A Cluster based Approach with N-Grams at Word Level for Document Classification

Apeksha Khabia
M. Tech Student
CSE Department
SRCOEM, Nagpur, India

M. B. Chandak
Associate Professor and Head
CSE Department
SRCOEM, Nagpur, India

ABSTRACT

A breakneck progress of computers and web makes it easier to collect and store large amount of information in the form of text; e.g., reviews, forum postings, blogs, web pages, news articles, email messages. In text mining, growing size of text datasets and high dimensionality associated with natural language is great challenge which makes it difficult to classify documents in various categories and sub-categories. This paper focuses on cluster based document classification technique so that data inside each cluster shares some common trait. The common approach for document clustering problem is bag of words model (BOW), where words are considered as features. But some semantic information is always lost as only words are considered. Thus we aim at using vector-space model based on N-grams at word level which helps to reduce the loss of semantic information. The problem of high dimensionality is solved with feature selection technique by applying threshold on feature values of vector space model. The vector space is mapped into a modified one with latent semantic analysis (LSA). Clustering of documents is done using k-means algorithm. Experiments are performed on Stack Exchange data set of some categories. R is used as text mining tool for implementation purpose. Experiment results show that tri-grams give better clustering results than words and bi-grams.

General Terms

Data Mining, Text Mining

Keywords

Document clustering, N-grams at word level, dimensionality reduction, Latent Semantic Analysis

1. INTRODUCTION

Today advances in information and communication technologies offer ubiquitous access to large amounts of information and are causing an exponential increase in the number of text documents available on web. As more and more textual information is available electronically, effective classification of these text documents is getting more and more difficult with the growing size of datasets and high dimensionality associated with text data. Among different approaches used to tackle classification problem, clustering based classification of text data is of great importance and enabling approach. The main idea is to perform text clustering followed by classification with trained clusters with selected features [1].

Generally, given a collection of text documents, document clustering is automatic grouping of documents together in such a way that the documents within each cluster are similar to each other. Traditional clustering methods include k-means and its variants.

For document clustering, the Vector-space model can be used. In the vector-space model [2], each document is considered as a vector, where vector components represent certain feature weights. Traditionally, components of vectors are unique words. However, it has the challenge of high dimensionality. k-means type algorithms are more efficient than hierarchical algorithms [3] for document clustering as they are significantly computational effective when dataset is large with high dimensions.

Traditionally, in vector-space model unique words are considered as the components of vectors, that is bag of words (BOW) model is used. Another approach is to use N-grams as the vector components instead of only unique words. An N-gram is a sequence of symbols extracted from a long string [4]. These extracted symbols can be a byte, character, or word. For extracting character N-grams from a document N-character wide window is to be moved across the document character by character. The advantage of the character N-gram representation is that, it is more robust and less sensitive to grammatical and typographical errors, and also it requires no linguistic preparation, making it more language independent than other representations. Word based N-grams can be extracted from document by moving window of length N-words across the document word by word. The collection of only words cannot capture phrases or multi-word expressions, while word based N-grams have shown to be helpful features in several text classification tasks [5, 6, 7]. We have used word based N-gram for building vector-space for documents and reviewed the distance measure to make it suitable for document clustering.

When any one of these approach is used for representing vector space model, it is not surprising to find thousands or tens of thousands of different words or N-grams, of which a very small subset appears in an individual document, even for a relatively small sized text document collection of a few thousand documents. As a result, very sparse and very high-dimensional feature vector is formed for describing a document. Because of high dimensionality of the feature vector, dimension reduction technique is to be applied to feature vector of N-gram. Various dimension reduction techniques are proposed in [8, 9]. Dimensions can be reduced by selecting features from feature vector above some threshold value.

LSA combines the classical vector space model with a Singular Value Decomposition (SVD), a two-mode factor analysis. Thereby, bag-of-words representations of texts can be mapped into a modified vector space that is assumed to reflect semantic structure [10]. Thus, vector space model based on N-gram can be mapped to vector space with LSA, so that more semantic information will be captured which can help while clustering with k-means.

This paper is organized as follows. Section 2 gives general background of N-gram representation, dimensionality reduction technique, document clustering by k-means. Section 3 presents the proposed system for cluster based classification with N-grams. Section 4 provides the experimental setup and procedure followed by experiment results, while section 5 gives the conclusion and future research directions.

2. RELATED WORK

There is a variety of work that has been carried out by researchers in the field of document clustering. In [11] real world example is given where the k-means clustering technique is applied on questions and answers of Stack Overflow website using Apache Mahout. Once grouped, a common picture of stackoverflow data with relationships between questions can be seen [11]. The Yingbo Miao et. al. [12] have proposed the novel method for document clustering using most frequent character N-grams and compared the results with term based and word based document clustering. A systematic study is conducted for document representation with word, multi-word term and character N-grams by Mahdi Shafiei et. al. [13]. They also studied three methods for dimensionality reduction i.e. independent component analysis, latent semantic indexing and document frequency. Other works are also present that shows examples on document clustering and preprocessing operations in text mining [14, 15, 16].

3. THE PROPOSED SYSTEM

The proposed system consists of the following steps: Text pre-processing, morphological analysis using N-grams, vector space model of document and N-gram, then dimensionality reduction by applying threshold to feature vector and then vector space model based on N-gram is mapped with LSA space. Finally, applying k-means clustering to modified vector space and obtain the cluster of documents. In the end, these document clusters are used for training in classification step. These all steps are shown in figure 1.

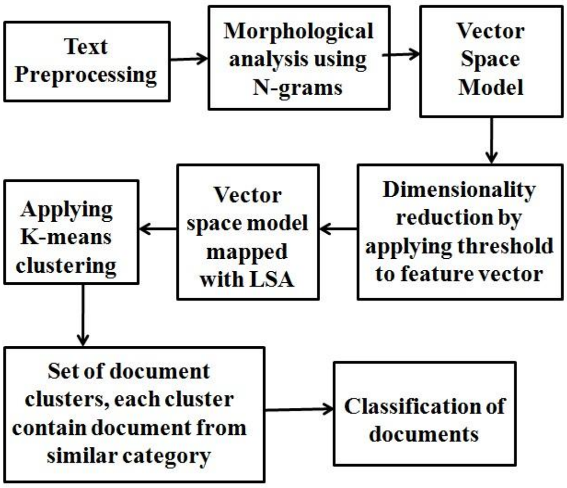

Text Preprocessing
Morphological analysis using N-grams
Vector Space Model
Dimensionality reduction by applying threshold to feature vector
Vector space model mapped with LSA
Applying K-means clustering
Set of document clusters, each cluster contain document from similar category
Classification of documents

Figure 1: Various Steps of System

3.1 Pre-processing of Text Dataset

Document clustering one of task of text data mining deals with unstructured data, a large amount of data is stored as unstructured or semi-structured text format like in books, research papers, news articles, blogs, web pages, email messages, XML documents etc. For obtaining the structured format from unstructured long sequence of operations are needed such as converting to lower case, punctuations removal, stop words removal, URL removal, stemming, and white space removal. These operations are known as the pre-processing operations.

3.1.1 For 1-gram Representation

This is typical practice for text document representation as bag of words. The dimensions of vector space model comprised of documents on one side and these bag of words as other dimension. For this purpose of extracting words from text document, various operations of text processing are applied, followed by removal of stop words and stemming of remaining words.

3.1.2 For n-gram Representation

The sequence of symbols extracted from a long string is N-gram. This sequence of symbols can be a character, byte or word. If word sequence is taken into account, semantic of text is better captured. Thus we are considering word level N-grams, adjacent N words acts as N-grams. That means bi-grams, tri-grams, etc. can be retrieved. For N-grams representation stop words are not removed and stemming is also not performed. Thus N-gram representation would be less sensitive to typographical and grammatical errors.

3.2 Vector Space Model with Feature Weights

In order to get processed by document clustering algorithm, the text document dataset should be represented using an appropriate numerical model. For this vector space model is represented with suitable feature weights using a term weighting technique. We have used the term frequency inverse document frequency (TFIDF) weighting scheme as term weights. Also the TFIDF weightings are normalized to unity. TFIDF combines both document frequency and term frequency.

3.3 Dimensionality Reduction

Several N-grams are formed even from small length of text document. Thus, N-gram representation model have huge number of features. For the sake of computation time and complexity dimension reduction is necessary task. As we are performing document clustering, dimensions are reduced from feature vector only, that means we are reducing the number of features to be used for clustering. Features are selected, by applying threshold on TFIDF values of vector space model. The N-grams which have more total TFIDF weight in the text document collection are selected; half of the total N-grams are selected as featured for document clustering. Thus dimensions are reduced up to 50%.

3.4 Vector Space Mapping with LSA

Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text [17]. LSA uses singular value decomposition. The document term matrix is then decomposed via singular value decomposition into: term vector matrix comprising of left singular vectors, the document vector matrix comprising of right singular vectors and the diagonal matrix comprising of singular values. The basic idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints. To the great extent these constraints determines the similarity of meaning of words and sets of words to each other. So the reduced document term matrix is mapped to new vector space with LSA. So, terms and documents that are closely associated are placed near one another.

3.5 Document Clustering

Document clustering groups the text document collection into different groups such that documents within the same group share similar features. For document clustering k-means algorithm is used. k-means algorithm is based on the unsupervised learning approach. When k-means is used for document clustering documents are automatically partitioned into k different clusters. To determine the value of k to be used while clustering, sum of squared error method is used. In this we use that value of k at which the within group sum of squares (WSS) distance is less, from the calculated value of WSS for different k, because the main goal of clustering is to minimize the WSS distance.

4. IMPLEMENTATION DETAILS AND DISCUSSIONS

We implemented the system using R statistical and machine learning tool. In this section, at first dataset is described, then the pre-processing procedure for word based and N-gram based representation are described followed by description of dimensionality reduction procedure and clustering results when applied to both representation as explained in the system description.

4.1 Dataset Description

For implementation purpose text dataset from Stack Exchange [18] website is used. Stack Exchange is a network of question and answer websites on multiple topics in different fields. The Stack Exchange data dump is present at the internet archive. Each site can be downloaded individually, and includes an archive with Posts, Users, Votes, Comments, Badges, PostHistory, and PostLinks. As we need text dataset, we have used Posts and Comments. These files are present in XML format. The topics with more similar semantics are chosen, like Astronomy, Aviation, Earth Science and Space.

4.2 Pre-processing of text dataset with R

R is a statistical analysis and machine learning tool which is used for performing various pre-processing operations on text document dataset. XML documents are parsed to plain text with the help of ‘XML’ package of R and posts and comments are extracted. ‘tm’ package of R helps to carry out pre-processing operations on text document corpus. For word based representation, text pre-processing operations performed are converting to lower case, removal of URLs, removal of stop words, removal of punctuations, removal of extra white space followed by stemming of words. While in case of N-gram based representation stop words are not removed and word stemming is not performed.

4.3 Document clustering with R

Now the document term matrix is formed for word, bi-gram and tri-gram with appropriate tokenization control. Normalized TFIDF weighting scheme is used as feature weights. Then the dimensions of document term matrix are reduced by reducing the number of features (that are word, bi-gram or tri-gram). The dimensions are reduced up to 50%. The reduced dimensionality for all three matrices is presented in following table.

Table 1: Results of dimension of features in case of all three representations (word, bi-gram, tri-gram) of document term matrix before and after dimension reduction

	Original number of featues	Number of features after dimension reduction	Percentage decrease in number of featues
Word	49403	24613	50.17
Bi-gram	557535	276609	50.38
Tri-gram	1237146	614268	50.35

After this the value of k to be used in k-means clustering algorithm is determined from the plot of number of clusters (different values of k) to within group sum of squares for particular k.

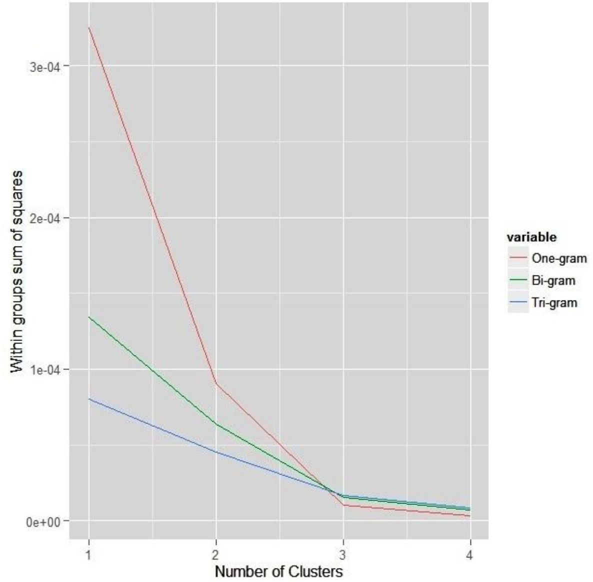
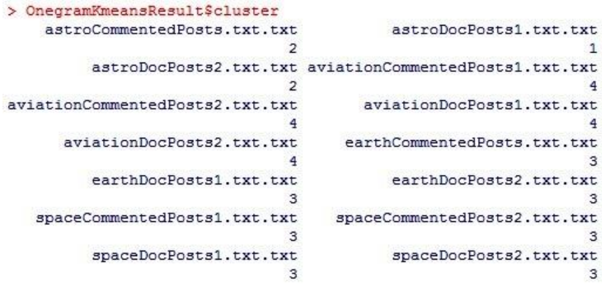


Figure 2: Graph between different values of k and within group sum of squares to determine the value of k, for each word, bi-gram and tri-gram

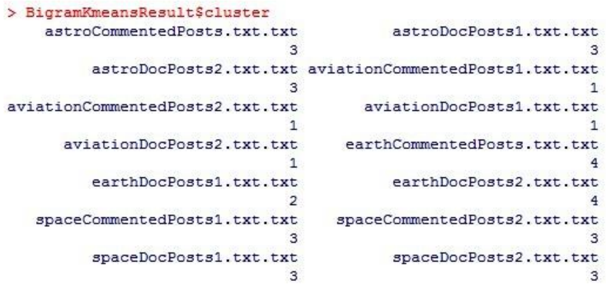
In above figure 2, minimum value of WSS for all three representations scheme is at 4. So the value of k to be passed to k-means is 4.

Then k-means clustering is applied on each document term matrices of three representations. The results shows that when clustering is performed by using tri-gram representation of vector space model, documents are clustered properly with documents of same category in same clusters and documents of different category in different clusters. For other representations that is, word and bi-gram documents are not properly clustered. Results for one-gram, bi-gram and tri-gram are shown in following figure 3, 4 and 5.



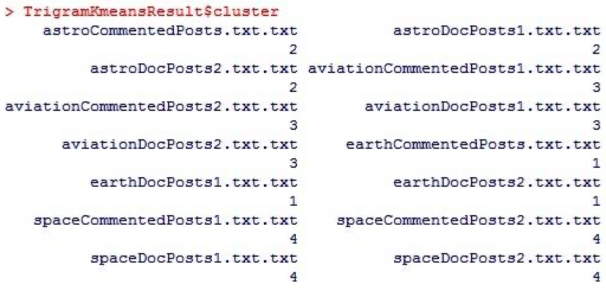
```
> OnegramKmeansResult$cluster
    astroCommentedPosts.txt.txt          astroDocPosts1.txt.txt
                              2                               1
         astroDocPosts2.txt.txt aviationCommentedPosts1.txt.txt
                              2                               4
aviationCommentedPosts2.txt.txt       aviationDocPosts1.txt.txt
                              4                               4
      aviationDocPosts2.txt.txt     earthCommentedPosts.txt.txt
                              4                               3
         earthDocPosts1.txt.txt          earthDocPosts2.txt.txt
                              3                               3
    spaceCommentedPosts1.txt.txt    spaceCommentedPosts2.txt.txt
                              3                               3
         spaceDocPosts1.txt.txt          spaceDocPosts2.txt.txt
                              3                               3
```

Figure 3: Clustering result for vector space model based on one-gram



```
> BigramKmeansResult$cluster
    astroCommentedPosts.txt.txt          astroDocPosts1.txt.txt
                              3                               3
         astroDocPosts2.txt.txt aviationCommentedPosts1.txt.txt
                              3                               1
aviationCommentedPosts2.txt.txt       aviationDocPosts1.txt.txt
                              1                               1
      aviationDocPosts2.txt.txt     earthCommentedPosts.txt.txt
                              1                               4
         earthDocPosts1.txt.txt          earthDocPosts2.txt.txt
                              2                               4
   spaceCommentedPosts1.txt.txt    spaceCommentedPosts2.txt.txt
                              3                               3
         spaceDocPosts1.txt.txt          spaceDocPosts2.txt.txt
                              3                               3
```

Figure 4: Clustering result for vector space model based on bi-gram

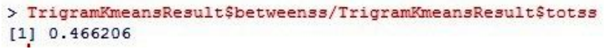


```
> TrigramKmeansResult$cluster
    astroCommentedPosts.txt.txt          astroDocPosts1.txt.txt
                              2                               2
         astroDocPosts2.txt.txt aviationCommentedPosts1.txt.txt
                              2                               3
aviationCommentedPosts2.txt.txt       aviationDocPosts1.txt.txt
                              3                               3
      aviationDocPosts2.txt.txt     earthCommentedPosts.txt.txt
                              3                               1
         earthDocPosts1.txt.txt          earthDocPosts2.txt.txt
                              1                               1
   spaceCommentedPosts1.txt.txt    spaceCommentedPosts2.txt.txt
                              4                               4
         spaceDocPosts1.txt.txt          spaceDocPosts2.txt.txt
                              4                               4
```

Figure 5: Clustering result for vector space model based on tri-gram (properly clustered)

The goal of cluster analysis is that the objects within a group be similar to one another and different from the objects in other groups. The greater the similarity within a group and the greater the difference between groups, the better or more distinct is the clustering [19]. So, the measure of goodness of the classification by k-means has been defined by the ratio between SS (BSS) to total SS (TSS), where SS stands for Sum of Squares. That means ideally we want a clustering that has the properties of internal cohesion and external separation, i.e. the BSS/TSS ratio should approach to 1.

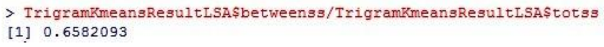
In the results of above k-means clustering procedure on tri-grams, the ratio of BSS/TSS found is 0.466206 as shown in following figure 6.



```
> TrigramKmeansResult$betweenss/TrigramKmeansResult$totss
[1] 0.466206
```

Figure 6: BSS/TSS ratio when k-means is applied on tri-gram representation model.

Now we calculated the LSA space from the previous vector space model of word level and applied the k-means clustering to this new mapped vector space model. The text documents are clustered appropriately and also the ratio of BSS/TSS has increased (shown in figure 7)



```
> TrigramKmeansResultLSA$betweenss/TrigramKmeansResultLSA$totss
[1] 0.6582093
```

Figure 7: BSS/TSS ratio when k-means is applied on modified vector space model mapped with LSA.

5. CONCLUSION AND FUTURE SCOPE

Classification problems on text data mainly focus on feature space that is words and semantics between these words. Also we know that there is great issue of growing size of text and high dimensionality in text mining. In this paper, we have applied k-means clustering algorithm using N-grams on some semantically similar categories of dataset from Stack Exchange website. For dimensionality reduction, feature selection technique is used, by applying threshold on TFIDF values of vector space model. Our experiments were conducted on four categories; Astronomy, Aviation, Earth Science and Space. The implementation of system was carried out with R tool. Results demonstrated that:

- R tool is helpful for all pre-processing operations, k-means clustering.
- Vector space model based on tri-gram with word level gives more accurate cluster results than word and bi-gram. This is because with the help of N-grams semantics of text are better captured, so that text documents are accurately clustered.
- Semantics of text can be better captured with LSA by purely statistical computation. Thus clustering results of modified vector space model mapped with LSA progress towards goodness.

In future, each category of documents can further be divided into sub-categories. For this classification semantics of text will be a great challenge as most of the words inside a category are semantically similar.

6. ACKNOWLEDGEMENTS

The authors are grateful to the Principal, Shri Ramdeobaba College of Engineering and Management for providing adequate facilities to conduct a research. Authors are thankful to the faculty members of Computer Science and Engineering Department for their continuous support and cooperation during the work.

7. REFERENCES

[1] Khabia A., Chandak M. B., “A Cluster Based Approach for Classification of Web Results”, International Journal of Advaanced Computer Research, December 2014. Vol. 4, No. 4, Issue 17.

[2] Salton G., Buckley C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management. Vol. 24, No. 5, Pages 513–523.

[3] Agrawal C. C., Zhai C. 2012. A Survey of Text Clustering Algorithms. In:Mining Text Data. Springer US. ISBN: 978-1-4614-3222-7 (Print) 978-1-4614-3223-4 (Online).

[4] Canvar W. B. 1994. Using an n-gram-based document representation with a vector processing retrieval model. In TREC. Pages 269–278.

[5] Tan C., Wang, Y., and Lee, C., “The use of bigrams to enhance text categorization”, Journal of Information Processing and Management, 2002.

[6] Wang S. I., Manning, C. D. 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In Proceedings of ACL.

[7] Lin D., Wu, X. 2009. Phrase clustering for discriminative learning. In Proceedings of ACL.

[8] I. K. Fodor. 2002. A survey of dimension reduction techniques. Technical Report UCRL-ID-148494. Center for Applied Scientific Computing. Lawrence Livermore National Laboratory.

[9] Y. Yang, J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, Proceedings of ICML. 14th International

Conference on Machine Learning. Pages 412–420. Nashville, US.

[10] Wild F., Stahl C. 2006. Investigating Unstructured Texts with Latent Semantic Analysis. In Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V. Springer. Berlin Heidelberg.

[11] Owen S., Anil R., Dunning T., Friedman E. 2012. Real-world applications of clustering. In: Mohout In Action. Manning Publications, Shelter Island.

[12] Yingbo M., Vlado K., Evangelos M. 2005. Document Clustering using Character Ngrams: A Comparative Evaluation with Termbased and Wordbased Clustering. In the proceedings of the 14th ACM international conference on Information and knowledge management (CIKM). Pages 357-358. ISBN:1-59593-140-6.

[13] Mahdi S., Singer W., Roger Z, Evangelos M, Bin T., Jane T., Ray S. 2007. Document Representation and Dimension Reduction for Text Clustering. 23rd International Conference on Data Engineering Workshop. IEEE. Pages 770 – 779.

[14] Zho Y. 2012. R and Data Mining: Examples and Case Studies. Elsevier. http://www.rdatamining.com/

[15] Feinerer I., Hornik K. 2014. Text Mining Package. http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf.

[16] Stewart B. M. 2010. Practical Skills for Document Clustering in R*. http://faculty.washington.edu/jwilker/tft/Stewart.LabHandout.pdf

[17] Landauer T., Foltz, P., and Laham, D. 1998. Introduction to Latent Semantic Analysis. In: Discourse Processes 25, Pages 259–284.

[18] http://creativecommons.org/licenses/by-sa/3.0/legalcode

[19] Tan P., Steinbach M., Kumar V. 2006. Introduction to Data Mining. Errata.