ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Recommendation System Based On Clustering and Collaborative Filtering

K. Dhanalakshmi, A.Anitha, G. Michael, K.G.S. Venkatesan

Dept. of C.S.E., Bharath University, Chennai, India

Dept. of C.S.E., Bharath University, Chennai, India

Associate Professor, Dept. of C.S.E., Bharath University, Chennai, India

Associate Professor, Dept. of C.S.E., Bharath University, Chennai, India

ABSTRACT: Cluster-based recommendation is best thought of as a variant on user-based recommendation. Instead of recommending items to users, items are recommended to clusters of similar users. This entails a pre processing phase, in which all users are partitioned into clusters. Recommendations are then produced for each cluster, such that the recommended items are most interesting to the largest number of users. The upside of this approach is that recommendation is fast at runtime because almost everything is pre computed. In this paper, we describe the problem of recommending conference sessions to attendees and show how novel extensions to traditional model-based recommender systems, as suggested in Adomavicius and Tuzhilin can address this problem. We introduce Recommendation Engine by Conjoint Decomposition of items and Users (RECONDITUS)-a technique that is an extension of preference-based recommender systems to recommend items from a new disjoint set to users from a new disjoint set.

I. INTRODUCTION

Wal-Mart claimed to have the largest data warehouse with 500 terabytes storage (equivalent to 50 printed collections of the US Library of Congress). In 2009, eBay storage amounted to eight petabytes (think of 104 years of HD-TV video). Two years later, the Yahoo warehouse totalled 170 petabytes1 (8.5 times of all hard disk drives created in 1995). Since the rise of digitisation, enterprises from various verticals have amassed burgeoning amounts of digital data, capturing trillions of bytes of information about their customers, suppliers and operations. Data volume is also growing exponentially due to the explosion of machine-generated data (data records, web-log files, sensor data) and from growing human engagement within the social networks [2]. The growth of data will never stop. According to the 2011 IDC Digital Universe Study, 130 exabytes of data were created and stored in 2005. The amount grew to 1,227 exabytes in 2010 and is projected to grow at 45.2% to 7,910 exabytes in 2015.3The growth of data constitutes the "Big Data" phenomenon – a technological phenomenon brought about by the rapid rate of data growth and parallel advancements in technology that have given rise to an ecosystem of software and hardware products that are enabling users to analyse this data to produce new and more granular levels of insight [5].

II. EXISTING SYSTEM

The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. The basic assumption of user-based CF is that people who agree in the past tend to agree again in the future [6].

Different with user-based CF, the item-based CF algorithm recommends a user the items that are similar to what he/she has preferred before In traditional CF algorithms, to compute similarity between every pair of users or services may take too much time, even exceed the processing capability of current RSs. Consequently, service recommendation based on the similar users or similar services would either lose its timeliness or could not be done at all [7].

10.15680/ijircce.2015.0303164

III. PROPOSED SYSTEM

We proposed a Agglomerative Hierarchal Clustering or Hierarchal Agglomerative Clustering. Clustering are such techniques that can reduce the data size by a large factor by grouping similar services together. A cluster contains some similar services just like a club contains some like-minded users. This is another reason besides abbreviation that we call this approach ClubCF [10]. This approach is enacted around two stages. In the first stage, the available services are divided into small-scale clusters, in logic, for further processing. At the second stage, a collaborative filtering algorithm is imposed on one of the clusters. This similarity metric computes the Euclidean distance *d between two such user points This value* alone doesn't constitute a valid similarity metric, because larger values would mean more-distant, and therefore less similar, users. The value should be smaller when users are more similar [12].

IV. ARCHITECTURE DIAGRAM

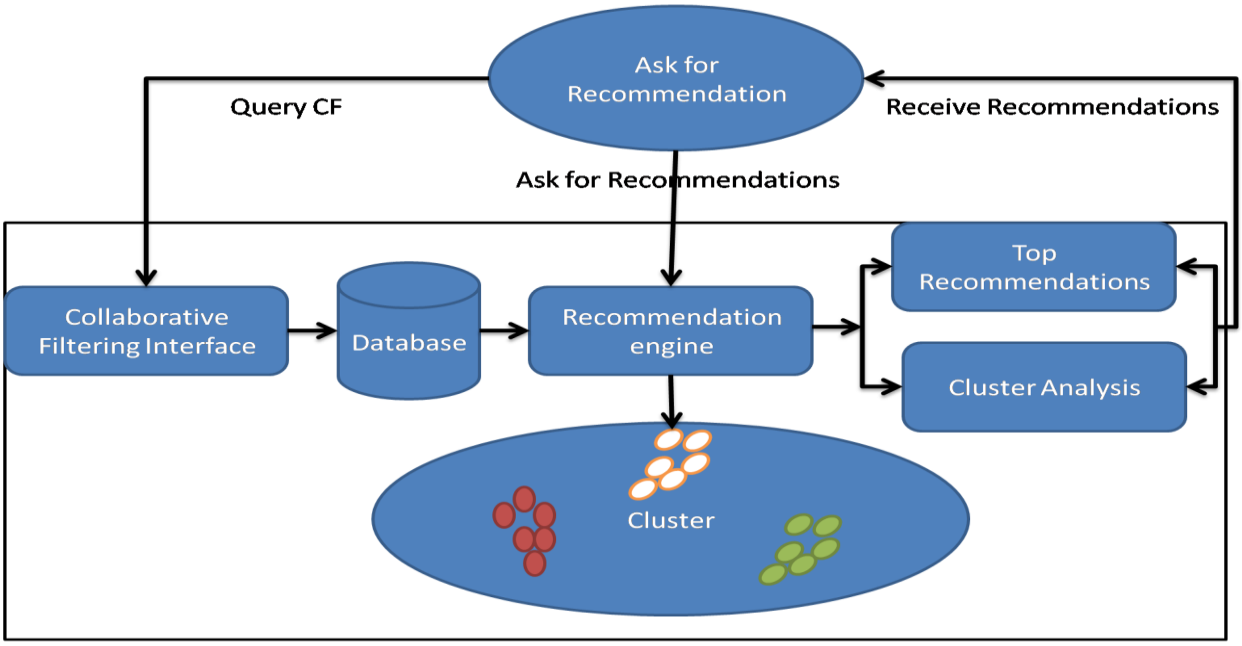


Fig 1: Architecture diagram for collaborative filtering interface with database cluster analysis.

V. MODULES

A. LOGIN AND ADD MOVIE DETAILS

The Login Form module presents site visitors with a form with username and password fields. If the user enters a valid username/password combination they will be granted access to additional resources on your application. Which additional resources they will have access to can be configured separately [15].

In this module admin can add new movie titles and their release date and their genre details. These details will added to the existing details. User can select the movie details added in this module and they will rate the movie based on their reviews. These details are used for cluster the data based on their ratings [19].

B. DATA PREPROCESSING:

The training data, we are given a list of vectors (u; m; r; t), where u is a user ID, m is a movie ID, r is the rating u gave to m, and t is the date. After training, we output predictions for a list of user-movie pairs. We measure error by using the root mean squared error. After pre processing we output the movie ids with the corresponding users and their ratings with ; separated files [20].

C. DATA CLUSTERING:

We cluster the people based on the movies they watched and then cluster the movies based on the people that watched them. The people can then be re-clustered based on the number of movies in each movie cluster they watched [25]. Movies can similarly be re-clustered based on the number of people in each person cluster that watched them. On the first pass, people are clustered based on movies and movies based on people. On the second, and subsequent passes, people are clustered based on movie clusters, and movies based on people clusters. A cluster contains some similar services just like a club contains some like-minded users [30].

D. RECOMMENDATION:

This similarity metric computes the Euclidean distance *d between two such user points This value* alone doesn't constitute a valid similarity metric, because larger values would mean more-distant, and therefore less similar, users [33]. The value should be smaller when users are more similar. Therefore, the implementation actually returns 1 / (1+*d).* The upside of this approach is that recommendation is fast at runtime because almost everything is pre computed. One could argue that the recommendations are less personal this way, because recommendations are computed for a group rather than an individual. This approach may be more effective at producing recommendations for new users, who have little preference data available [29].

VI. **ALGORITHM**

A. COLLABORATIVE FILTERING:

It is a technique used by some recommender systems. Collaborative filtering has two senses, a narrow one and a more general one. In general, collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc [35]. Applications of collaborative filtering typically involve very large data sets. Collaborative filtering methods have been applied to many different kinds of data including: sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors; financial data, such as financial service institutions that integrate many financial sources; or in electronic commerce and web applications where the focus is on user data, etc. The remainder of this discussion focuses on collaborative filtering for user data, although some of the methods and approaches may apply to the other major applications as well [36].

B. K-MEAN CLUSTERING:

Clustering is the process of partitioning a group of data points into a small number of clusters. For instance, the items in a supermarket are clustered in categories. Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have n data points xi,i=1...nthat have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μi,i=1...k of the clusters that minimize the *distance* from the data points to the cluster. K-means clustering solves

$$\mathrm{argminc}\sum i=1k\sum x\in cid(x,\mu i)=\mathrm{argminc}\sum i=1k\sum x\in ci // x-\mu i // 22$$

where ci is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance d(x,μi)= // x−μi // 22. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution [37].

VII. EXPERIMENTAL SETUP AND RESULT

A Hard drive of twenty G and a RAM memory of 512 MB (min) square measure used for the implementation. Java JDK 1.7 is employed because the front-end java and five.0 is employed because the back-end with MySQL [22]

A. SCREENSHOTS

Fig 1 login on user name

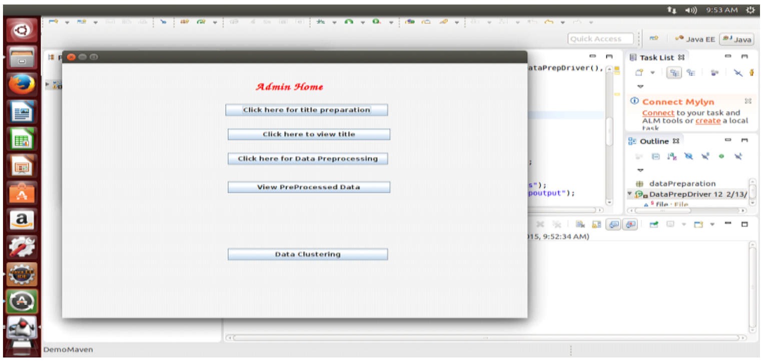
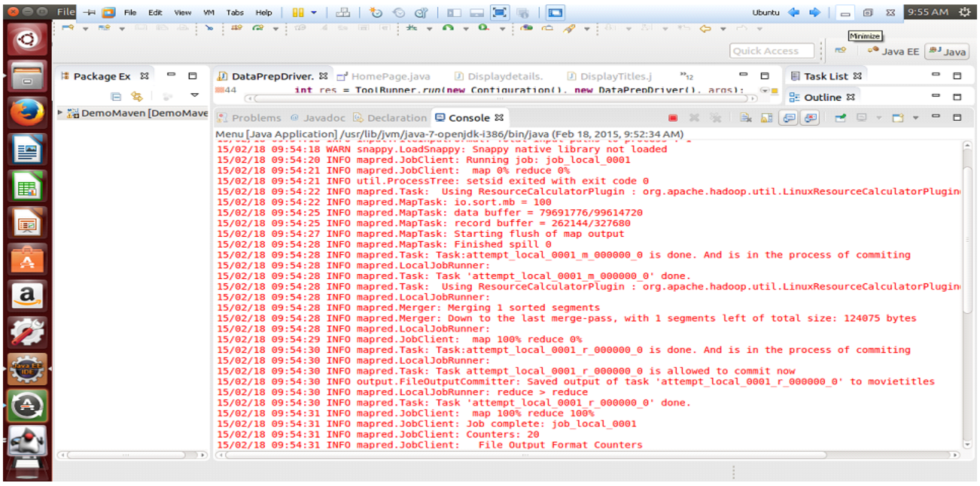


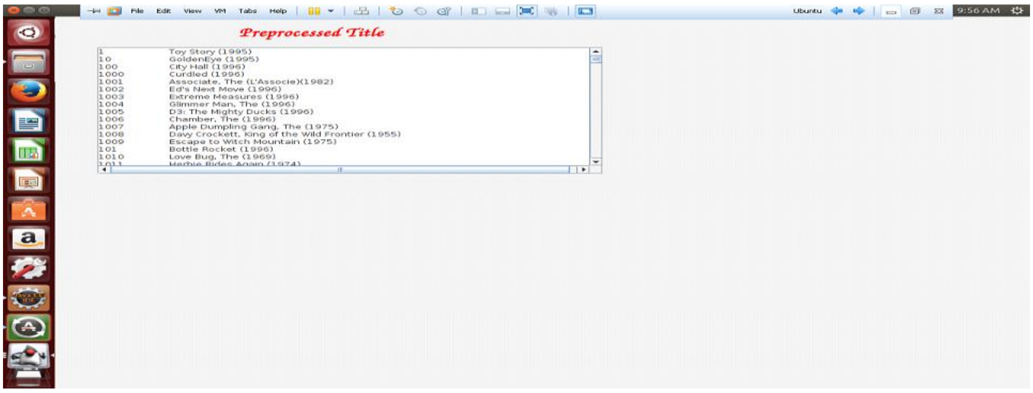
FIG .2 ADMIN HOME



```
Menu [Java Application] /usr/lib/jvm/java-7-openjdk-i386/bin/java (Feb 18, 2015, 9:52:34 AM)
15/02/18 09:54:18 WARN snappy.LoadSnappy: Snappy native library not loaded
15/02/18 09:54:20 INFO mapred.JobClient: Running job: job_local_0001
15/02/18 09:54:21 INFO mapred.JobClient:  map 0% reduce 0%
15/02/18 09:54:21 INFO util.ProcessTree: setsid exited with exit code 0
15/02/18 09:54:22 INFO mapred.Task:  Using ResourceCalculatorPlugin : org.apache.hadoop.util.LinuxResourceCalculatorPlugin
15/02/18 09:54:22 INFO mapred.MapTask: io.sort.mb = 100
15/02/18 09:54:25 INFO mapred.MapTask: data buffer = 79691776/99614720
15/02/18 09:54:25 INFO mapred.MapTask: record buffer = 262144/327680
15/02/18 09:54:27 INFO mapred.MapTask: Starting flush of map output
15/02/18 09:54:28 INFO mapred.MapTask: Finished spill 0
15/02/18 09:54:28 INFO mapred.Task: Task:attempt_local_0001_m_000000_0 is done. And is in the process of commiting
15/02/18 09:54:28 INFO mapred.LocalJobRunner:
15/02/18 09:54:28 INFO mapred.Task: Task 'attempt_local_0001_m_000000_0' done.
15/02/18 09:54:28 INFO mapred.Task:  Using ResourceCalculatorPlugin : org.apache.hadoop.util.LinuxResourceCalculatorPlugin
15/02/18 09:54:28 INFO mapred.LocalJobRunner:
15/02/18 09:54:28 INFO mapred.Merger: Merging 1 sorted segments
15/02/18 09:54:28 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 124075 bytes
15/02/18 09:54:28 INFO mapred.LocalJobRunner:
15/02/18 09:54:29 INFO mapred.JobClient:  map 100% reduce 0%
15/02/18 09:54:30 INFO mapred.Task: Task:attempt_local_0001_r_000000_0 is done. And is in the process of commiting
15/02/18 09:54:30 INFO mapred.LocalJobRunner:
15/02/18 09:54:30 INFO mapred.Task: Task attempt_local_0001_r_000000_0 is allowed to commit now
15/02/18 09:54:30 INFO output.FileOutputCommitter: Saved output of task 'attempt_local_0001_r_000000_0' to movietitles
15/02/18 09:54:30 INFO mapred.LocalJobRunner: reduce > reduce
15/02/18 09:54:30 INFO mapred.Task: Task 'attempt_local_0001_r_000000_0' done.
15/02/18 09:54:31 INFO mapred.JobClient:  map 100% reduce 100%
15/02/18 09:54:31 INFO mapred.JobClient: Job complete: job_local_0001
15/02/18 09:54:31 INFO mapred.JobClient: Counters: 20
15/02/18 09:54:31 INFO mapred.JobClient:   File Output Format Counters
```

```
44    int res = ToolRunner.run(new Configuration(), new DataPrepDriver(), args);
```

FIG NO .3 ADMIN CODING



Preprocessed Title

```
1           Toy Story (1995)
10          GoldenEye (1995)
100         City Hall (1996)
1000        Curdled (1996)
1001        Associate, The (L'Associe)(1982)
1002        Ed's Next Move (1996)
1003        Extreme Measures (1996)
1004        Glimmer Man, The (1996)
1005        D3: The Mighty Ducks (1996)
1006        Chamber, The (1996)
1007        Apple Dumpling Gang, The (1975)
1008        Davy Crockett, King of the Wild Frontier (1955)
1009        Escape to Witch Mountain (1975)
101         Bottle Rocket (1996)
1010        Love Bug, The (1969)
1011        Herbie Rides Again (1974)
```

Fig no .4 preprocessed title

The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. The basic assumption of user-based CF is that people who agree in the past tend to agree again in the future.

XI. CONCLUSION

In this paper, we present a ClubCF approach for big data applications relevant to service recommendation. Before applying CF technique, services are merged into some clusters via an AHC algorithm. Then the rating similarities between services within the same cluster are computed. As the number of services in a cluster is much less than that of in the whole system, ClubCF costs less online computation time. Moreover, as the ratings of services in the same cluster are more relevant with each other than with the ones in other clusters, prediction based on the ratings of the services in the same cluster will be more accurate than based on the ratings of all similar or dissimilar services in all clusters. These two advantageous of ClubCF have been verified by experiments on real-world data set.

X. ACKNOWLEDGEMENT

The author would like to thank the Vice Chancellor, Dean-Engineering, Director, Secretary, Correspondent, HOD of Computer Science & Engineering, Dr. K.P. Kaliyamurthie, Bharath University, Chennai for their motivation and constant encouragement. The author would like to specially thank **Dr. A. Kumaravel**, **Dean , School of Computing**, for his guidance and for critical review of this manuscript and for his valuable input and fruitful discussions in completing the work and the Faculty Members of Department of Computer Science &Engineering. Also, he takes privilege in extending gratitude to his parents and family members who rendered their support throughout this Research work.

REFERENCES

1. R. S. Sandeep, C. Vinay, S. M. Hemant, “Strength and Accuracy Analysis of Affix Removal Stemming Algorithms,” International Journal of Computer Science and Information Technologies.
2. M. C. Pham, Y. Cao, R. Klamma, et al., “A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis,” *Journal of Universal Computer Science*, Vol. 17, No. 4,PP. 583-604, April 2011 .
3. N. Mittal, R. Nayak, M. C. Govil, et al., “Recommender System Framework using Clustering and Collaborative Filtering,” in Proc. 3rd Int’l Conf. on Emerging Trends in Engineering and Technology, pp. 555-558, November 2010 .
4. S. Agarwal, and A. Nath, “A study on implementing Green IT in Enterprise 2.0,” International Journal of Advanced Computer Research, Vol. 3, No. 1,PP. 43-49, March 2013 .
5. X. Li, and T. Murata. “Using Multidimensional Clustering Based Collaborative Filtering Approach Improving Recommendation Diversity,” in Proc. 2012 IEEE/WIC/ACM Int’l Joint Conf. on Web Intelligence and Intelligent Agent Technology, pp. 169-174, December 2012
6. X. Liu, Y. Hui, W. Sun, et al., “Towards service composition based on mashup,” in Proc. of IEEE Congress on Services .
7. R. D. Simon, X. Tengke, and W. Shengrui, “Combining collaborative filtering and clustering for implicit recommender system,” in Proc. 2013 IEEE 27th Int’l Conf. on. Advanced Information Networking .
8. K.G.S. Venkatesan. Dr. V. Khanna, Dr. A. Chandrasekar, “Autonomous System(AS) for mesh network by using packet transmission & failure detection”, Inter. Journal of Innovative Research in computer & comm. Engineering, Vol. 2, Issue 12, PP. 7289 – 7296, December - 2014.
9. K.G.S. Venkatesan and M. Elamurugaselvam, “Design based object oriented Metrics to measure coupling & cohesion”, International journal of Advanced & Innovative Research, Vol. 2, Issue 5, PP. 778 – 785, 2013.
10. Teerawat Issariyakul • Ekram Hoss, “Introduction to Network Simulator NS2”.
11. S. Sathish Raja and K.G.S. Venkatesan, “Email spam zombies scrutinizer in email sending network Infrastructures”, International journal of Scientific & Engineering Research, Vol. 4, Issue 4, PP. 366 – 373, April 2013.
12. G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” IEEE J. Sel. Areas Communication., Vol. 18, No. 3, PP. 535–547, Mar. 2000.
13. K.G.S. Venkatesan, “Comparison of CDMA & GSM Mobile Technology”, Middle-East Journal of Scientific Research, 13 (12), PP. 1590 – 1594, 2013.
14. P. Indira Priya, K.G.S.Venkatesan, “Finding the K-Edge connectivity in MANET using DLTRT, International Journal of Applied Engineering Research, Vol. 9, Issue 22, PP. 5898 – 5904, 2014.
15. K.G.S. Venkatesan and M. Elamurugaselvam, “Using the conceptual cohesion of classes for fault prediction in object-oriented system”, International journal of Advanced & Innovative Research, Vol. 2, Issue 4, PP. 75 – 80, April 2013.
16. Ms. J.Praveena, K.G.S. Venkatesan, “Advanced Auto Adaptive edge-detection algorithm for flame monitoring & fire image processing”, International Journal of Applied Engineering Research, Vol. 9, Issue 22, PP. 5797 – 5802, 2014.
17. K.G.S. Venkatesan. Dr. V. Khanna, “Inclusion of flow management for Automatic & dynamic route discovery system by ARS”, International Journal of Advanced Research in computer science & software Engg., Vol.2, Issue 12, PP. 1 – 9, December – 2012.

18. Needhu. C, K.G.S. Venkatesan, "A System for Retrieving Information directly from online social network user Link ", International Journal of Applied Engineering Research, Vol. 9, Issue 22, PP. 6023 – 6028, 2014.

19. K.G.S. Venkatesan, R. Resmi, R. Remya, "Anonymizimg Geographic routing for preserving location privacy using unlinkability and unobservability", International Journal of Advanced Research in computer science & software Engg., Vol. 4, Issue 3, PP. 523 – 528, March – 2014.

20. Selvakumari. P, K.G.S. Venkatesan, "Vehicular communication using Fvmr Technique", International Journal of Applied Engineering Research, Vol. 9, Issue 22, PP. 6133 – 6139, 2014.

21. K.G.S. Venkatesan, G. Julin Leeya, G. Dayalin Leena, "Efficient colour image watermarking using factor Entrenching method", International Journal of Advanced Research in computer science & software Engg., Vol. 4, Issue 3, PP. 529 – 538, March – 2014.

22. T. Niknam, E. Taherian Fard, N. Pourjafarian, et al., "An efficient algorithm based on modified imperialist competitive algorithm and K-means for data clustering," Engineering Applications of Artificial Intelligence.

23. K.G.S. Venkatesan. Kausik Mondal, Abhishek Kumar, "Enhancement of social network security by Third party application", International Journal of Advanced Research in computer science & software Engg., Vol. 3, Issue 3, PP. 230 – 237, March – 2013.

24. Annapurna Vemparala, Venkatesan.K.G., "Routing Misbehavior detection in MANET'S using an ACK based scheme", International Journal of Advanced & Innovative Research, Vol. 2, Issue 5, PP. 261 – 268, 2013.

25. K.G.S. Venkatesan. Kishore, Mukthar Hussain, "SAT : A Security Architecture in wireless mesh networks", International Journal of Advanced Research in computer science & software Engineering, Vol. 3, Issue 3, PP. 325 – 331, April – 2013.

26. Annapurna Vemparala, Venkatesan.K.G., "A Reputation based scheme for routing misbehavior detection in MANET"S ", International Journal of computer science & Management Research, Vol. 2, Issue 6, June - 2013.

27. K.G.S. Venkatesan, "Planning in FARS by dynamic multipath reconfiguration system failure recovery in wireless mesh network", International Journal of Innovative Research in computer & comm. Engineering, Vol. 2, Issue 8, August -2014.

28. K.G.S. Venkatesan, AR. Arunachalam, S. Vijayalakshmi, V. Vinotha, "Implementation of optimized cost, Load & service monitoring for grid computing", International Journal of Innovative Research in computer & comm. Engineering, Vol. 3, Issue 2, PP. 864 – 870, February -2015.

29. R. Karthikeyan, K.G.S. Venkatesan, M.L. Ambikha, S. Asha, "Assist Autism spectrum, Data Acquisition method using Spatio-temporal Model", International Journal of Innovative Research in computer & communication Engineering, Vol. 3, Issue 2, PP. 871 – 877, February -2015.

30. K.G.S. Venkatesan, B. Sundar Raj, V. Keerthiga, M. Aishwarya, "Transmission of data between sensors by devolved Recognition", International Journal of Innovative Research in computer & comm. Engineering, Vol. 3, Issue 2, PP. 878 – 886, February -2015.

31. K.G.S. Venkatesan, N.G. Vijitha, R. Karthikeyan, "Secure data transaction in Multi cloud using Two-phase validation", International Journal of Innovative Research in computer & comm. Engineering, Vol. 3, Issue 2, PP. 845 – 853, February -2015.

32. K.G.S. Venkatesan, "Automatic Detection and control of Malware spread in decentralized peer to peer network", International Journal of Innovative Research in computer & comm. Engineering, Vol. 1, Issue 7, PP. 15157 – 15159, September -2013.

33. Satthish Raja, S K.G.S. Venkatesan, "Electronic Mail spam zombies purify in email connection", International Journal of Advanced Research in Computer Science Engineering & Information Technology, Vol. 1, Issue 1, PP. 26 – 36, June – 2013.

34. K.G.S. Venkatesan. Dr. V. Khanna, S.B. Amarnath Reddy, "Providing Security for social Networks from Inference Attack", International Journal of Computer Science Engineering & Scientific Technology, March – 2015.

35. K.G.S. Venkatesan, Dr. Kathir. Viswalingam, N.G. Vijitha, " Associate Adaptable Transactions Information store in the cloud using Distributed storage and meta data manager", International Journal of Innovative Research in computer & communication Engineering, Vol. 3, Issue 3, PP. 1548 – 1555, March -2015.

36. M. A. Beyer and D. Laney, "The importance of "big data": A definition," Gartner, Tech. Rep., 2012.

37. X. Wu, X. Zhu, G. Q. Wu, et al., "Data mining with big data," IEEE Trans. on Knowledge and Data Engineering.