iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

Big Data Summarisation of Textual Data using NMF Methodology

Ravishankar V[1], Komal M[2], Harshitha S R[3], D Vishwanatha[4]
[1]*Additional Asst. Professor, R.V College of Engineering*
[2, 3, 4]*R.V. College of Engineering*

Abstract: ***Data summarisation allow organisations to make better decisions .The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researchers already knows. Text summarisation has emerged as important research area in recent past.***
This paper proposes a method to represent inherent structure of big data (textual data) set using the feature of semantic by using non-negative (NMF) technique. The existing methods such as text rank algorithm have been reviewed to discover the drawbacks.
The objective is to propose a new method using both NMF and K-means clustering to summarise with regard to a given data. Distributed NMF based summarization module and parallel processing systems are adapted to summarise the textual data. The NMF can find parts of representation of the data because non-negative constrains of NMF are compatible with the intuitive notations of combining parts to form a whole which is way NMF parts based representation is done.

1) ***It can extract sentence reflecting the inherent semantics of a document***
2) ***Improve quality of summarization***
3) ***Find important events in the data***

The textual big data is collected and the data is cleaned with the help of pandas (data creation, manipulation, wrangle) after which the data can be readily encoded. Later the encoded textual data is segregated for which the NMF methodology is implemented with the help of programming language. Python programming language is been used for the supervised learning and that allow easy utilization scripting and library called pandas is used, which aids the manipulation and analysis of data also helps performing many fact without writing the actual code. It results in the effective segregation of textual data based on the titles. The summarised textual data can be easily visualised using various visualization tools such as pycharm, ploty [12] etc to draw inference and a statistical study can be made. The inherent structure of textual big data could be represented using the proposed method with the help of semantic feature by the distributed NMF based parallel processing.

I. INTRODUCTION

In the era of rapid development area of internet data is freely available to be use for readers in the visible shape of e-newspaper, journal articles, Technical report and transcription dialogues etc. There are huge number of data available within the digital media and confiscate the acceptable information from all those media is difficult job for people in an exceedingly demand of time.

Big data deals with huge amount of data which flows rapidly from various data source and has different formats. Big data is large and complicated which makes it difficult to process using traditional approaches were summarization becomes a tedious process. One of the existing method of data summarization of natural language processing (NPL) are speech tagging, N-grams, text rank algorithm, key word extraction. These traditional document summarization methods put a limit to summarize accurate information exploring big data (i.e., message, email, blog, collection data of smart phone, SNS). The drawbacks of existing data summarization techniques are algorithm complexity, failing to detect outliers, improper handling of spatial data sets and time consuming.

Therefore an effective data summarization technique is needed to overcome these drawbacks. This paper proposes a new method for data summarization which utilizes NMF (non-negative matrix factorization) technique for effective data summarization which would represent the inherent structure of big documents set using the feature of semantic by using NMF technique. The method is also used for summarizing the big data size of document for Iot using the distributed parallel processing based on hadoop.

Data summarization is process of decreasing the dimensions of data while maintaining their basic summary. The three methods of data summarization are selection of important data, rejection of unnecessary data and substitution of large data with a single sentence which describes the large data. Summarization may be done on the query based or generic summary. It may split into single document summarization or multi document summarization, which supports the scope of summary target. Multi document is to supply single summary from the varied document, where as single document summarization performs only one document summarization at time.

Initially a textual big data document is considered for the implementation of NMF methodology for data summarization [4]. The first step is to clean the textual big data which includes unwanted data deletion, data addition, data manipulation and wrangling. In the second stage the encoded textual data is represented in the vector form, which gives rise to the metrics. In the third stage the zero values in the vectors are considered and they are replaced by the null value which belongs to a particular semantic.

After the completion of all the three stages the data is now ready for the implementation of NMF technique. At this point the procedure of the NMF technology has to be studied to arrive at the algorithm which can be implemented on textual big data. The procedure, algorithm and the clustering property of NMF technology are briefly explained.

A. Non-Negative-Matrix Factorization.

NMF is a combination of group of algorithm in multi variant analysis and linear algebra. When the matrix V is factorized into two respective matrixes W and H, with the property that all the three matrix are non negative element. These non-negative matrixes make the result easy to inspect. Since the problem is not solvable in general, it is commonly approximately numerically.

As we all know that NMF has a long history under the name of 'Self modelling curve resolution'. In this method the matrix is curved rather than discreet.

B. Procedure

Let V be the product of W and H.

V=WH

Matrix Multiplication is applied here.

Dimensions of a matrix may be lower than the product matrix and it's the basic property of NMF. NMF generates factors which significantly reduce the dimension compared to the original matrix.

Let V=m (rows) n (column)

W=m (rows) p (column)

H=p (rows) n (column)

Which is p can be significantly less than both m and n.

There are some examples based on this application.

1) Let us assume the input matrix be V with 100 rows and 50 columns. Where words are in row and letters are in column. That is we have 50 letters indexed by 100 words. And vector v in V represents a letter.
2) If the algorithm is made to find 10 featured in order to generate a feature matrix W .W consist of 100 rows and 10 column in addition to coefficients matrix K with 10 rows and 50 column.
3) The product of W and K is a matrix with 100 rows and 50 columns. By applying matrix multiplication we can say that this linear combination of W and K gives V.

C. Clustering Property of NMF

V=WH.

More specifically the approximation of v which is V~WH is only achieved by finding W and H which minimize the error function.

||V-WH||, which is W is greater than equal to 0, H is greater than equal to 0.

If we further proceed constraints on H that is $HH^T=I$, which is equivalent to minimization of K-mean clustering.

Further the H gives the cluster membership which is $H_{KJ} > H_{IJ}$ for all i is not equal to k, which suggests that the input data V_j belongs to K^{th} cluster. Whereas W gives the cluster centroids, which is K^{th} column gives the cluster centroid K^{th} cluster. These centroids give a very significant enhancement to NMF. Also $\mathbf{HH^T=I}$ is not explicitly imposed, it holds to a large extent, and the clustering property holds too.

D. Algorithm

There are so many ways to find W and H. But multiplicative rule has been popular method, because it is easy to implement .In this algorithm we implement W and H as non-negative [2].

$H^{n+1}_{[i,j]} = H^{n}_{[i,j]} \{((W^n)^T V))_{[I,J]}/((W^n)^T W^n H^n))_{[i,j]} \}$ and

$W^{n+1}_{[i,j]} = W^{n}_{[i,j]} \{((V(H^{n+1})^T)_{[I,J]}/(W^n H^{n+1} (H^{n+1})^T)_{[I,J]}\}$

Till W and H are stable.

Also W and H factor is identity matrix with V=WH.

Some more interesting algorithm is developed. Few approaches are based on altering non-negative least square .

Steps for this algorithm

1) H is fixed where as W is found by non-negative least square solver, and vice versa.

2) Some of the NMF variants regularise one of W and H. Therefore the method used to solve W and H may be the same or different. There is other method also such as Gradient descent method, Active set method, optimal gradient method etc. Present algorithm is sub-optimal in that they only find a local minimum, rather than a global minimum.

II. ISSUES AND CHALLENGES OCCURES IN TEXTUAL DATA SUMMARIZATION

Some of the following research issues and challenges may occurs during the implementation

A. Research Issues

1) While performing a multi document textual data summarization several problems may occurs in which evaluation of summary such as sentence ordering, redundancy, temporal dimension, etc. which makes the very rigid to accomplish the quality summary. Were other issues occurs such as cohesion grammatically, coherence which may harmful for summarization [6].

2) The standard of summary varied from person to person or system to system. Some of the people sense some set of sentence are effective for the summarization, at same time other people feels other set of words are crucial for the needed summarization.

B. Implementation Challenges

1) For achieve of quality summary, quality keywords are needed for the textual data summarization.

2) There is no accurate standard to recognize quality of key words in the document. The extracted keywords are varying while applying different approaches of keywords data extraction.

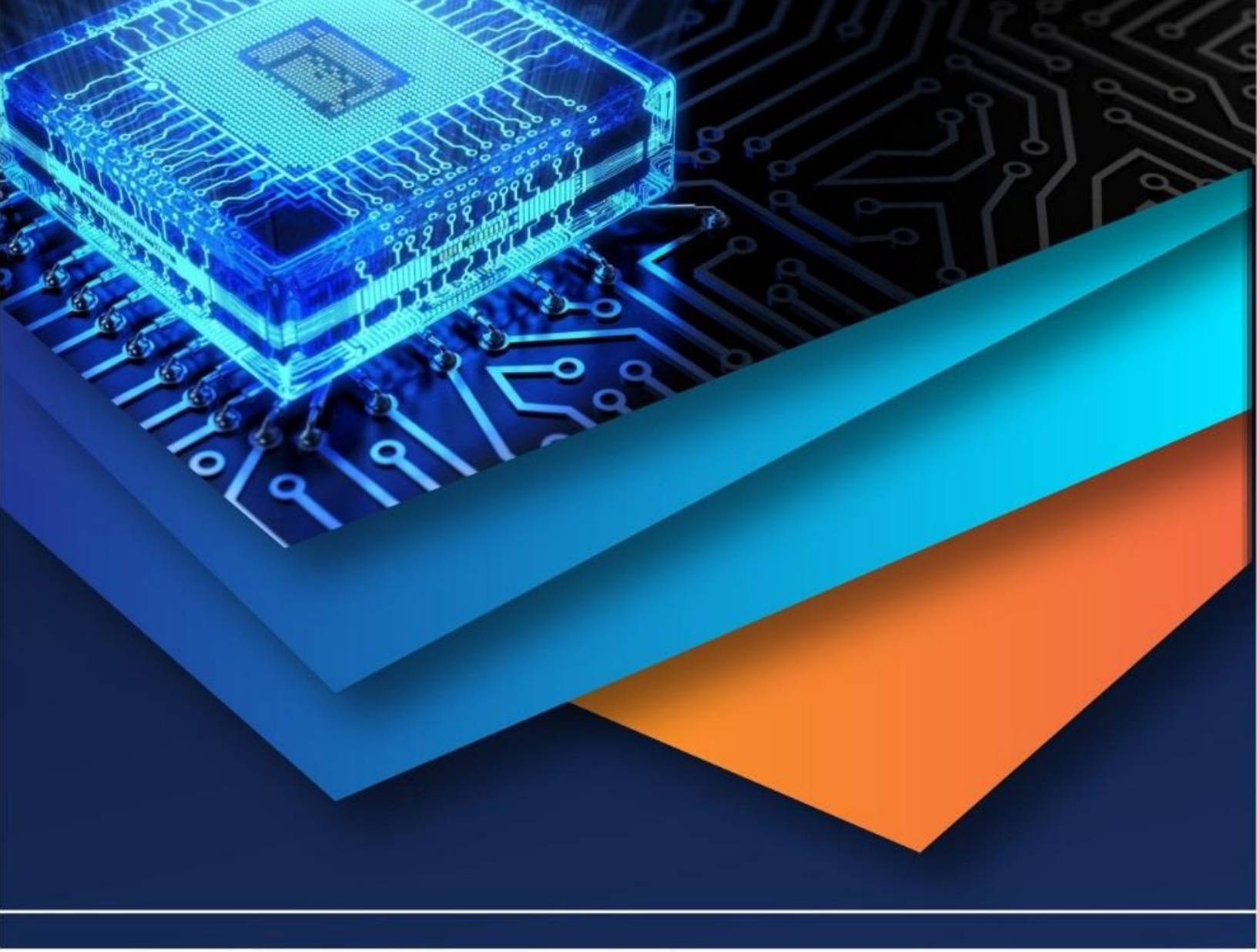
3) Multi-lingual textual data summarisation is challenging task.

III. CONCLUSION

Textual data summarization is very useful for the users to extract only needed information in demanded time. In this area significant amount of work has been done in the recent year. Due to shortage of information and standardization lot of research is still taking place. Traditional document summarization procedure are restricted for the summarizing the approximate information from big document data, for enhancing the summarization quality it may uses various natural language processing methods based on single node computer environment which is complicated and time consuming process. Therefore in order solve limitation of the summarization for big document data summarization, this paper is proposed in which information summarized from a big data is done in a effective way which overcomes most of the drawbacks of traditional summarization technique, the proposed methodology can represent inherent structure of big data document set using the semantic feature by NMF methodology in addition to that it can summarize big data using the distributed parallel processing. This paper provides a novel text summarization process which includes summarization methodologies and evaluation of matrices in addition to some of the important research issues in this region of text summarization are also discussed in this paper.

REFERENCES

[1] M.W.Berry, M.Browne, A.N.Langville, P.V.Pauce and R.J.Plemmons.Algorithms and application for approximate non negative matrix factorization. Computational Statistics & Data Analysis, 52(1):155-173, September 2007.
[2] E.Y.Chang, K.Zhu and H.Bai. Parallel algorithms for mining large scale datasets.In CIKM, 2009. S.Few, Multivariate Analysis using parallel co ordinates,https://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf.
[3] Mani,"Automatic Summarization",John Benjamins Publishing Company,2001.
[4] W.L.C. David A Kenny,Deborah A.Kashy. Dyadic Data Analysis. The Guilford Press,2006.
[5] P.O.Hoyer Non Negative matrix factorization with sparseness constraints. J. Mach. Learn. Res.,5,2004.
[6] K Kanjani. Parallel non negative matrix factorization for document clustering. Technical report, Texas A& M University, May 2007.
[7] D Kim,S.Sra, and I.S Dhillon. Fast Newton-type methods for the least square non negative matrix approximation problem.In SDM, 2007
[8] W.Xu,X.Liu, and Y. Gong. Document clustering based on non negative matrix factorization. In SIGIR.
[9] D Kim,S.Sra, and I.S Dhillon.Fast projection based methods for the least squares non negative matrix approximation problem. Statistical Analysis ana Data Mining,2008
[10] P. Zikopoulos, Understanding Big Data : Analytics for Enterprise Class Hadoop and streaming Data, McGraw-Hil, New Ypurk.USA, 2012
[11] J.Heer and B.Shneiderman."Ineractive dynamics for visual analysis",Communications of the ACM.
[12] Evgeniy Yur'evich Gorodov and vasiliy Vasil,evich Gubarev –Analytical revoew of data visualization methods in application to Big Data



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY