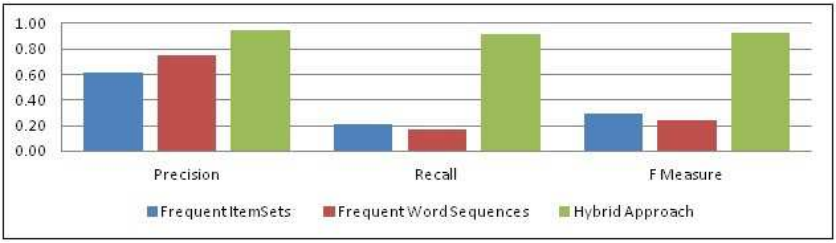

Fig 1 Precision, Recall and F-Measure

3.3 Error Analysis

During the development of this algorithm, several problems for improving clustering results were encountered. These problems and reason for errors in clustering result are discussed below.

Different Spellings in Different Documents results in True Negative. In case of words, which are originally from other languages than the one under purview, e.g. English word 'football' can be written as ਫੁਟਬਾਲ or ਫੁੱਟਬਾਲ. Now, during clustering phase, efforts are made to find similarity between two documents about football, but having different spellings, that do not match. To overcome this problem, we have used normalization of Cluster Titles in pre-processing step.

Phrases containing Important Terms but not coming in Top k Frequent Phrases, results in True Negatives. For example, a document contains news about football. But word 'football' is appearing only one or two times in whole document, then it is very hard to capture this desired information in top k Frequent phrases. To overcome this problem, VSM approach is utilized after creating Initial clusters. In this step, top k Frequent Terms are identified. Advantage of applying this step is utilizing those meaningful terms which are not captured in top k Frequent phrases, but very vital for efficient, effective & correct clustering of documents.

Multiple Key Phrases matches with Multiple Cluster Titles results in False Positive and True Negative. For example, a document contains an article on football, but uses some terms common with other sports e.g. team, goal, match referee etc. then it becomes difficult to identify the exact cluster for the document. To overcome this problem, the number of matching Cluster frequent Terms are counted for each matching cluster. Document is, then, placed in that cluster which has maximum number of matching Cluster frequent Terms.

4. Conclusion

Domain based Punjabi Text clustering software is logically feasible, efficient and practical for Punjabi text documents. It is more feasible and has a better performance than Frequent Itemsets and Frequent Word Sequences with reference to Punjabi Text Documents. The results are validated and drawn from the experimental data. This approach focuses on the semantics of a sentence. Proposed work shows better results as it uses a list of Cluster Title candidates, which does not allow the construction of huge number of clusters with meaningless names. This algorithm was not tested with benchmark data set, because all available data sets are for English language only. Dataset for Punjabi language is created manually because no such benchmark dataset is available for Punjabi language.

References

AGRAWAL R. AND SRIKANT, R. (1994)). *Fast Algorithms for Mining Association Rules*. In Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. , 487 - 499. ISBN:1-55860-153-8.

BHARATI, A. AND SANGAL, R. (1990). *A karaka based approach to parsing of Indian languages*. In Proceedings of the 13th conference on Computational linguistics. Association for Computational Linguistics Stroudsburg, PA, USA. 3, 25-29. ISBN:952-90-2028-7 doi>10.3115/991146.991151

BHARATI, A. AND SANGAL, R. (1993). *Parsing free word order languages in the Paninian framework*. In Proceedings of the 31st annual meeting on Association for Computational Linguistics. Association for Computational Linguistics Stroudsburg, PA, USA. 105-111. doi>10.3115/981574.981589

CARDONA, G. (1976). *Panini: A Survey of Research*, Mouton, Hague-Paris.

CARDONA, G. (1988). *Panini: His Work and Its Tradition* (Vol. 1: Background and Introduction), Motilal Banarsidas, Delhi.

FUNG, B.C.M., WANG, K. AND ESTER, M. (2003). *Hierarchical Document Clustering Using Frequent Itemsets*. In Proceedings of SIAM International Conference on Data Mining.

KIPARSKY, P. (1982). *Some Theoretical Problems in Panini's Grammar*. Bhandarkar Oriental Research Institute, Poona, India.

LI, Y., SOON M. CHUNG, S. M. AND HOLT, J. D. (2008). *Text document clustering based on frequent word meaning sequences*. Data & Knowledge Engineering, 64, 1, 381-404.

MISHRA, S. K. (2007). *Sanskrit Karaka Analyzer for Machine Translation*. M.Phil dissertation. Jawaharlal Nehru University, New Delhi.

SALTON, G., WONG, A. AND YANG, C. S. (1975). *A vector space model for automatic indexing*. Communications of the ACM. ACM New York, NY, USA. 18, 11, 613 - 620. doi>10.1145/361219.361220