ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com
Journal DOI: 10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

A Framework for Ranking of Colleges Based on Unstructured Data using k-Anonymity Algorithm in Hadoop.

Ravuri Daniel, G.Mani, P. Prudhvi Kiran, and S. Praveen Kumar.
Department of Information Technology, Vignan's Institute of Information Technology, Duvvada, Visakhapatnam-530049, Andhrapradesh, India.

Manuscript Info

Manuscript History:

Received: 11 February 2016
Final Accepted: 19 March 2016
Published Online: April 2016

Key words:
Big data, Data mining, Hadoop, Hive, JOUM, K-Anonymity, MapReduce.

Corresponding Author
Ravuri Daniel.

Abstract

Choosing a right measure in assessing the education system might be a great challenge in present scenario. As the numbers of technical institutions are increasing enormously, students and parents are uncertain to take up higher education in a reputed institution. The information regarding colleges or institutions is graded by many institutional stake holders and others through social networking sites like Twitter. The large volume of data generated through social networking sites is unstructured which is posted by different kinds of people. Processing the unstructured data is a tedious process. To rank the institutions based on the unstructured data would be a difficult process using traditional or conventional data mining techniques and tools. The proposed framework is for ranking the institutions based on K-anonymity algorithm which is implemented in HADOOP and HIVE. This method improves the efficiency and accuracy of the data processing compare to the traditional methods.

Copy Right, IJAR, 2016,. All rights reserved.

Introduction:-

In recent years it has been observed that the growth of technical institutes is increasing tremendously. The career options for the students are also great in number. Consequently the system is creating uncertainty for the students as well as parents in choosing the right institution to continue higher education.

Many of us depend on reviews, comments, feedback, opinions and outlook given by previous preceding bodies or social networking sites. The institutional statistical data can be manipulated and there is a chance of false data generation by outward bodies when compared to the institutional stake holders. Thus, Generic and factual data which is a backbone of any institution has to be captured and projected to the students for making right decision.

Our paper Extract, Transform and utilize structured and unstructured data generated by stakeholders and others through social networking sites respectively. The generated data is then analyzed using Hadoop and Hive to produce classification of the institutions.

The remaining part of the paper is organized as follows: Section 2 gives a brief description of the important papers that are reported. Section 3 introduces proposed system model for ranking the colleges based on unstructured data in Hadoop. Section 4 discusses the implementation of the system. Section 5 shows the experimental results and discussions. Section 6 presents conclusion and future work.

Literature review:-

We live in on-demand, on-command Digital universe with data prolifering by Institutions, Individuals and Machines at a very high rate. This data is categories as "Big Data" due to its sheer Volume, Variety, Velocity and Veracity. Most of this data is unstructured, quasi structured or semi structured and it is heterogeneous in nature.

The volume and the heterogeneity of data with the speed it is generated, makes it difficult for the present computing infrastructure to manage Big Data. Traditional data management, warehousing and analysis systems fall short of tools to analyze this data. Due to its specific nature of Big Data, it is stored in distributed file system architectures. Hadoop and HDFS by Apache is widely used for storing and managing Big Data.

Harshawardhan et al,[5] said that, the term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte or larger-sized datasets with high velocity and different structures. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers.

It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.
Efthymios Kouloumpiset al, [3] described on "Twitter Sentiment Analysis: The Good the Bad" is investigate the utility of linguistic features for detecting the sentiment of Twitter messages. They use three different corpora of Twitter messages in their experiments.

For development and training, they use the hash tagged data set (HASH), and the emotion data set (EMOT). For evaluation they use a manually annotated data set produced by the iSieve corporation (ISIEVE).

T.K.Das et al, [10] wrote on "BIG Data Analytics: A Framework for Unstructured Data Analysis" is nowadays, most of information saved in companies are unstructured models. Retrieval and extraction of the information is essential works and importance in semantic web areas. Unstructured data targeted in this work to organize, is the public tweets of Twitter.

Building a Big Data application that gets stream of public tweets from Twitter which is latter stored in the HBase using Hadoop cluster and followed by data analysis for data retrieved from HBase by REST calls is the pragmatic approach.

By follow all the above papers we proposed an approach for process unstructured data based on K-Anonymity algorithm which is implemented in HADOOP and HIVE. This method improves the efficiency and accuracy of the data processing.

Proposed system architecture and methodology:-

Architecture:-

In the Figure 1 proposed system is providing rating and feedback of the colleges by processing unstructured data taken from social networking sites like Twitter. The application inputs dataset that contains the information taken from Twitter. The application also inputs dataset of the individual college data of the stakeholders like students, faculty of the colleges which is used to identify the internal students and faculty.

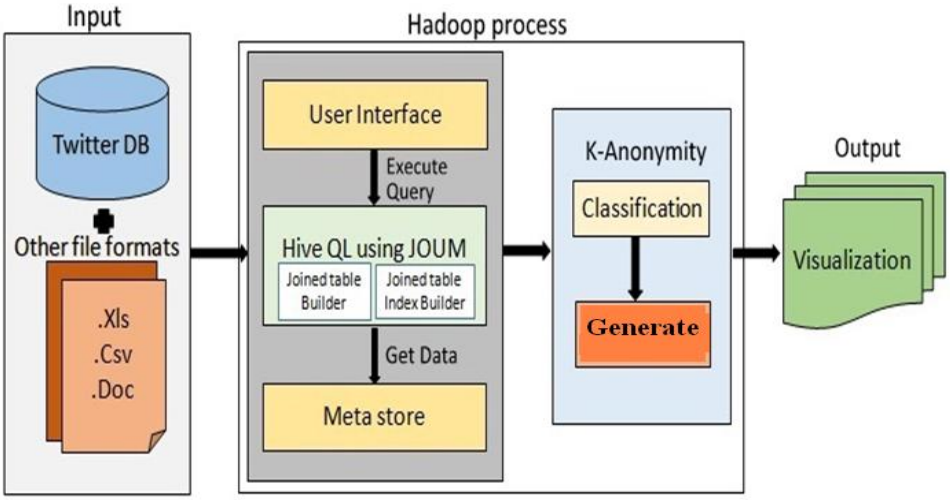

Figure 1: Architecture Diagram of Framework for Ranking of colleges

Then the input data is processed by using Hadoop. In Hadoop we write the Hive Queries by using user interface. By these queries we performed JOUM [7] (Join Once Use Many) operations. After that the processed data stored in Meta store. Then apply K-Anonymity algorithm on the processed data. Further the data is classified and clustered which results a data set. Finally the resulted data is visualized.

Methodology:-

For accessing the position of the institution, we need to classify the institutions based on important attributes such as feedback from all aspects and this has to be analyzed properly to fit into our framework. Hence we use K-Anonymity methodology that can attain efficient ranking among the colleges from the data set.

K-Anonymity Algorithm:-

K-Anonymity is one of the algorithm, tailored to solve the problem of identity disclosure. An individual is indistinguishable from at least (k-1) individuals in a k-anonymous dataset correspondingly, A dataset satisfies K-Anonymity, if every record in the data set is identical to at least (k-1) other tuples with respect to the set of quasi identifier attributes, and such a dataset is so called k-anonymous. The K-Anonymity algorithm limits the ability to link or match the published data with existing external information. These attributes in the private information can be used for linking with external data used for other specific purpose.

Algorithm Inducing k-Anonymity in Hadoop

Input:	User dataset
Output:	List of nodes according to K
	Where K is anonymity parameter chosen by user
Step1:	Procedure (T,A,k)
	T is a dataset,
	A is a list of attributes,
	k is a anonymity parameter
Step2:	Obtain d-> data_node from T.
Step3:	Generate college_List {(a, d) : a 2 A}
	while college_List contains college with
	positive_gain using equation (1)do
Step4:	split best_college from college_List with highest_gain.
Step5:	if best_college maintains k-anonymity then go to step6
Step6:	Apply the split and generate new data_nodes N.
Step7:	else
Step8:	remove best_college from college_List.
Step9:	remove colleges from data_node with Negative_gain
Step8:	end if
Step10:	end while
Step11:	return best_college
Step12:	end procedure

The detailed implementation steps of the proposed algorithm as follows:

Step I:

An unstructured dataset about colleges (Twitter data) is taken as input for the system.

Table 1: Twitter Data about colleges

87	95476210	ravi	ravi@gmail.com	2014-12-0 <a href="h college architecture is very good
88	23587405	hemu	hemu@gmail.com	2014-12-0 <a href="h we never seen such a college architecture
89	94852460	spiky	spiky143@gmail.com	2014-12-0 <a href="h lab programmers are doing some extras in vignan
90	34587624	raju ravi	rajuravi@gmail.com	2014-12-0 <a href="h holidays are very less in vignan waste
91	94852476	raju krish	rajukrish@gmail.com	2014-12-0 <a href="h some times students are need to come to college on sundays also
92	96589658	raju	raju@gmail.com	2014-12-0 <a href="h students are always occurred on ground waste
93	65485268	parvathi	paru@gmail.com	2014-12-0 <a href="h some students are escaping classes by siiting outside
94	45612308	prakash	pakish@gmail.com	2014-12-0 <a href="h if anystudent is lete comes to college they are not allowed
95	98457218	dinu	din@gmail.com	2014-12-0 <a href="h we cant go out side with out permission
96	64582452	shanti	shanti@gmail.com	2014-12-0 <a href="h vignan is always good in encouraging students
97	78452136	manu	manu@gmail.com	2014-12-0 <a href="h college provides some money for doing some projects to college
98	98742575	kisha	kish@gmail.com	2014-12-0 <a href="h seminar hall is too small
99	47586912	krishna	kai@gmail.com	2014-12-0 <a href="h NSS is doing some good things in college
100	34751280	sakhi	sakhi@hotmail.com	2014-12-0 <a href="h all feculty is very helpful to students
101	96325874	naveen	nav@gmail.com	2014-12-0 <a href="h very less time for lunch break
102	36541287	pranav	prachi@gmail.com	2014-12-0 <a href="h vignan is best college for engineering
103	85412035	kartik	kartikmutyam@gmail.com	2014-12-0 <a href="h thank you for vignan to be a part in my life
104	45879546	ram	ramravi@gmail.com	2014-12-0 <a href="h vignan is waste college
105	85476247	rajtarun	tarunraj@hotmail.com	2014-12-0 <a href="h faculty is bad in viit
106	87542394	premkumar	premkumar46@rediff.com	2014-12-0 <a href="h studies are worst
107	47985467	swamy	swamy@gmial.com	2014-12-0 <a href="h girls are too bad
108	68542137	rithu	rithu12@gmail.com	2014-12-0 <a href="h dirty college I have ever seen

tweets praveen2

Step II:
The unstructured data is interleaved in Hadoop Process to obtain a structured format using JOUM.

Step III:
The dataset is split into no. of nodes according k parameter. We have considered good and bad as k anonymity parameters for the data set considered.

Step IV:
The data set we have taken can be divided according to the colleges by running a query in HIVE using K-Anonymity algorithm. College_list is generated using following formulae

$$G_i = E_p - \text{avg}\,(E_c) \text{ ---------- } (1)$$

Where, G_i is Information gain,
E_p is Parent entropy,
E_c is child entropy.

$$E = -\,(\sum P_i \log (P_i)) \text{ ---------- } (2)$$

Where E is Entropy,
P_i is probability distribution of k-parameter

Table 2: After applying K-Anonymity Algorithm in Hive

	t.tweet_ic	t.name	t.text	s.id	s.branch	s.year	s.college
2	64873893	praveen k	vignan is good college for studies	1.33E+08	it	2	Vignan
3	54873893	prakash	Faculty is good in college	3.56E+08	ece	4	Vignan
4	84257698	parvathy	college looks good at outside	5.98E+08	ece	4	Vignan
5	45872694	sabeena g	very good in conducting cultural programs	9.96E+08	ecm	2	Vignan Lara
6	58245869	hemanth	campus is good	4.36E+08	mech	4	Vignan Nirula
7	87548962	raghunadl	principal is too good	3.92E+08	cse	1	Vignan University
8	42187945	renuka	good in campus placements	3.26E+08	ecm	4	Vignan Womens
9	47568475	kalyani.m	college is good for seminars	2.07E+08	mech	1	Vignan University
10	98758546	santhoshk	we can improve good communucation skills in college	1.8E+08	cse	2	Vignan University
11	50640564	prathyush	PD is good in college	3.88E+08	mech	4	Vignan University
12	40657854	jyothi	cse Hod is very good	7.9E+08	cse	2	Vignan
13	44402186	geethanja	college buses are very good	2.28E+08	ecm	1	Vignan University
14	21054863	geetha	T&P is not good in VIIT	1.53E+08	cse	4	Vignan
15	48521703	rakesh	I.T brach is very good in vignan	3.42E+08	mech	2	Vignan
16	98421057	rajesh	vignan womens college very good	58379816	civil	1	Vignan University
17	35124860	fatima	pharmacy also having in vignan is good	9.8E+08	civil	4	Vignan
18	42571350	PANDU	every faculty is good with students	5.03E+08	ece	1	Vignan
19	65820150	naveen	we can't get good internal marks	9.82E+08	cse	2	Vignan
20	60142204	prashanth	some faculty only doing good for their job some are escaped	4.33E+08	it	1	Vignan
21	84502105	gopika	college atmosphere is too good	5.57E+08	ecm	2	Vignan Womens
22	95476210	ravi	college architecture is very good	6.01E+08	civil	2	Vignan University
23	64582452	shanti	vignan is always good in encouraging students	7.68E+08	eee	4	Vignan

System implementation:-

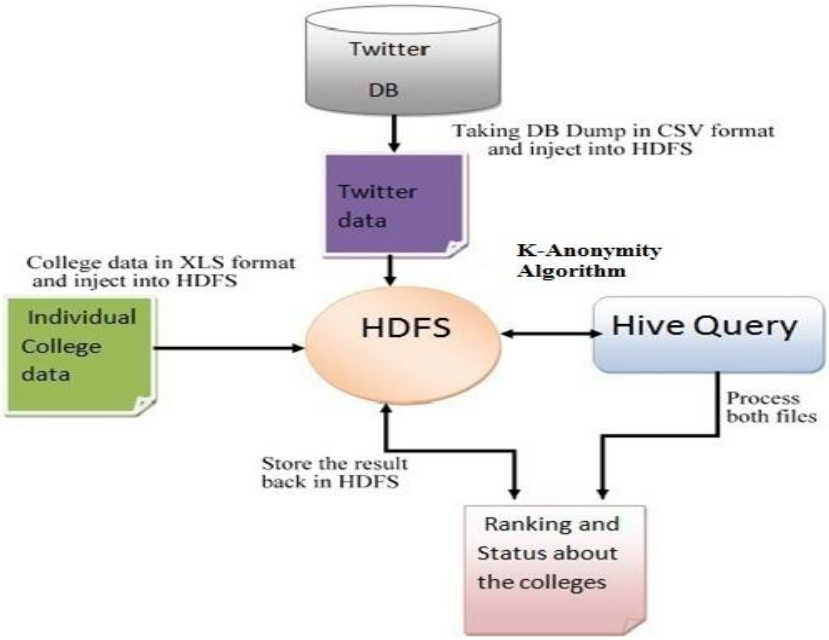

Figure2: Process flow Diagram for Ranking of colleges

- Taking database dump in .CSV format and inject into HDFS.
- Taking the colleges data in .XLS format and inject into HDFS.
- Then, read .CSV and .XLS files from HDFS through HIVE script.
- Apply k-anonymity Algorithm, process both files and get the results.
- Store the result data in a CSV file into HDFS.
- Then, download resulted file from HDFS.

Databases are abundant for small sets of data and low latency queries. However, when it comes to Big Data and large data sets in terabytes, traditional SQL database is not the perfect solution. Hive queries data in parallel across multiple nodes using MapReduce, distributing the database across multiple hosts as load increases. Hive can also be used as an alternative to writing java MapReduce jobs, because it provides an SQL-like interface to run complex queries against Big Data. By providing a simple, SQL like wrapper, complex MapReduce code can be avoided with a few lines of SQL-like entries.

The traditional RDBM system cannot scan a vast amount of data (tera, peta, zeta bytes) and assist for plan/ parse/ execute query using normal joins. On the other hand, HIVE is one of the important and efficient components in HADOOP for processing SQL queries. We introduce bucketmap join instead of normal joins in hive for query optimization. The bucketmap join as shown below.

set hive.optimize.bucketmapjoin = true;
This transition allows translating our SQL queries into Map/Reduce tasks and could speeding up HIVE and giving us efficient results. In this project we take both structured and unstructured datasets as input and transform (join) into one table according to one attribute.

According to this project we use JOUM methodology that join the tables in star schema data and build an index f or Joined data. Based on JOUM, SQL queries execution time in HIVE has been improved without changing HIVE framework. JOUM performance is improved even by increasing data size. The bucket map Join Operation is used to match the rows of two or more tables and produces all rows from all tables related to some specific fields or properties.

The joined data table contains schema Fact/Dimension tables which will be uploaded into HDFS.

```
SELECT t.tweet_id, t.name, t.text,s.id, s.branch, s.year, s.college
FROM tweets_praveen1 t
BUCKET MAP JOIN students_data s
ON (t.email_id = s.email_id)
```

Table 3: The joined data table

	t.tweet_id	t.name	t.text	s.id	s.branch	s.year	s.college
2	64873893	praveen kumar	vignan is good college for studies	1.33E+08	it	2	Vignan
3	54873893	prakash	Faculty is good in college	3.56E+08	ece	4	Vignan
4	55485493	srinivas	there is big ground in college	9.96E+08	ecm	2	Vignan
5	35849752	daniel	labs are not sufficient and bad	8.35E+08	it	1	Vignan University
6	84257698	parvathy	college looks good at outside	5.98E+08	ece	4	Vignan
7	45872694	sabeena grace	very good in conducting cultural programs	9.96E+08	ecm	2	Vignan Lara
8	58245869	hemanth	campus is good	4.36E+08	mech	4	Vignan Nirula
9	65848267	sriya	in vignan campus places are not provided and bad	7.06E+08	ecm	3	Vignan University
10	44478548	kalavani	no extra fees is taken by the college its awesome	8.03E+08	eee	1	Vignan
11	96584752	sindhu	exams are going in very strict manner	1.95E+08	civil	3	Vignan
12	87548962	raghunadh	principal is too good	3.92E+08	cse	1	Vignan University
13	42187945	renuka	good in campus placements	3.26E+08	ecm	4	Vignan Womens
14	69845278	satish	college looks like big. Better join in vignan	2.75E+08	mech	2	Vignan
15	35487954	radha	sports meets are always going	1.66E+08	ecm	4	Vignan
16	47568475	kalyani.m	college is good for seminars	2.07E+08	mech	1	Vignan University
17	15489645	hemalatha	many workshops are conducting impressive	8.91E+08	eee	2	Vignan University
18	54878964	chitti babu	all HODs are very helpful	3.57E+08	eee	3	Vignan University
19	34587954	subhasini.k	some faculty are bad in vignan	3.88E+08	ecm	4	Vignan University
20	98758546	santhoshkumar	we can improve good communucation skills in college	1.8E+08	cse	2	Vignan University
21	50640564	prathyusha	PD is good in college	3.88E+08	mech	4	Vignan University
22	45065872	pravalika	I am enjoying the study in vignan	5.27E+08	ecm	2	Vignan
23	60549842	prasanna	faculty takes lead with students	4.35E+08	eee	4	Vignan
24	40657854	jyothi	cse Hod is very good	7.9E+08	cse	2	Vignan

Results and discussion

The performance of the proposed algorithm in Hadoop is accurate and processing time is less compared with different Traditional data mining tools and techniques. The performance evaluation of data in Hadoop is shown in Table 4. In this the CPU cumulative time is very less in processing the data while compared with traditional data mining tools.

Table 4: Performance Evaluation table

Size of data	Total input paths to process	No of splits	No of Mappers	No of Reducers	Cumulative CPU time
40960 bytes	1	2	3	1	6.56 sec
5234Kilobytes	29	29	29	1	191.29

Table 4 contains the information of size of data, no of splits, total no of mappers, no of reducers and cumulative CPU time. Table 5 shows how many people are posting good and bad comments about institutions. By analyzing this data we can provide ranking of the institutions.

Table 5: Comments like Good and Bad

t.tweet_id	t.name	t.text	s.id	s.branch	s.year	s.college
64873893	praveen kumar	vignan is good college for studies	1.33E+08	it	2	Vignan
54873893	prakash	Faculty is good in college	3.56E+08	ece	4	Vignan
84257698	parvathy	college looks good at outside	5.98E+08	ece	4	Vignan
45872694	sabeena grace	very good in conducting cultural programs	9.96E+08	ecm	2	Vignan Lara
58245869	hemanth	campus is good	4.36E+08	mech	4	Vignan Nirula
87548962	raghunadh	principal is too good	3.92E+08	cse	1	Vignan University
42187945	renuka	good in campus placements	3.26E+08	ecm	4	Vignan Womens
47568475	kalyani.m	college is good for seminars	2.07E+08	mech	1	Vignan University
98758546	santhoshkumar	we can improve good communucation skills in college	1.8E+08	cse	2	Vignan University
50640564	prathyusha	PD is good in college	3.88E+08	mech	4	Vignan University
40657854	jyothi	cse Hod is very good	7.9E+08	cse	2	Vignan
44402186	geethanjali	college buses are very good	2.28E+08	ecm	1	Vignan University
21054863	geetha	T&P is not good in VIIT	1.53E+08	cse	4	Vignan
48521703	rakesh	I.T brach is very good in vignan	3.42E+08	mech	2	Vignan
98421057	rajesh	vignan womens college very good	58379816	civil	1	Vignan University
35124860	fatima	pharmacy also having in vignan is good	9.8E+08	civil	4	Vignan
42571350	PANDU	every faculty is good with students	5.03E+08	ece	1	Vignan
65820150	naveen	we can't get good internal marks	9.82E+08	cse	2	Vignan
60142204	prashanth	some faculty only doing good for their job some are escaped	4.33E+08	it	1	Vignan
84502105	gopika	college atmosphere is too good	5.57E+08	ecm	2	Vignan Womens
95476210	ravi	college architecture is very good	6.01E+08	civil	2	Vignan University
t.tweet_id	t.name	t.text	s.id	s.branch	s.year	s.college
35849752	daniel	labs are not sufficient and bad	835115314	it	1	Vignan University
65848267	sriya	in vignan campus places are not provided and bad	706356914	ecm	3	Vignan University
34587954	subhasini.	some faculty are bad in vignan	387993367	ecm	4	Vignan University
60547824	saikrishna	library is very bad in viit	190122749	cse	2	Vignan
58426842	bhaskar	college premices is very bad	591891246	eee	4	Vignan Lara
20147856	uma	transport is very bad in viit	425045057	civil	3	Vignan
35021846	bhanu	sports department is very bad	720814702	ecm	4	Vignan
85476247	rajtarun	faculty is bad in viit	NULL	NULL	NULL	NULL
47985467	swamy	girls are too bad	NULL	NULL	NULL	NULL
78457854	kavitha	architecture is bad	NULL	NULL	NULL	NULL

Figure 3: Analysis of Information Gain

By analyzing the Table 5 like taking good and bad comments we provide the result like ranking of the institutions shown in Table 6.

Table 6: Result analysis table

S.no	Good	Bad	Institution	Rank
1	70%	30%	VIgnan	1
2	60%	40%	Vignan University	2
3	50%	50%	Vignan Womens	3
4	30%	70%	Vignan Nirula	4
5	20%	80%	Vignan Laura	5

Figure 4 shows the analysis of the results Table 6. The X-axis shows the information about the ranking of colleges. The Y-axis shows the percentage of the good and bad comments.

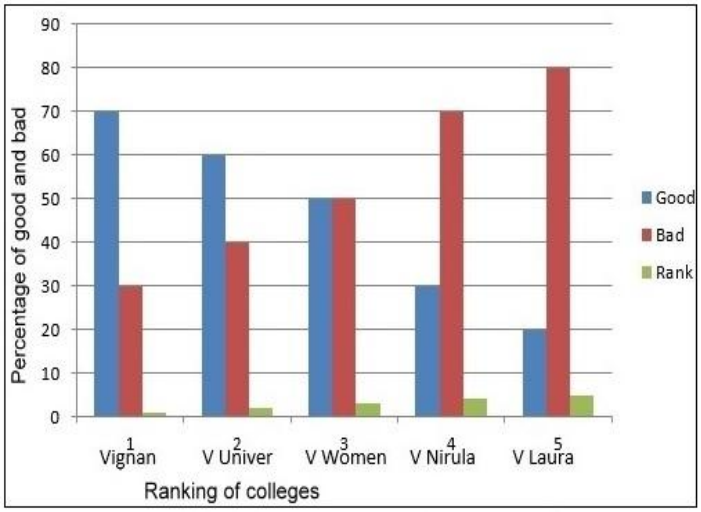

Figure 4: Analysis Graph of the Resulted data set

When the percentage of good increases then the ranking of college increases with the calculation in the framework. When the percentage of bad is increases then the ranking of colleges decreases.

Conclusion:-

We performed the paper using the K-Anonymity algorithm in Hadoop and acquired results fast and accurate. The key concept of K-Anonymity algorithm is the algorithm limits the ability to link or match published data with existing external information, those attributes in the information that could be used for linking with external data. It gives effective results in less time taken to process data. After analyzing the data we gave the ranking to the colleges.

Future work:-

Best functionality of the paper is processing Twitter and institutional data and provides the ranking of institutions. We would try to enhance the project by processing the data of all remaining institutions and provides ranking and categorization of the system in future based on different calculative attributes.

Acknowledgement:-

This work was supported partially by Vignan's Institute of Information Technology. We thank our Department of IT, who provided insight and expertise that greatly assisted the proposed system, although they may not agree with all of the interpretations/conclusions of this paper. We thank International journal of advanced research for assistance with k-Anonymity methodology in Hadoop and linger for comments that greatly improves the manuscript.

References:-

1. AnjaGruenheid, Edward Omiecinski (September 21-23, 2011), in IDEAS11, Lisbon, Portugal
2. Arik Friedman, Ran Wolff, Assaf Schuster (july 2008) in VLDB Journal, Vol 17,number 4
3. Efthymios Kouloumpis, Theresa Wilson, Johanna Moore Proceedings of the Fifth International, AAAI Conference on Weblogs and Social Media
4. G Jyothi, V.DurgaPrasada Rao, P.SureshBabu (September-October 2012) International Journal of Engineering Research and Applications (IJERA) ISSN:2248-9622. www.ijera.com Vol. 2, Issue 5, pp.793-795
5. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar/ (October 2014), International Journal of Scientific and Research Publications, Volume 4, Issue 10, 1 ISSN 2250-3153
6. Hortonworks, Inc. | 455 W. Maude Ave | Suite 200 | hortonworks.com
7. HussienSH March-2015 in International Journal Of Scientific & Engineering Research, Volume 6, Issue3, /ISSN2229-551
8. K.V.Kanimozhi1, Dr.M.Venkatesan (March 2015) in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3.
9. Ted Garcia and Taehyung ("George") Wang 2013 IEEE Seventh International Conference On SemanticComputing.
10. T.K.Das et al. / International Journal of Engineering and Technology (IJET)
11. Twinkle Antony et al, International Journal of Computer Science and Mobile Computing,Vol.3 Issue.2, February- 2014, pg. 459-462
12. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati (2007) Springer US,
13. Advances In Information Security