Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 8, August 2015, pg.17 – 25

RESEARCH ARTICLE

Classification of Big Data Through Artificial Intelligence

Divya, Amit Jain, Gagandeep Singh

Panchkula Engineering College, Mouli, Panchkula, Haryana & Kurukshetra University
Panchkula Engineering College, Mouli, Panchkula, Haryana & Kurukshetra University
Panchkula Engineering College, Mouli, Panchkula, Haryana & Kurukshetra University
divyash.sharma74@rediffmail.com ; amit014@gmail.com

*Abstract— By **technology innovations, there has been a large increase within the utilization of Bigdata knowledge, joined of the foremost most well-liked styles of media thanks to its content richness, for several vital applications. To sustain Associate in Nursing current ascension of knowledge Bigdata, there's Associate in Nursing rising demand for a complicated content-based knowledge classification system. Thanks to the chop-chop increasing massive knowledge, abundant analysis effort has been dedicated to develop classification primarily based massive knowledge retrieval ways which may efficiently retrieve knowledge of interest. Considering the restricted man-power, it's abundant expected to develop retrieval ways that use options mechanically extracted from massive knowledge. Through Architecture-Algorithm co-design for Bigdata processing Applications, a scalable. Manycore processor consists of classification of heterogeneous cores with stream process capabilities, and zero-overhead inter-process communication through computer science with a hardware-software mechanism has been designed. This is often designed for achieving superior and low-power consumption, particularly thus on cut back access needed for Bigdata processing Applications.***

Keywords— classification , Bigdata , PBO(pollination based optimization) , BBO(biogeography based optimization) , Apriori.

© 2015, IJCSMC All Rights Reserved

INTRODUCTION

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business. To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real time and can protect data privacy and security. There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology:

A. Operational Big Data

This includes systems like Mongo DB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored. NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

B. Analytical Big Data

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data. MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines. These two classes of technology are complementary and frequently deployed together.

BENEFITS OF BIG DATA

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

USING THE INFORMATION KEPT IN THE SOCIAL NETWORK LIKE FACEBOOK, THE MARKETING AGENCIES ARE LEARNING ABOUT THE RESPONSE FOR THEIR CAMPAIGNS, PROMOTIONS, AND OTHER ADVERTISING MEDIUMS.

USING THE INFORMATION IN THE SOCIAL MEDIA LIKE PREFERENCES AND PRODUCT PERCEPTION OF THEIR CONSUMERS, PRODUCT COMPANIES AND RETAIL ORGANIZATIONS ARE PLANNING THEIR PRODUCTION.

USING THE DATA REGARDING THE PREVIOUS MEDICAL HISTORY OF PATIENTS, HOSPITALS ARE PROVIDING BETTER AND QUICK SERVICE.

REVIEW

Behrouz et. al.[15] A combination of multiple classifiers leads to a significant improvement in classification performance. Furthermore, by learning an appropriate weighting of the features used via a genetic algorithm (GA), we further improve prediction accuracy. The GA is demonstrated to successfully improve the accuracy of combined classifier performance, about

© 2015, IJCSMC All Rights Reserved

10 To 12% when comparing to non-GA classifier. This method may be of considerable usefulness in identifying students at risk early, especially in very large classes, and allow the instructor to provide appropriate advising in a timely manner. **Riccardo et al. [14]** proposed cognitive, and behavioural aspects of distance students. Course Vis is presented in the paper, and several examples of pictorial representations generated by the tool. **Luo et. al. [21]** Efficient meaning for sampling of data, reduction of data also needed to develop. Newly develop mining technique and searching algorithms that are suitable for extracting more different or complex relationship between fields.

Youssef M.ESSA et. al. [25] The proposed framework is developed by using mobile agent and MapReduce paradigm under Java Agent Development Framework (JADE). JADE is a promising middleware based on the agent paradigm because it supports generic services such as communication support, resource discovery, content delivery, data encoding and agents mobility. Indeed, there are seven reasons for using mobile agents as follows:

(1) Reduce the network load,

(2) Overcome network latency,

(3) Encapsulate protocols,

(4) Execute asynchronously and autonomously,

(5) Adapt dynamically,

(6) Naturally heterogeneous and robust, and

(7) Fault-tolerant

So, the mobile agent is used with Hadoop to overcome the problems faced Hadoop. In the proposed strategy, mobile agents send both code and data to any machine. The machine can react dynamically for any changes in the environment. Furthermore, if a machine or environment down, the mobile agent can migrate to another machine with code and data.

PROPOSED DESIGN

The first step is acquiring the data set of our choice. Data set can be in form of numeric or text or others. Then Association rule mining is done using Apriori algorithm. Apriori algorithm basically consists of two steps. First is CANDIDATE GENERATION phase and the second is count of SUPPORT factor with THRESHOLD. If Support factor is greater than Threshold, Positive rules are generated and are accepted, otherwise Negative rules are generated and are rejected. The rules which are Positive are valid rules and rules which are declared negative by Support and Threshold are invalid .Valid i.e. Positive rules are presented as a input for Pollination based optimization. Further, rules can be changed on the basis of REPRODUCTION FECTOR (R) based on given formulae.

$$R= [(A\times D)/(\text{œ} + A\times D)] +[((\text{œ}/(\text{œ} + A\times D)\times N^{P})/(A^{P} + N^{P}))] – C(N + D)$$

On the basis of Pollination based Optimization Classification is performed and rules will be optimized and then results will be compared with BBO.

© 2015, IJCSMC All Rights Reserved

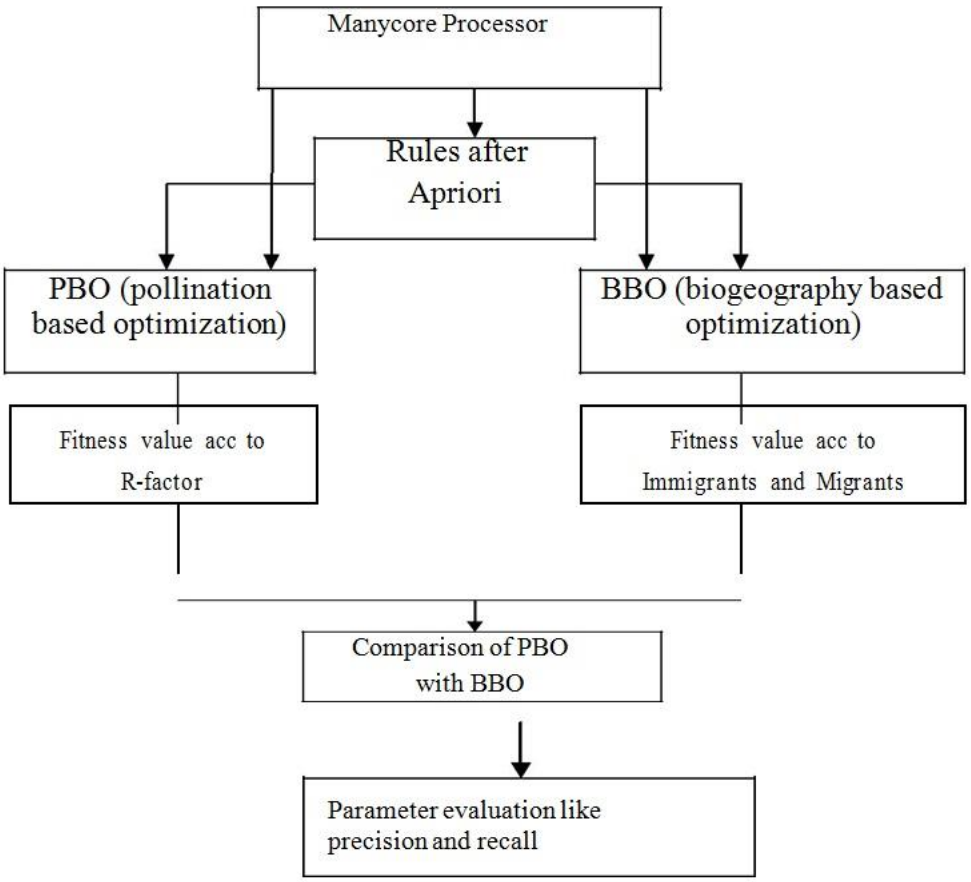

Figure 1: Design of Proposed Work

Optimization is a natural process embedded in the living beings. Pollination is a process of transfer of pollen from male parts of flower called anther to the female part called stigma of a flower. Self pollination results in the development of some flower seeds, when flower of one plant gives the pollen and pistil both then it is said to be a self pollination and when the pollen and pistil are from different flowers of different plants then it is said to be a cross pollination. Pollinators are responsible for movement of the pollens which are further responsible for the reproduction by setting of seeds. But pollinators don't have any information about the benefit of the plant. Pollination in plants is done for the energy requirement and to produce new plants. The floral display, fragrance and nectar lure pollinators and leads to pollination. Some species of plants optimize their nectar, display and fragrance producing resources. If pollination process is proceeding smoothly the plants spend average resources. If pollination process is above normal the plants reduce expenditure on resources for producing nectar, floral display and fragrance in the flowers. If the pollination success goes below normal, plants increase the resource expenditure such that more floral display, fragrance and nectar to attract pollinator. As more pollinators and their number of visits increase the pollination success rate increases.

PBO ALGORITHM IN DATA MINING:

Step 1: Initialize PBO Parameter

i. Initial Population=Number of Plants = no. of attributes
ii. Number of weeks = no. of iterations
iii. Number of seasons = pollination value after every iteration depends on R
iv. Pollination weekly goal= cluster formation

© 2015, IJCSMC All Rights Reserved

Step 2: Randomly generate vectors.

i. For season = 1 : number of seasons (iterations)
ii. For week = 1: number of weeks
iii. For k = 1: number of plants

Step 3: Evaluate Reproduction Vector(R)

$R= [(A\times D)/(\text{œ} + A\times D)] +[((\text{œ}/(\text{œ} + A\times D)\times N^P)/(A^P + N^P))] – C(N + D)$

Step 4: Based on R, update number of seasons.

Evaluate Error = Goal – R

Step 5: Based upon error update N, D, A

Step 6: Exit, if Error acceptable.

After implementing PBO algorithm, we have implemented BBO algorithm and basic flow of this comparison work is given below. Rules generated by Apriori algorithm are sent as a input for PBO and BBO algorithm for optimization. After getting optimized rules from both of the optimization algorithms, these results are compared by using parameters like Precision, Recall, F-Measure.

Results

Now, we will see the various results ,i.e., output of apriori algorithm with PBO, then we shall study the graph of support vs. confidence .Similarly, we study the output of apriori algorithm with BBO, then we shall study the graph of support vs confidence. In last we compare the various results obtained from Apriori algorithm with PBO, Apriori algorithm with BBO and Apriori only through comparing precision, recall and F-factor.

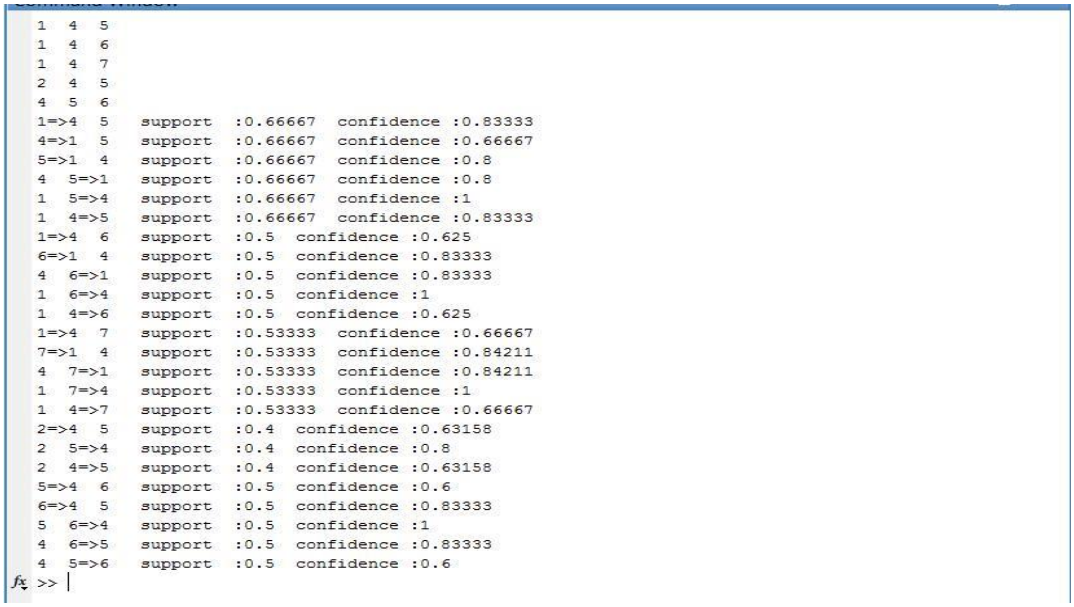

Figure 2: Result of PBO + Apriori Algorithm

© 2015, IJCSMC All Rights Reserved

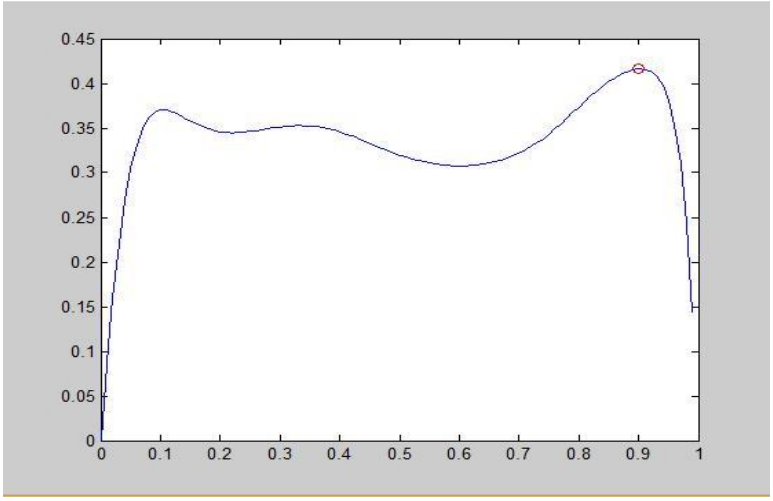

Figure 3: Graph of algorithm is shown above. The vertical line of graph represents Support and horizontal line represents Confidence. As the value of confidence increases, the fitness vector also increases.

RESULTS OF BBO+APRIORI ALGORITHM

```
Simulation # 1/3 - PopSize = 4
Simulation # 2/3 - PopSize = 16
Simulation # 3/3 - PopSize = 64
mean min cost = 41.1711      21.4938      17.2331
mean generation count = 37.27           100           100
# samples = 45
mean improvement segment generations  = 4.2667      5.5778
mean improvement amounts = 6.0797      6.2997
# samples = 46
mean improvement segment generations  = 15.7174      16.3478         13.5      12.3261
mean improvement amounts = 3.9618      5.3382      4.8117      4.8949
# samples = 49
mean improvement segment generations  = 23.0612      20.4898      22.8367
mean improvement amounts = 4.2717      3.4287      4.0039
Simulation # 1/3 - PopSize = 4
Simulation # 2/3 - PopSize = 16
Simulation # 3/3 - PopSize = 64
mean min cost = 1073.6435      594.54324       478.4415
mean generation count = 34.91           100           100
# samples = 40
mean improvement segment generations  = 5.025         4.4
mean improvement amounts = 88.52186      125.8258
# samples = 42
mean improvement segment generations  = 15.0952      12.2381       11.881      13.9286
mean improvement amounts = 108.9235      109.4561       82.9641      107.2277
# samples = 42
mean improvement segment generations  = 21.0714      23.1667       16.881
mean improvement amounts = 111.0086      97.23543      95.82203
Simulation # 1/3 - PopSize = 4
Simulation # 2/3 - PopSize = 16
Simulation # 3/3 - PopSize = 64
```

```
Simulation # 1/3 - PopSize = 4
Simulation # 2/3 - PopSize = 16
Simulation # 3/3 - PopSize = 64
mean min cost = 12.8434      3.37037      2.31475
mean generation count = 37.16           100           100
# samples = 44
mean improvement segment generations  = 6.0909         4.5
mean improvement amounts = 4.5254      5.6123
# samples = 42
mean improvement segment generations  = 13.9524      13.0952      13.8333       13.619
mean improvement amounts = 1.9449      2.2709      2.0923      1.9036
# samples = 40
mean improvement segment generations  = 19.325       21.525        17.95
mean improvement amounts = 1.4616      1.1178      1.2573
Simulation # 1/3 - PopSize = 4
Simulation # 2/3 - PopSize = 16
Simulation # 3/3 - PopSize = 64
mean min cost = 3990.69       1264.55        811.09
mean generation count = 35.65           100           100
# samples = 41
mean improvement segment generations  = 6.2927      4.8293
mean improvement amounts = 2003.7561       1244.878
# samples = 46
mean improvement segment generations  = 15.587      15.5652      14.7391       11.913
mean improvement amounts = 726.1739      661.6522      843.4783      724.6304
# samples = 40
mean improvement segment generations  = 23.5        19.95       18.825
mean improvement amounts = 486.925         669.3       587.025
Simulation # 1/3 - PopSize = 4
Simulation # 2/3 - PopSize = 16
```

© 2015, IJCSMC All Rights Reserved

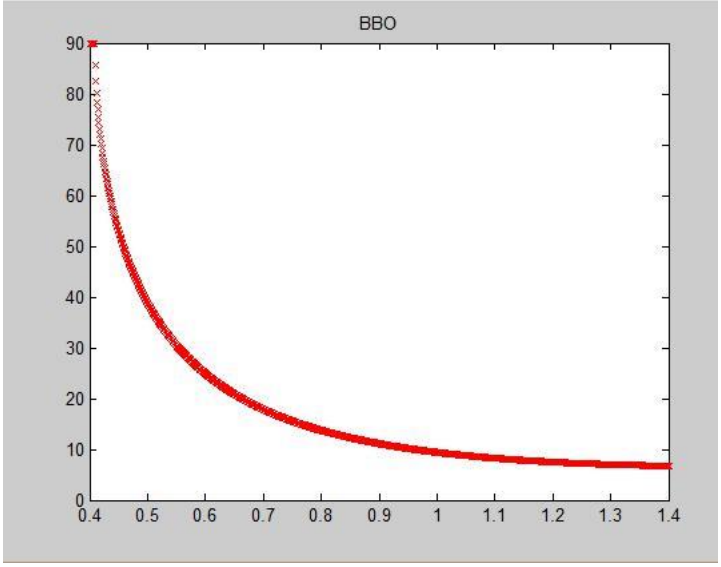

Figure 4: Graph of BBO represents Support Vs Confidence. The vertical line represents Support and horizontal line represents Confidence. As value of Support factor increases, it represents fitness vector high. Optimum value of each sample is plotted.

COMPARISON OF (PBO +APRIORI) ALGORITHM WITH (BBO+APRIORI) AND APRIORI

Data name	Performance Measures	Apriori	Modified BBO+Apriori	Modified PBO+Apriori
Dataset 1	Precision	.20	.32	.41
	Recall	.29	.31	.40
	F-measure	.1634	.2910	.3190
Dataset 2	Precision	.32	.36	.39
	Recall	.26	.34	.46
	F-measure	.4008	.6221	.7890
Dataset 3	Precision	.49	.52	.64
	Recall	.31	.35	.40
	F-measure	.5087	.8012	.9490

Table 1: Comparison of PBO+Apriori, BBO+Apriori and Apriori

© 2015, IJCSMC All Rights Reserved

CONCLUSION:

After evaluation of parameters like precision, recall and f-measure, we analyze artificial intelligence are better optimizer and can reduce calculation with efficient results for big data classification in comparison to traditional systems and methodology. We have also compared the results of apriori algorithm; apriori with PBO; apriori with BBO where apriori with PBO gives the best results. So we can conclude it is a unique blend of a classification algorithm, parallelism and artificial intelligence,

FUTURE SCOPE

In future we can focus on security of the big data like for security we can apply new secure algorithms like RSA and Hill Cipher methods. For further security we can even try password secure methods.

REFERENCES

[1]AsmaaBenghabrit, BrahimOuhbi, HichamBehja"Text Clustering Using Statistical and Semantic Data"-IEEE 2013.

[2]Marisa S. Viveros, John P. Nearhos, Michael J. Rothman-"Applying Data Mining Techniques to a Health Insurance Information System".2012

[3]Linna Li, Bingru Yang, Faguo Zhou "A Framework for Object-Oriented Data Mining"-Fifth International Conference on Fuzzy Systems and Knowledge discovery2010.

[4]DezhenFeng, ZaimeiZhang, FangZhou, JianhengJi[4] "Application Study of Data Ming on Customer Relationship Management in E-commerce".

[5]BhavaniThuraisingham"Data Mining for Malicious Code Detection and SecurityApplications"-IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops 2012.

[8] He YueShun, Ding QiuLin "Application research of data mining architecture for intelligent decision" -Asia-Pacific Conference on Information Processing 2009

[12] Luo Fang, QiuQizhi "The Study on the Application of Data Mining Based on Association Rules" -International Conference on Communication Systems and Network Technologies (2012).

[13] Xindong Wu "Data Mining: Artificial Intelligence in Data Analysis" -Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology 2004

[14] Ricardo Mazza, Vania Dimitrova "Course Vis: Externalizing Student Information to Facilitate Instructors in Distance Learning" Proceedings of the International conference in Artificial Intelligence in Education.2004

[15] Behrouz Minaei-Bidgoli, Deborah A. Kashy, GerdKortemeyer, William F. Punch

"Predicting Student Performance: An Application Of Data Mining Methods With The Educational Web-Based System Lon-Capa" 33rd ASEE/IEEE Frontiers in Education

Conference 2003

[16] ShusakuTsumoto, ShogiHirano, YukoTsumoto."Towards Data-Oriented Hospital Services: Data Mining in Hospital Information System." IEEE 2011.

[18] Heiner, C., Beck, J. E., &Mostow, J.(2004, June 17-19). Improving the Help Selection Policy in a Reading Tutor that Listens. Proceedings of the InSTIL/ICALL Symposium on NLP and Speech Technologies in Advanced Language Learning Systems.

[19]http://www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools.

© 2015, IJCSMC All Rights Reserved

[20] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html2006

[21]Qi Luo "Advancing Knowledge Discovery and Data Mining"-Workshop on Knowledge Discovery and Data Mining 2008.

[22] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE "Data Mining with Big Data" IEEE Transactions on knowledge and Data Engineering, Vol. 26, No. 1, JANUARY 2014

[23] Zaiying Liu, Ping Yang, Lixiao Zhang School of Information Science and Technology Shanghai Sanda University Shanghai, China "A Sketch of Big Data Technologies" 2013 Seventh International Conference on Internet Computing for

Engineering and Science.

[24] Jinsong Zhang, Yan Chen, Taoying Li College of Transportation Management Dalian Maritime University Dalian, China," Opportunities of Innovation under Challenges of Big Data" 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)

[25] Youssef M. ESSA Software Engineering Department, Etisalat Corporation Cairo,Egypt " Mobile Agent based New Framework for Improving Big Data Analysis" 2013International Conference on Cloud Computing and Big Data.

© 2015, IJCSMC All Rights Reserved