(IJISE) 2018, Vol. No. 7, Jan-Jun **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

SINGULAR VALUE DECOMPOSITION: EMPLOYABILITY OF INDEPENDENT COMPONENT ANALYSIS –TOPIC DETECTION, CLUSTERS, LATENT SEMANTIC INDEXING IN ENHANCING DATA INTELLIGENCE USABILITY

Karan Gupta

Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Haryana,131039

ABSTRACT

Concept detection plays an important role if there is a huge amount of data available. We know that cluster analysis, topic detection, opinion mining has got a major role in the product marketing, online shopping, E-commerce. In this paper, we have conducted the topic detection and clustering experiments on the News samples which were sourced from online newspapers. Our aim is to find out the topics which also available in the text documents as a group of words and apply a clustering technique using the Singular value decomposition method. Then opinions are extracted from the comments, collected on a particular subject of interest like the comments for Smartphone. Finally, the clustering technique is applied on these sentiments to figure out the opinions of the people towards different features of the Smartphone. The results obtained here are competitive with the technology available.

Keywords- *Singular value decomposition; Topic detection; Clusters; Sentiments; Term document matrix; Latent semantic Indexing; Independent Component analysis; Machine learning.*

In the modern world of mobile devices and social networks, data are generated at an unprecedented rate. Users consume, generate and search for information on the internet. Timely and relevant information access is a major requirement for users to help with the information overload. Though this is a desirable requirement, it is not easy to achieve given the unstructured nature of information The Significant portion of the desired information is in the form of text. Users have to wade through documents to get to what they are looking for.

A keyword-based search solves the problem, but if the exact keyword is not known, information extraction becomes difficult. Semantic understanding of the textual data source becomes important. Businesses would like to know what consumers are commenting on their product. They would like to know what the consumers are complaining about and what the consumers are praising. Sources like Facebook, Twitter provide a wealth of information where the users comment on their experiences with products.

Our goal is to go beyond sentiment analysis and extract semantic information from the user comments. Ideally, semantic extraction in here is the process of taking out the important topics which are hidden in the news also separating out the documents and making the groups with their similarity called as Clustering. Once the clustering is done, we come with the different clusters of documents which similar to one another in some topic, with this we can achieve semantic discovery in terms of topics of same clusters will put aside. The text is a major source of information disseminated on the web.

Users also generate text-based data in the form of comments, feedback, and blogs. Analysis of text data is useful for domains like product marketing, after-sales service, campaign management. The goal of the current project is to extract semantic information from a given text data corpus. We intend to explore techniques for clustering, ontology extraction to derive semantic information.

The organization of the paper rest is as: Section II speaks about the literature survey made, followed by section III, which deals with the mathematical background required for the experiments, followed by section IV showing the experiments conducted, followed by section V where the results and discussions are kept and finally we have concluded in section VI.

RELATED WORK

Topics are in general the group of the words which are distributed over the documents. Latent Dirichlet (LDA) Allocation is one of the Topic models which say the above-mentioned sentence. Also, the process of fetching these topics from a mixture of documents is ideally NP-hard, [1] shows the first implementable algorithm for this problem. [2] Gave the solution using the tensor-clever method. Singular value decomposition (SVD) can be applied to information retrieval as Latent Semantic Indexing (LSI) [3].

Clustering is one of the classical approaches and upon which there is some concept detection research work based [4] [5]. In [6] [7] [8], different uses of Independent Component Analysis (ICA) has been discussed but comparison with other clustering strategies has not done.

while working assumes that every cluster within the document has got a precise shape, on the other hand, Naïve Bayes method assume that all the vectors of the feature vector space which represents the whole document corpus are not dependent on one another. There are different methodologies which could perform clustering based on Latent Semantic Indexing (LSI)[12], where this method all the documents are projected on a vector space by using SVD, and later which conducts clustering using the clustering algorithms which were traditional.

(IJISE) 2018, Vol. No. 7, Jan-Jun **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

MATHEMATICAL BACKGROUND

In this section we will discuss all the significant equations and formulas which are necessary to conduct our experiments. There is a little knowledge about the techniques like singular value Decomposition, Term Frequency (tf), Inverse Document Frequency (IDF), Cosine similarity some accuracy methods are very much needed.

A. Singular value Decomposition

Linear algebra has got a decomposition method in which the matrix factorization will take place. Before we proceed with SVD, Eigenvector, Term vector and Document Vector need to define. From the theory of Linear Algebra, we can define the SVD as It is a matrix decomposition of the form;

$X=U\sum V^T$

Where d represents the documents and q represent the query. In here, every dimension relates to a term.

The term is generally a keyword or phrase; the dimensionality here in a vector space is the total number of the words presents in the vocabulary. This vector operation application comes when there is a need for a comparison of two documents.

B. Tf-Idf

Term frequency (Tf), is the number of times the particular word appears in a document [13]. The weight of the term in a document is directly proportional to that of the term count or frequency, which is represented as tf (t, d);

$$f(t,d)=\sum_{x\in d}^{\cdot} fr(x,t)$$

Where its value will be 1 if x=-1 and otherwise 0. Inverse document frequency shows that the importance of word the particular document, it can be treated as the log scale function which contain the word, we can formulate it as;

$$idf(t,D)=\log\frac{N}{\left|\{d\in D:t\in d\}\right|}$$

Where N is the number of the text documents which are present in the data corpus, number of property how they are built with, i.e., essentially the first column of the matrix represents the terms which tends to occur in all of the documents, which is called as the Term vector.

Rows of the V matrix are arranged in such a way that, vectors documents when term t appears. If the term is absent in the document, its value leads to a divided by zero. So it is quite common to set the denominator.

$$tf \cdot idf(t,d,D) = tf(t,d) \times idf(t,D)$$

represent the all the documents in the data text corpus into the Vector Space are called as the Document vectors.

C. Cosine Similarity

By considering the inner product in the space between two vectors as the documents, we find the similarity between these documents is ideally called as the cosine similarity. Usually this is done at the vector level with which the documents are represented [14].

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

It can be calculated by using the dot product formulas like Euclidean (7) and (8):

7

$$\frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_I \times B_I}{\sqrt{\sum_{i=1}^{n} (A_{i=1}{}^2)} \times \sqrt{\sum_{i=1}^{n} (B_{i=1}{}^2)}}$$

8

Where attributes of the vectors are A, B and cos( ) is represented as the dot product with the use of magnitude.

PROPOSED METHODOLOGY

In this section, we will explain the methods and algorithms used for our experiments. Basically the process starts with the corpus of the documents, i.e. Newspaper samples which are collected from Internet sources. As we wanted to know the features or Topics in other words, which are hidden in documents, we carried out the procedure which is given in the below algorithms. We have also found the opinions for a specific product of interest and applied a clustering technique

to find out the sentiments of the customers towards specific features of a product. We can formulate this section into subsections of the experiments.

A. Concept Detection

Assume we know that the text documents are the collection of topics which are distributed in the documents. Using the SVD functionality we applied the steps. We know that by the definition of SVD, U matrix which is the representation of term vectors, so these terms will contribute their information to the documents. We assumed that the generation of U matrix is with respect to some concepts, ideally the number of columns which represents a number of the concepts. So we extracted the terms of highest value for a particular concept and in turn which makes a group of the words for a particular concept which is distributed in the documents. Finally, we went through the terms of a particular column and we collected the higher valued terms which are depicted as the words which are making particular concepts. These are explained with the outputs generated and showed in the results section.

B. Clustering

It can be treated as the unsupervised classification, where without any role model we are able to classify the related documents and kept aside with the similarity in their text content Here for the clustering, we are not applying the technique directly to the text documents; instead, we are considering the Hierarchical Clustering approach. We consider the document vectors which are generated by the SVD method for the corpus of documents, where each vector is documented vector and which are drawn in space model. Then we calculate the cosine value between these two vectors. Before we apply this, we consider the highest ordered singular values from the SVD, i.e. we look into the singular value matrix and consider that many rows in the V matrix as the document vectors and this process we call as the dimension reduction. Instead of considering all the vectors, we consider only those are giving their highest contribution to the document corpus.

In here, for the calculation of the similarity using the agglomerative approach we have used the Apache Lucene open sources [15]. This is based on the Hierarchical clustering.

C. Opinion mining and Cluster analysis

As we have already discussed the opinion mining in the above section, here we are aimed to find out the sentiments which are hidden in comments given by the customer for a particular product or subject. For this purpose, we collected comments for Smartphone and applied the below strategy.

The below algorithm shows the process of analysis of the opinions in the comments

Here we have used Stanford core Natural Language processing (NLP) open sources for our reference in coding [16]. We collected the comments from various sources like Twitter and also Online Newspaper websites. Once we have comments lines, these individual lines we made them get placed in an individual text document. So now, these documents in the collection form a corpus of the data where we put them for separation of positive and negative comments.

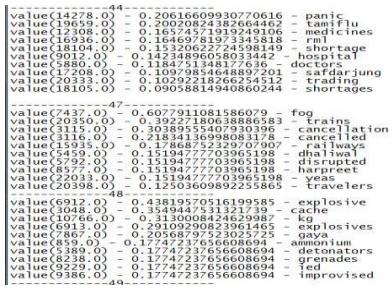
Once we are done separation of the comments as positive and negative, we analyzed them manually even programmatically and checked for the similarity so the scores we got with the code were competitive.

We wanted to know the comments about particular of Smartphone like battery life, screen resolution, etc. so we applied our clustering code to these comments and we were able to identify the clusters of the text comments which are in text file speaking something related features. All the people comments for particular features were put aside.

So in the next section, we have included the results and various snapshots to show that results achieved were correct and competitive.

RESULTS AND DISCUSSIONS

In this section we have shown the results of our experiments conducted. According to our



proposed methodology, we started with the concept extraction which is shown in the below.

Here with the above figure, we can analyze that, there are different concepts as a group of words is shown. For instance, the concept 47 in the figure shows that there are some discussions in the news which are related to the cancellation of Trains because of the Fog. So in this, we have come with many concepts which are meant important in the documents and which are hidden.

Our next goal was to apply an SVD based efficient Hierarchy approached algorithm to these News samples to get the related documents at one side. The below figure 6 shows that their

different clusters we obtained and next subsequent Figures 7 and 8 show that a proof of clustered documents was speaking about a particular topic.

Again, we separately applied the Document similarity measure between all the documents to check the correctness of the clustering algorithm; we were able to see that documents 47 and 7 are most identified with their similarity values which showed.

We also calculated the time efficiency of the algorithm by increasing the number of documents. we can conclude that the algorithm is efficient, i.e. there is a constant increase in the time as there is an increment in the number of the documents.

Next, we applied the opinion analysis technique on the Comments which were collected for a Samsung Smartphone.

We also wanted to know that, how accurately the comments are getting separated. For that we went through the comments manually and compared with programmatic results, so the results obtained were showing an average percentage of matching, and the value was around 84 percent.

The graph shows that, the line in violet color for manual analysis, which is supposed to be 100 percent accurate and on the other hand, the line in red color with Programmatic analysis which was around 84 percent accurate.

Finally, with these separated sentiments, we again moved back to clustering technique to know the opinions of the customers towards the specific features of a Smartphone. We got the clusters of the comment documents speaking towards a particular feature of a phone. There were clusters separated for battery life, screen resolution, etc. In this way, we achieved the expected results.

CONCLUSION AND FUTURE WORK

With the experiments conducted and results shown in the above sections, we can conclude that we found an SVD based technique for clustering the documents and extracting the concepts which are hidden in the documents. Also, we found a novel approach to finding the opinions of the customers towards specific features of Smartphone using Opinion analysis and clustering technique. We also showed that the algorithms and the experiments conducted were competitive and efficient. This work can be made language agnostic so that it can have wide uses in applications like Product marketing, Online shopping, etc.

REFERENCES

[1] Arora, S., Ge, R., and Moitra, A. learning topic models-going beyond SVD. In Foundation of computer science

[2] Arora, S, R. Halpen, D., Moitra, A., Sontag, D., Wu, Y., and Zhu M. A practical slgorithm for topic modelling with provable guarantees. In International conference on machine learning, 2013

[3] S. Deerwester, S. Dumais, G. Furans, and T. landaer. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391-407, 1990.

[4] J. M. Schultz and M. Liberman. Topic detection and tracking using idf- weighted cosine coefficient. In proc. DARPA Broadcast News Workshop, pages 189-192, Hemdon, Virginia, 1990.

[5] F.Walls, H. Jin, S. Siesta, and R. Schwartz. Topic detection in broadcast news. In Proc. DARPA Broadcast News Workshop, pages 193–198,Hemdon, Virginia, 1999

[6] E. Bingham. Topic identification in the dynamical text by extracting Minimum complexity time components. In Proc. 3rd Int. Conf. Independent Component Analysis and Blind Signal Separation, pages 546–551, San Diego, California, 2001.

[7] A. Kab´an and M. Girolami. Unsupervised topic separation and keyword identification in document collections: a projection approach. Tech. rep. 10, Dept. of Computing and Information Systems, Univ. of Paisley, 2000.

[8] T. Kolenda, L. Hansen, and J. Larsen. Signal detection using ica: Application to chat room topic spotting. In Proc. 3rd Int. Conf. Independent Component Analysis and Blind Signal Separation, pages 540–545, San Diego, California, 2001.

[9] P. Willett. Document clustering using an inverted file approach. Journal of Information Science, 2:223–231, 1990.

[10] L. Baker and A. McCallum. Distributional clustering of words for text classification. In Proceedings of ACM SIGIR, 1998.

[11] X. Liu and Y. Gong. Document clustering with cluster refinement and model selection capabilities. In Proceedings of ACM SIGIR 2002, Tampere, Finland, Aug. 2002.

[12] D. Cutting, D. Karger, J. Pederson, and J. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of ACM SIGIR, 1992.

[13] Manning, C. D.; Raghavan, P.; Schutze, H. (2008). "Scoring, term weighting, and the vector space model". Introduction to Information Retrieval (PDF). p. 100. doi:10.1017/CBO9780511809071.007.ISBN 9780511809071

[14] Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43

[15] Stanford core NLP Group: http://nlp.stanford.edu/software/corenlp.shtml

[16] The Apache Software Foundation, Licensed under the Apache License, Version 2.0 ; https://lucene.apache.org/core/