A CLUSTERING OF FACEBOOK NOTIFICATION BY USING K-MEANS CLUSTERING

Badval Satvirkaur J¹, MS.Bhagyashree S.Patil², Mr Vinitkumar Gupta³, Mr. Indra Jeet Rajput⁴ Department of Computer Engineering, Hashmukh Goswami College of Engineering, Vahelal

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, Text minig, 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 systematic 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 business analytics, information access, corporate investigation, national security. Now there is a wide range of cluster algorithms and different modifications. The problem of objects splitting into clusters assumes 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 parameters 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 cmeans, Gustafson-Kessel) which use the sum of squares of the weighed deviations of objects coordinates from the centers of required clusters as target function. 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 similar 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, Knowledge Server, Megaputer Intellidgens TextAnalyst). Such software represents scalable systems with different linguistic and mathematical methods of the text analysis. Similar systems have visualization and data manipulations tools, graphic interfaces, 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 commercial 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. We first cluster the facebook likes and in other section we compare the likes of other post with own facebook post like to decided how many person from my friend list follow the same stuff I follow.

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 $d_i(xi, xj)$. The problem of the analysis for text in a natural language 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 models. 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.



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 vector wiht length N in which each element corresponds term from a set T. The coefficients specifying the frequency of occurrence of the term in the document can be values of elements. Texts in vector model are considered as set of the terms constituting them. This approach is called a bag of words. Application of vector model assumes the choice of a method of terms weighing. There are several standard methods for numerical estimation of the document term. Term frequency (TF) defines term weight depending on the number of occurrences 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 promotes 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 reciprocalof 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 subject. 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 RapidMiner 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 meansof 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, creating 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; Tokenize Linguistic; Filter Stopwords; Filter Tokens (by Length); Word stemming (Stem); Transform Cases). The operators Tokenize Nonletters and Tokenize Linguistic are created 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 necessary attributes thanks to different types of filters. The selected attributes will be on the operator's output. This conversion is necessary for the following operator k-means who carries out a clustering only for numerical values. For a clustering 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 followed by the analysis of model on a test set. Results of this process given below.

Edit Parameter Text text The text		EdifParameter Text text The text		Edi Parameter Text text The text			
id name		1 id name		1	id name		
1872358486283123 Nic	kson Joyson	2 2244035475768612	Kangu Singn Koun Kanghan Chourasia	3	402/02000/02940	Kaveri Guota	
30030441/032807 Pete	Clask	4 1348884088536809	Privatka Das	4	10200385007963818	Kavita Kumar	
AMENEDEDEDEDEDED Pup	s sangn sthi Baddy	5 939812636147236	Shathank Sharea	5	361975387487813	Harry Sandhu	
1007733014401000	Gerni	6 238658346552281	lous Masrin	6	589288122792286	Heenal Rathod	
738183998476749 Tev	ti Sharma	392223191146877	Alina Jasmeen	1	188482531668387	Chandramouli Sriram	
616711811853571 Sco	nia Sinch	182995235538348	Dita Des		10154378724381806	Hushtag Ahmed	
972817432826442 Lub	u Setar	9 186893786513618	Garv Resson	9	283371375355468	Pragati Varshney	
1132987858187823 Shr	inath Goodalia	10 912345842166472	Abdul Mannan Saeed	10	1434188329995375	Shaheen Arefa	
1748226848841327 Aka	rksha Chavan	11 1686218668283123	Syed Thahir	11	238658346552281	Joya Nasrin	
773347195166584 Kar	Lshma Thakre	12 168129897839859	Dyoti Dhivar	12	392223191146827	Alina Jasmeen	
787683412733889 Gol	di Sinha	13 279618769119488	Satabdi Dhara	13	912345842166472	Abdul Mannan Saeed	
738828856274398 Rum	a Pratihar	14 1119491838161189	Rahul Agarwal	14	1686218668283123	Syed Thahir	
1238649559564824 Jyo	ti Mishra	15 170120546837079	Neetu Norya	15	1218724644862354	Bharu Bahot	
198237370661783 Nit	in Sharma	16 237287976677899	Svara Saroj	16	279618769119488	Satabdi Ohara	
1686647714968922 Sus	wa Kumari Nishad	17 1639851949376982	Rama Singh Barhat	17	1621571981488577	Pankaj Rehar	
1723865217987419 Kan	chan Negi	18 687577614677993	Gaurav Iglesias	18	1772743969728669	Darshita Wehta	
846683075490812 Shil	char Singh	19 1379426788767512	Hitarthi Bakshi	19	1639851949376902	Rana Singh Barhat	
1376374075779803 Div	ya Mudaliar	20 1796732673913293	Nikitha Reddy	20	1053498343778288	Seena Raykar	
198234871468412 //cm	u Naggo	21 118541458678474	Khushi Arora	21	986235318177729	Rakesh Dubey	
398364871346225 Nam	rata Jamdar	22 251454151874586	Krati Sharma	22	1355/09441145093	vaisnaii Pippai	
581892951982910 Rx 1	/ibha Gajera	23 286703821762473	Raman Sinha	23	18285694223854989	Sniveta Uniabra	
1110017992460206 Pru	gakshi Parikh	24 778888536581469	Vibha Sahu	24	4010/33/83628/4	Suntuality Suntra	
1758470781874530 Raji	ni Sharma	25 1863498343778288	Seena Raykar	25	52365311/811100	Recikant danjare	
829645463791329 Lak	shmi Ganesh	26 18/8448889867837	Sanskriti Basnet	26	120113040/22/555	Reta Object	
1876748472618978 Par	ta Nahadik	27 284/11591959792	Pooga Dalal	21	10010002142/10//	Shakers Bernik	
1758214724439501 Hem	ant Giri	22 1562968488696892	Vinuta Koraddi Naw filesh	20	100300001/50005	Giran Ginta	
18/88/91885254533 Serv	uti singh	27 995821667384595	njey slûgt Charte Orbeker	42	580728521223560	Middl I genius	
2588858864598799 Ang	21 18118	0 20203094223034909	Sivela Unter 3	45	1220002268895122	tiliki Starea	
11:95/882/857886 Jud	istir Honanty	10 172210470647170	Vanue Justa Caselita Gueta	12	209666376178775	Thuisha Bauak	
138/8690/66169/1 Raj	para kusawata	10 \$125224/304/1/6	Manifest Engines	12	834565963347436	lierra Ita Richra	
10213101920091420 Bft	Long denge	1/ 1774541341061110	Recondition using all the	34	928877358835751	Deenika Sharma	
/30120/0/0/0/09/8 N1K	ite retei Hillio	35 18135753578278253	Swith Singh	35	223812771581228	Tithi Sarkar	
03003003/40034/ Jy0	11 208	% 684703390735741	Suffa Varea	36	323977297975757	Kawitha M Kawi	
110011+002+20518 5MV	d Chahai	17 18212545553681478	Varonica Tatka	37	1935827989974288	Saniksha Saxena	
9200040040//451 Fan	ij uletri	20 102123403030014/0	veronace verold		1000002080000000	Taught Solly	

Word	Attribute Name	Total Occurences V	Document Occurences	
SHARMA	SHARMA	2358	3	
PRIYANKA	PRIYANKA	739	3	
MISHRA	MISHRA.	584	3	
AGARWAL	AGARWAL	475	3	
KUMARI	KUMARI	448	3	
THAKUR	THAKUR	399	2	
CHAUHAN	CHAUHAN	370	2	
ANKITA	ANKITA	357	2	
ANJALI	ANJALI	340	2	
PANDEY	PANDEY	332	2	
SHWETA	SHWETA	328	3	
AGRAWAL	AGRAWAL	317	2	
SHIVANI	SHIVANI	299	3	
TIWARI	TIWARI	291	2	
MANIBHA	MANISHA	290	3	
SARKAR	SARKAR	283	2	
MONIKA	MONIKA	278	2	
POONAM	POONAM	275	2	
JAIBWAL	JAJSWAL	260	3	
RAJPUT	RAJPUT	245	3	
PREETI	PREETI	243	2	

Cluster Model

Cluster 0: 1 items Cluster 1: 1 items Cluster 2: 1 items Total number of items: 3

The correct number of clusters is three. This is not the case with Eucledean measure which identifies three clusters.

V. CONCLUSION AND FUTURE WORK

We have to apply the method of Facebook clustering to cluster the data that we collected from Facebook. A Facebook notification has different types as social page notification, group notification.Facebook are open to Internet. user can read and receive updates from a Fan Page by becoming a fan of the page. I researched about the k- means and these are what I got: k-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues. It works really well with large datasets. Using RapidMiner Studio on the basis of this technique allow the analyst quickly without labor-consuming development to carry out document clustering and selection of number of clusters. In Dissertation phase-I and msr survey we work on vector creation from the text data. Kmeans. In dissertation phase-2 we will going to get desire output of the k-means clustering.

5.1 Future Work

In Dissertation phase-I and msr survey we work on vector creation from the text data. K-means. In dissertation phase-2 we will going to get desire output of the k-means clustering.

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 whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my 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 successfully conducted. I am gratefully indebted to his for his very valuable comments on this dissertation.

REFERENCES

- 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 International Conference on Data Mining (ICDM'05)
- [3] Kun Liu, Hillol Kargupta, Senior Member, IEEE, and Jessica Ryan," Random Projection-Based Multiplicative Data Perturbation 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.