International Journal For Technological Research In Engineering
Volume 4, Issue 8, April-2017 *ISSN (Online): 2347 - 4718*

A CLUSTERING OF FACEBOOK POSTS BY USING K-MEANS CLUSTERING

MS.Badval Satvirkaur[1], Mr J Vinitkumar Gupta[2]

Abstract: It is very important to find out useful information from big amount of data. In this paper we clustering the data using the k-means algorithm. There has been a very important analysis area that how to cluster the facebook data and get the result; data mining is the a technique of extracting hidden predictive imformation from the large data .a social media platform such as facebook is main steam which give the information and many knowledge about data.there is so many textmining method technique. By using k-means approach we get result.
Index Terms: Data mining, Social Network, Facebook, k-means clustering

I. INTRODUCTION

Data mining helps to identify valuable information in such huge databases. Data Mining is an analytic process designed to explore data in search of consistent patterns and/or system-atic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Social network websites such as Facebook, Tweeter, etc. have become a useful marketing toolkit. Many companies find that it can provide new opportunities. Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning. Text clustering is in use in information search and providing information access, busi-ness analytics, corporate investigation, national security. Now there is a wide range of cluster algorithms and different mod-ifications. The problem of objects splitting into clusters as-sumes a set of solutions and the choice of a clustering method is connected with the evaluation of clustering results. Cluster algorithms applicability is carried out in relation to specific data sets. Choosing the suitable algorithm and setting its pa-rameters for text data analysis needs special consideration. In this work the features of k-means algorithm application in document clustering are investigated.

II. RELATED WORK

Clustering algorithms is subdivided into several types such as hierarchical, partitional, density-based, grid-based, fuzzy clustering. Classical hierarchical algorithms allow create full tree of the overlapping clusters. Nonhierarchical algorithms are based on optimization of some objective function which defines splitting objects set [1]. In this group there is special set of clustering algorithms k-means (k-means, fuzzy c-means, Gustafson-Kessel) which use the sum of squares of the weighed deviations of objects coor-dinates from the centers of required clusters as target func-tion. Clusters are looked for a spherical or ellipsoidal shape. The algorithm of k-means is considered as one of the most effective tools for carrying out clustering of text data, the efficiency of application for this method for sim-ilar data types are supported with experiments. There are a lot of software of Text Mining now (RapidMiner Studio, IBM Intelligent Miner for Text, SAS Text Miner, Semio Corporation SemioMap, Oracle Text, KnowledgeServer, and Megaputer Intellidgens Text Analyst). Such soft-ware represents scalable systems with different linguistic and mathematical methods of the text analysis. Similar systems have visualization and data manipulations tools, graphic in-terfaces, provide access to different data sources. RapidMiner Studio is the environment for carrying out experiments for data analysis and machine learning, including data loading and data translations (ETL), visualization, modeling. Processes of data analysis are represented randomly by the enclosed operators in created XML files due to the graphic user interface RapidMiner. GUI generates the XML file which contains analytical processes which the user applies to data. The graphic interface can be used for interactive management and check of the started processes. The platform is available both in a cloud, and in the client-server option. For commer-cial versions an opportunity to work with Big Data is given, connection to different data sources is provided. The platform easily extends by means of libraries, BI platforms and web applications.

III. PROPOSED WORK

We have to apply the method of Facebook clustering to cluster the data that we collected from Facebook . A Facebook has different types of data like comments ,posts ,like, share etc. we retrieve the facebook comments and likes. The goal of the clustering of facebook data is that to manage and make structured collection of the information. Data The results can be used in ad-vanced marketing for the company and to satisfy more users. Data mining is the process in which extract the knowledge from the data. In this research we are use k-means clustering method to build the cluster of facebook data. So the main objective is to develop the system is to cluster the facebook data and convert it into structure data.

Data extract

For data import we use Microsoft 2016. We import likes and comment of the chef vikas Khanna post. Document clustering is a process of detection of natural groups in a collection of documents. Let there Xn is data set {x1,..., xn} and the function defining degree of similarity of objects in most cases is function of distance between objects d(xi, xj). The problem of the analysis for text in a natural lan-guage is complexity of selection useful information except for its size and metadata. To make possible using of traditional cluster

algorithms is necessary to transform an array of text documents to numerical form. There are two common models for representation of text collections: treelike and vector mod-els. The treelike model is sets of the chains following one after another words. Such way allows create similar chains among different documents and reveal their similarity.

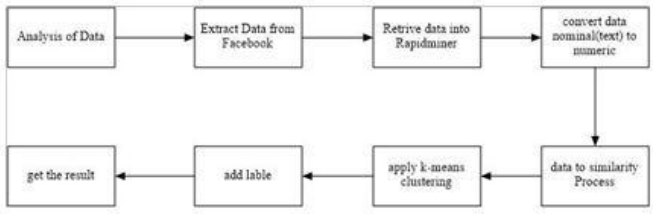


Fig 1 flow chart fot proposed work

The vector model is a matrix with frequencies of the words occurrence in a document. Let T is an array of text data, N – total quantity of terms in all arrays. Let T is a set of terms in an array of text documents. Then document presents as a vec-tor wiht length N in which each element corresponds term from a set T. The coefficients specifying the frequency of oc-currence of the term in the document can be values of ele-ments. Texts in vector model are considered as set of the terms constituting them. This approach is called a bag of words. Ap-plication of vector model assumes the choice of a method of terms weighing. There are several standard methods for nu-merical estimation of the document term. Term frequency (TF) defines term weight depending on the number of occur-rences in the document. Thus importance of the word in the document is estimated. Inverse frequency of the document (IDF) represents the return frequency of the document with which some word occurs in documents of a collection pro-motes reduction of weight of the most common words.TF-IDF (term frequency - inverse document frequency) is the statistical measure used for an assessment of importance of a term in the context of all set of documents. TF-IDF is calculated as product of the number of word entries into the document and functions from value reciprocal of the documents number TF-IDF value increases for terms which often occur in the specific document and seldom used in other documents of a collection. The vector space model allows define quickly with a small error a key word in the text and document sub-ject. The vector model despite of the shortcomings remains the most often used in text analysis.

IV. EXPERIMENT AND RESULT

As the tool for document clustering we selected Rapid Miner Studio 6.002, at the last version there is a set of modern algorithms, tools and approaches for text analysis [7]. For document rubrication we constructed cluster models using different types of metrics and compared results by means of criteria of accuracy. In such a way it is possible to define the most suitable way of calculation of distance for text data type. Initial data are presented by a set of text documents from a news line goarticles.com. Text documents contain from 420 to 650 words. For an assessment of cluster model documents were grouped by expert way in four thematic categories (education, web design, real estate, cars). The special operator of loading Loop Files is applied to import a collection of text documents. During the analysis of text data it is necessary to transform contents of all documents for separate words. The operator Process Documents carries out preprocessing of the text, cre-ating a bag of words, and also calculates the frequency of each word presenting the models of a vector space. In this process the operator Process Documents consists of 6 subprocesses which are consistently connected (Tokenize Non-letters; To-kenize Linguistic; Filter Stopwords; Filter Tokens (by Length); Word stemming (Stem); Transform Cases). The op-erators Tokenize Non-letters and Tokenize Linguistic are cre-ated by adding in subprocess of the operator Tokenize with the choice of different parameters. The first operator breaks into the lexemes based not on letters whereas the second breaks into lexemes based on linguistic sentences within this or that language. The operator Filter Stopwords deletes all words which have length less than 3 signs or more than 25. Stem carries out process of finding of a word stem. Transform Cases will transform all characters in selection to the lower register. Often there is an issue of attributes setting before application of some operators especially for big and difficult data sets. The operator Select Attributes allows select the nec-essary attributes thanks to different types of filters. The se-lected attributes will be on the operator's output. This conver-sion is necessary for the following operator k-means who car-ries out a clustering only for numerical values. For a cluster-ing of text collections the algorithm of k-means is used. In RapidMiner there are different operators capable to give help in selection of best value of parameter k and an assessment of clustering quality. Application of different distances was fol-lowed by the analysis of model on a test set. Results of this process given below. The correct number of clusters is three. This is not the case with Eucledean measure which identifies three clusters. main process of implementation

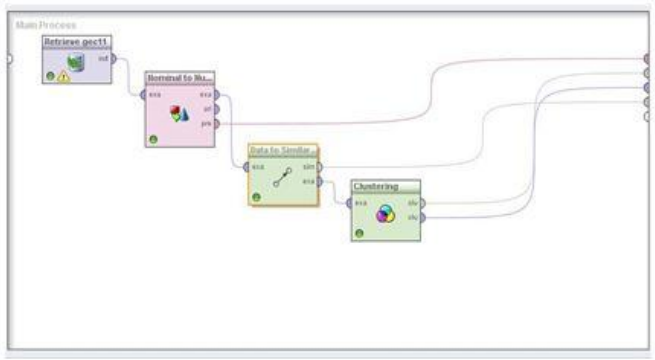
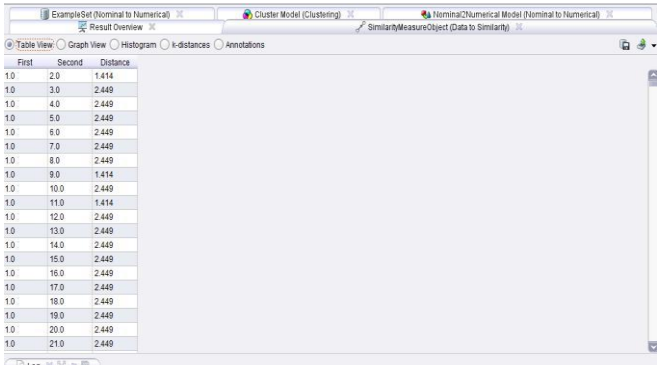


Fig 2 main process

implementation results

in implementation results we get the data to similarity, nominal to numeric result and k-means clustering algorithm output. In final k-means clustering algorithm, there is three different cluster as a output.

All output given below unsupervised learning method to solve known clustering is-sues. It works really well with large datasets.



First	Second	Distance
1.0	2.0	1.414
1.0	3.0	2.449
1.0	4.0	2.449
1.0	5.0	2.449
1.0	6.0	2.449
1.0	7.0	2.449
1.0	8.0	2.449
1.0	9.0	1.414
1.0	10.0	2.449
1.0	11.0	1.414
1.0	12.0	2.449
1.0	13.0	2.449
1.0	14.0	2.449
1.0	15.0	2.449
1.0	16.0	2.449
1.0	17.0	2.449
1.0	18.0	2.449
1.0	19.0	2.449
1.0	20.0	2.449
1.0	21.0	2.449

In future work we will cluster the other data of the face-book. We will also apply the different clustering method to improve the performance of the result.

Acknowledgment
I would like to thank my guide VinitKumar Gupta and Prof .Indr Jeet Rajput head of the computer department. The door to Prof. Vinitkumar Gupta office was always open when-ever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my Fig 3 Data to similarity own work, but steered me in the right the direction whenever he thought I needed it. Without their passionate participation and input, the validation survey could not have been success-fully conducted. I am gratefully indebted to his for his very valuable comments on this dissertation.

Fig 4 similarity measure's histogram view

Fig 5 clustering output

V. CONCLUSION AND FUTURE WORK

We have to apply the method of Facebook clustering to cluster the data that we collected from Facebook. First we import the facebook posts in the excel sheet. This data is retrieve in the Rapidminer tool. K-means cluster is not applicable on the text nominal data so we converted this data in nominal to numeric by using nominal to numeric operator. In k-means algorithm cluster is made based on the shortest distance. We measure all data similarity using data to similarity operator. After this we apply the k-means clustering algorithm and we get three different cluster. Link, status and photo this is the different type of the clusters. We researched about the k- means and these are what we got: k-means is one of the simplest algorithm which uses.

REFERENCES

[1] Vassilios S. Verykios1, Elisa Bertino2, Igor Nai Fovino2," State-of-the-art in Privacy Preserving Data Mining", CODMINE IST FET Project IST-2001-39151

[2] Keke Chen Ling Liu," Privacy Preserving Data Classification with Rotation Perturbation", Proceedings of the Fifth IEEE Inter-national Conference on Data Mining (ICDM'05)

[3] Kun Liu, Hillol Kargupta, Senior Member, IEEE, and Jessica Ryan," Random Projection-Based Multiplicative Data Perturba-tion for Privacy Preserving Distributed Data Mining", Published by the IEEE Computer Society 2006.

[4] S. Saitta, B. Raphael, I. F. C. Smith, "A bounded index for Cluster validity", in Proc. of Int. Conf. on Machine Learning and Data Mining in Pattern Recognition, pp. 174–187, Springer, 2007.

[5] U. Maulik, S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices", IEEE Transactions Pattern Analysis Machine Intelligence, vol. 24(12), pp 1650–1654, 2002.

[6] D. Davies, D. Bouldin, "A Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, No. 2, 1979, pp. 224–227.

[7] M. Hofmann, R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2013.

[8] G. Chernyshova, V. Gusyatnikov, "Application of Forecasting Technique for Economic Indicators", in Proc. Int. Conf. on Cloud, Big Data and Trust, pp. 23-24, Bhopal, India, 2013.

