Optimized Multi Class SVM Classifier for Named Entity Extraction for Workflow Scheduling in Cloud

Jyothi Bellary and E. Keshava Reddy

***Abstract*—This paper deals with the optimized Multi class SVM classifier (OMSC) with Named entity extraction in cloud environment. The proposed OMSC handles with scheduling workflow in cloud computing where the data and files are transferred between the participants based on the different set of rules, with having the additional advantage of rules formation capability in which it is going to follow 22 rules templates. This has shown an improved performance against the traditional Multi class SVM classifier. The average f-score for the tested data sets is 81.04 % when compared to the existing classifier it is improved sign. The time complexity is decreased and as per the scheduling is concerned the execution time and response time is improved.**

***Index Terms*—OMSC, workflow model, NE extraction and rules formation.**

I. Introduction

The number of available multi-media material has grown rapidly, because latest developments in cellular phones have enabled customers to quickly make their own material, called User Created Contents (UCC). These gadgets began to evolve to a context-aware one, so called as individual memory aid, to relieve storage issues. Improved portability and ubiquity follows that more and more market and academic categories are analyzing the chance of recording everything in one's everyday life. These personal recordings existing the technological task that search of contents must be finished on included ends, because the customers aim to prevent distribution of their personal information across the web servers.

Customers are progressively hanging out looking for contents, yet, until now, they cannot experience an automatic technique to easily accessibility related contents. Current search engines find contents based on their meta-data. Related content offered by content providers are effectively modified to have meta-data, so that they are relatively quickly discovered by Google. However, related content documented for storage aid reasons is normally unstructured and does not contain meta-data.

Named Entities (NEs), such as names of persons and locations, are regularly queried entities and bring the core details in data search. Therefore, offering an efficient method for NE extraction is very essential for related content search.

Manuscript received May 14, 2015; revised August 20, 2015.

Jyothi Bellary is with Aditya College of Engineering, Madanapalle, India (e-mail: jyothibellary@gmail.com).

E Keshava Reddy was with Jawaharlal Nehru Technological University Ananthapuramu. He is now with the Department of Mathematics JNTUCEA ,Jawaharlal Nehru Technological University Ananthapuramu, Ananthapuramu, India (e-mail: keshava_e@rediffmail.com).

NE process was first described at the Sixth Message Understanding Conference, as the "Recognition of names of individuals, places, and organization, as well as temporary and number expressions" [1] (B. Sundheim 1995). Fig. 1 defines the example text of NE tags.

```
The <ENAMEX TYPE="ORGANIZATION"> Jawaharlal technological university </ENAMEX> will hold a confirmation hearing <TIMEX TYPE="DATE"> next Friday </TIMEX> for vice chancellor. <ENAMEX TYPE="PERSON"> Ganesh</ENAMEX>.
```

Fig. 1. Example text of named entity tags.

Although these tasks seem obvious, the appropriate response is not correct in some situations due to the indecisiveness in natural language. For NE extraction, examples are mentioned in [2] (F. Kubala *et al* 1998) as "When is the Wall Street is opened, and when is it an organization or a place?", "When is the Richmond an organization, and when is it a location?". The system must generate a single, unambiguous outcome for any appropriate sequence in the writing. In order to motivate reliability and reduce indecisiveness regarding NE extraction, recommendations have been described in [3] (N. Chinchor 1997).

The jobs of NE extraction and of improved content extraction, such as automatic punctuation and capitalization creation, are substantially related to each other, because most capitalized words apart from first words in phrases are NEs. NE extraction experiments revealed that the NE recognition performance deteriorates when the capitalization and punctuation information are absent [2] (F. Kubala *et al* 1998).

The purpose of this paper is to propose a technique called Optimized Multiclass SVM Classifier for NE extraction (OMSC). In this OMSC we are going to generate the formation rules and attach this to the existing Multiclass SVM Classifier identify the exact content and it will be combined to the work flow scheduling model which is developed previously [4] (Jyothi Bellary *et al* 2014). The work is extended by using the OMSC technique.

This paper is structured as follows. A selection of related work is analyzed in Section II. In Section III, an optimized multi class SVM classifier is provided. Then, in Section IV, OMSC workflow scheduling Model is described. In Section V experimental evaluations is analyzed lastly, conclusion are mentioned in Section VI.

II. History and Related Work

In the last years, significant initiatives have been made and

DOI: 10.7763/IJET.2016.V8.932

amazing success has been acquired in the area of NE extraction. There is an increasing attention in NE extraction from multi-media information handling [5] (R. Basil *et al* 2005). Large-scale tasks such as the Global Autonomous Language Exploitation (GALE) [6] (D. Estival 2005) started out the way for the integration of NE extraction and machine translation.

The best resource of information with regards to NE extraction system explanations is the Automated Content Extraction (ACE) [7] (Doddington G *et al* 2004) and the proceedings of the Message Understanding Conference (MUC) [8], [9] (Chinchor N 1997) (Sundheim 1995), the 2002 and 2003 Conventions on Natural Terminology Learning (CoNLL) [10], [11] (Tjong Kim Sang E. 2003) (Tjong Kim Sang E. 2002) and the 1999 DARPA transmitted information work shop [12] (Przybocki M *et al* 1999). These proceedings contain the results of efficiency assessments as well as system explanations for each participating system in the assessment.

The assessments of MUC used domain specific data. For MUC techniques, since the domain is limited and capitalization details helpful for discovering NEs is available, many participating techniques of MUC were based on hand made rules. As the 1998 NIST Hub-4 assessment used transmitted news data, each individual in this assessment was needed to handle various domains in transmitted details. As the objective of this assessment was to identify organizations in content evaluation, each individual was also needed to deal with corruptions in feedback written text caused by conversation identification mistakes and details (such as capitalization) missing. CoNLL dedicated to language independent NE extraction for text input. At CoNLL-2003, 16 techniques presented a variety of machine learning techniques on written text data from newspaper.

The ACE system is designed to create technological innovation to automatically infer entities, the interaction among these entities, and the activities in which these entities take part. One of four difficulties which this system provided was the entity recognition and monitoring process. In this process, all mentions of an entity were to be gathered into the same entity and activities were calculated for text input.

Supervised learning-based NE extraction systems are usually classified according to whether they are stochastic or rule-based [13] (Bechet F *et al* 2004). As proven in the outcomes of evaluations, the Support Vector Devices (SVM)[14] (Asahara M and Matsumoto Y 2003) technique seems to outshine other methods. However, SVM needs extreme storage potential and calculations power. In concern of these specifications, , Maximum Entropy Models (ME) [15] (Bender O. *et al.* 2003), Conditional Random Fields (CRFs) [16] (McCallum A. and Li W. 2003) and Hidden Markov Model (HMM) [12] (Przybocki M. *et al*. 1999) seem to be genuine alternatives for NE extraction from large amount of verbal information. CRFs show better efficiency than HMM in many fields, which outcomes from the pleasure of the freedom presumptions needed by HMM. ME-based methods revealed the best efficiency at CoNLL-2003 and ME classifier is appropriate for the surroundings of a restricted storage because it needs smaller storage than CRFs. However, the best stochastic technique and the efficiency variations between classifiers differ across the evaluations.

III. Optimized Multi class SVM Classifier (OMSC)

In the Section II we have discussed about a problem of traditional handwritten rules in which it requires a large amount of human effort to write the rules. The proposed system OMSC deals with the formation Rules. This section defines the how to form the rules at the time of processing the workflow model.

A. Rule Formations

The syntactic structure of a phrase is in aspect indicated by punctuation marks, such as commas and full-stops. The system designed in this paper distinguishes all punctuation marks from successive words, and treats the punctuation marks as words. As some NEs involve more than one term, it is important in the execution of an NE extraction system to keep NE boundary information whether the word is along with its surrounding word.

The word characteristics itself speaks about word features; sometimes it gives good idea for NE extraction. For example, capitalization of the word first character, when it is not the first word of a phrase, reveals a greater probability of being a appropriate noun NE word. Table I shows possible form of word features. First, deterministic calculations are conducted to acquire word features. The Fst _Capital and All_ Capital are identified by whether the character in these terms is capitalized. The Not_in_Entry, Entry_in_L and Entry_in_R are used to notice the connections of non-NE words to NE words. These functions can be acquired by consulting to a table, which was designed when the word list was created. The last feature is Num termed as Numeric comes from the need of numerical and temporal functions.

TABLE I: The Arrangement of Channels

Type	Description
Fst_Capital	Words with capital letters at the first except the first word in the sentences
All_Capital	All capital letters in the word the word length is at least 2
Not_in_Entry	The word is not present in the NEs list
Entry_in_L	The word is present in the NEs Left entry
Entry_in_R	The word is present in the NEs Right entry
Num	The word is Numeric in the Numeric Dictionary

TABLE II: Word Featured Derived from the Name List

Type	Description
P_List	Words which is in the persons list
L_List	Words which is in the location list
O_List	Words which is present in the organization list

In this system designed here, word functions from name lists can be included as word functions at the rule formation. Table II reveals word functions depending on name lists.

Name details for individuals, places and organizations are used. When the rule-based system features this information, the program likes the more time factor, if more than one name-list's components are overlapped. If the same word seems to be on more than one name record, then a precedence concept is used. The location name record has the highest concern, the person name record has the next, and the organization name record has the smallest.

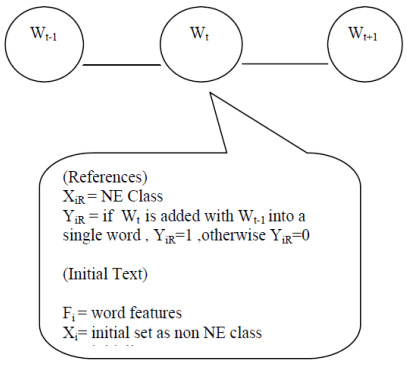

Fig. 2. Diagrammatic representation of rule formations [17] in OMSC.

Fig. 2 represents the rule formations, Wi denotes the word at position i in the rule generation table. In table 3 the rules are designed by comparing NE classes and their limitations in the present written text with those in the referrals. In order to execute this evaluation, the NE class of Wi is kept as XiR in the referrals. Moreover, YiR, which indicates whether Wi is along with Wi-1 into only one NE class is also saved in the referrals. If Wi is combined, YiR=1 and if not, YiR=0.

The appropriate rules are formed in accordance with the of Wi, Fi, Xi, and Yi to decrease the distance between Xi and Yi in the present written text and XiR and YiR in referrals. The initial value of Xi is initialized as non-NE, and that of Yi is set to 0. The detailed rule generation is given in Table III.

TABLE III: 22 RULE FORMATIONS FOR OMSC BASED ON TABLE I AND TABLE II

W: words F: features X: NE classes
W_0 [0 0], W_{-1} [0 0], W_1 [0 0], F_0 [0 0], W_0F_0 [0 0], W_0X_0 [0 0],
W_0W_{-1} [0 0], W_0W_1 [0 0], W_0W_{-1} [-1 0], W_0W_1 [0 1],
W_0F_{-1} [0 0], W_0F_1 [0 0], W_0F_{-1} [-1 0], W_0F_1[0 1],
W_0X_{-1} [0 0], W_0X_1 [0 0], W_0X_{-1} [-1 0], W_0X_1 [0 1],
$X_{-1}X_0$ [-1 0], X_0X_1 [0 1], $F_{-1}F_0$ [-1 0], F_0F1[0 1]

Table III shows about 22 rule formations [17] (Ji-Hwan Kim 2010) used in this system. Rule involves set of characters and a subscript. W, F, X signifies that templates are relevant to words, word functions and NE classes respectively. Y indicates whether the phrase is along with the past term into an individual NE class (if mixed, *Y*=1 and if not, *Y*=0). Subscripts display the comparative range from the current word; that is 0 indicates the present word, -1 indicates the previous word and 1 indicates the next word.

Each rule design has its own variety of program where the circumstances of the rule are met. For example, consider a generated concept 'if W0= RUPEES and F-1=NUM then change NE class to MONEY'. This belongs to the rule template of W0F-1[-1 0]. This implies that if the present word is 'RUPEES' and the function of the previous term is 'NUM' then modify the NE class of the previous and present terms into 'MONEY'. Then merge the previous term and the present term into only one NE class.

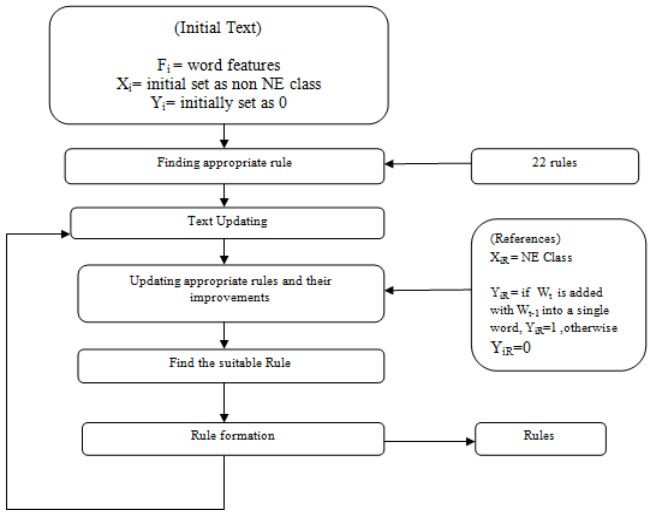

Fig. 3. Flow diagram for rule formations in OMSC.

In the Fig. 3 it has been shown that the enhancement for each possible rule is modified each time a rule is produced. From all the possible guidelines, the rule which causes the biggest enhancement is used to the current training data and the training data file is modified. If there are any changes in NE classes or NE boundaries which affect any of the other rules, then the developments from those other rules are also modified. In this program, the improvement is determined as the variety of terms which obtain their appropriate NE class or NE boundary after the concept is applied. These actions are recurring until no further changes can be created to the rules so as to decrease the variety of errors between the present NE classes and NE boundaries for the training information and the real NE classes and NE boundaries.

IV. OMSC WORKFLOW SCHEDULING MODEL

This section deals with OMSC Workflow scheduling in which we are extending the system frame work [4] (Jyothi Bellary *et al* 2014) . Here we are improving multi class SVM classifier by application rule formation function to it. This is shown in the Fig. 4.

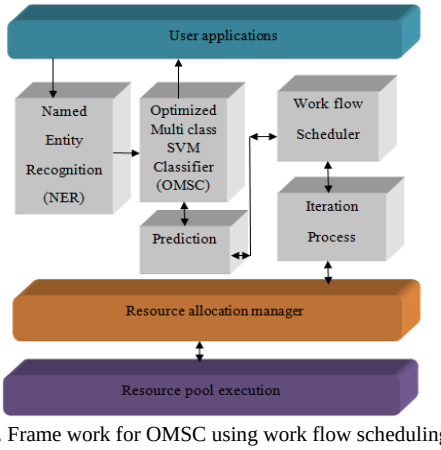

Fig. 4. Frame work for OMSC using work flow scheduling.

In the previous section we have discussed about the rule formations that will be enhanced to the multi class SVM classifier. This enhancement we call it as optimized multi class SVM classifier. As a result the time complexity of the OMSC is reduced and the response time will be improved.

V. Experimental Evaluation

To be able to evaluate the efficiency of the OMSC workflow scheduling, it was in comparison to that of multi class SVM classifier with workflow scheduling. As described in Section I, one of the reasons of this paper is to analyze the efficiency of the proposed NE recognition technique in information is to improve the precision and recall. Two essential metrics for evaluating the efficiency of an NE recognition are recall and precision. These terms are obtained from the Information Retrieval group. Recall R represents how much of the details that should have been resulted were actually correct. Precision P refers to the consistency of the information retrieved. The performance results of OMSC are shown in Table IV.

TABLE IV: Performance Results of OMSC with Different Datasets

English	Precision	Recall	F-score
1999IE	89.92%	90.17%	90.04 ±0.5
MUC-2	86.42%	86.21%	86.31 ±0.4
MUC-3	83.16%	82.67%	82.91 ±0.7
MUC-6	82.73%	64.97%	73.85 ±0.9
CoNill-03	78.15%	76.29%	77.20 ±1.2
Met2	74.98%	62.75%	68.86 ±0.3
Cora	87.56%	89.83%	88.84 ±1.0

A. Simulation Setup

To experimentally implement the work flow scheduling by using improved allocation algorithm we are going to use both FCFS scheduling policy and Round Robin scheduling policy. Here we are using the constant VMs count which is taken as 50,100 and 150. The experimental results for FCFS scheduling are shown in Table V.

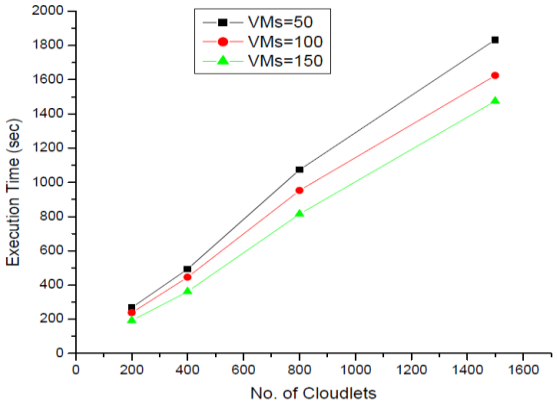

Fig. 5. Comparison of execution time with no. of cloudlets in FCFS.

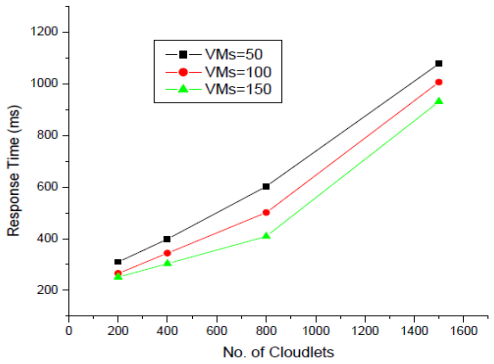

Fig. 6. Average response time in FCFS.

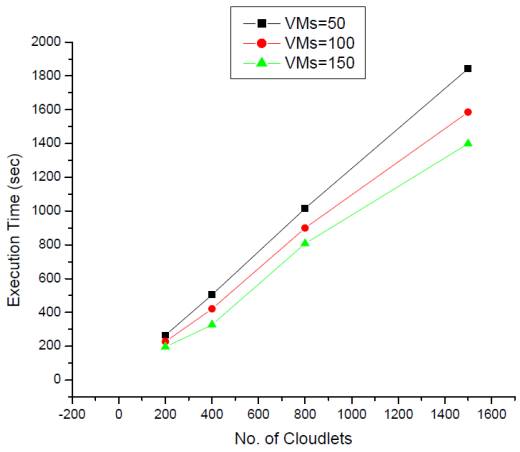

Fig. 7. Comparison of execution time with no. of cloudlets in round robin.

TABLE VI: OMSC with Round Robin Workflow Scheduling Using Improved Allocation

No. of cloudlets	**No. of VMs**	**Execution Time using Proposed Approach(sec)**	**Average response time (ms)**
200	50	264.37	274.57
	100	226.75	258.65
	150	194.60	243.91
400	50	504.71	342.52
	100	421.39	310.43
	150	326.84	299.02
800	50	1015.09	531.83
	100	899.67	498.45
	150	807.60	475.14
1500	50	1843.41	963.21
	100	1585.87	852.70
	150	1399.12	802.08

In Fig. 5 it shows the comparison graph of execution time and no. of cloudlets which is participated in the workflow

scheduling [18], [19] (Ashish Nagavaram *et al*. 2011) (Dhinesh Babu L. D. and Venkata Krishna P. 2013). It defines that execution time is increased automatically when the no of cloud lets increases. The average response time is also shown in Fig. 6. It can variable when we are using random VMs.

The experimental results of Round Robin scheduling policy are shown in Table VI. The Round Robin is similar to FCFS, but it has some time quantum for each and every task if the task is not finished within the time the process acts like FCFS and it swaps to the waiting queue.

In Fig. 7 it shows that the comparison of execution time in round robin with the no. of cloudlets. Here the round robin acts as efficient scheduling policy when compared to FCFS. The response time of round robin is smaller when compared to FCFS it is shown in Fig. 8.

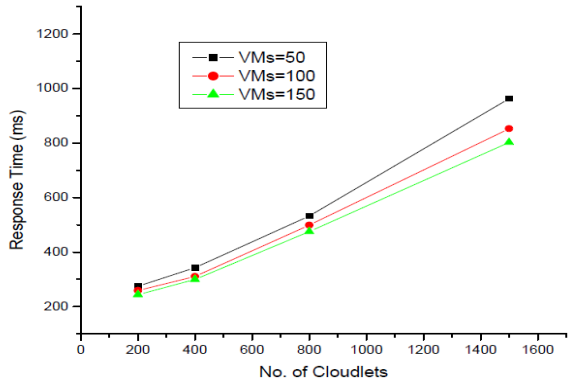

Fig. 8. Average response time in round robin.

VI. Conclusion

In this paper, an optimized multiclass SVM classifier for scheduling workflow in cloud, which forms automated rules to the named entity extraction, was suggested. Then its performance was in contrast to one of the successful Multi class SVM classifier. The suggested model has proven itself to be an affordable solution in NE extraction, while maintaining the key benefits of a rule-formation approach: its time complexity is improved and as far as scheduling is concerned the execution time and response time is improved.

References

[1] B. Sundheim, "Named entity task definition," in *Proc. 6th Message Understanding Conference*, Columbia, Maryland, 1995, pp. 317-332.
[2] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from speech," in *Proc. Broadcast News transcription and Understanding Workshop*, Lansdowne, Virginia, 1998, pp. 287-292.
[3] N. Chinchor, "MUC-7 named entity task definition (version 3.5)," in *Proc. 7th Message Understanding Conference*, Fairfax, Virginia, 1997.
[4] J. Bellary and E. K. Reddy, "Multiclass SVM classifier for named entity recognition with workflow scheduling system," *Journal of Machine Learning and Cybernetics and Awaiting Communication*, 2014.
[5] R. Basil, M. Cammisa, and E. Donati, "RitroveRAI: A web application for semantic indexing and hyperlinking of multimedia news," in *Proc. International Semantic Web Conference*, Sardinia, Italy, pp. 97-111, 2005.
[6] D. Estival, "The language translation interface: A perspective from the users," *Machine Transplantation*, vol.19, pp. 175-192, 2005.
[7] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The automatic content extraction (ACE) program -tasks, data, and evaluation," in *Proc. Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 837-840, 2004.
[8] N. Chinchor, "Overview of MUC-7/MET-2," in *Proc. 7th Message Understanding Conference*, Fairfax, Virginia, 1997.
[9] B. Sundheim, "Overview of result of the MUC-6 evaluation," in *Proc. 6th Message Understanding Conference*, Columbia, Maryland, pp.13-31, 1995.
[10] E. T. K. Sang, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," *in Proc. Conference on Natural Language Learning, Edmonton*, Canada, 2003, pp. 142-147.
[11] E. T. K. Sang, "Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition," in *Proc. Conference on Natural Language Learning*, Taipei, Taiwan, 2002, pp.155- 158.
[12] M. Przybocki, J. Fiscus, J. Garofolo, and D. Pallett, "1998 hub-4 information extraction evaluation," *DARPA Broadcast News Workshop*, Herndon, Virginia, pp. 13-18, 1999.
[13] F. Bechet, A. Gorin, J. Wright, and D. Tur, "Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How may I help you?," *Speech Communication*, vol. 42, pp.207-225, 2004.
[14] M. Asahara and Y. Matsumoto, "Japanese named entity extraction with redundant morphological analysis," in P*roc. Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, 2003, pp. 8-15.
[15] O. Bender, F. Och, and H. Ney, "Maximum entropy models for named entity recognition," in *Proc. Conference on Computational Natural Language Learning*, Edmonton, Candada, 2003, pp. 148-151.
[16] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, features induction and web-enhanced lexicons," in *Proc. Conference on Computational Natural Language Learning*, Edmonton, Canada, 2003, pp. 188-191.
[17] J. H. Kim, "Transformation-based named entity extraction from spoken content for personal memory aid," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, Nov 2010.
[18] A. Nagavaram *et al*., "A cloud-based dynamic workflow for mass spectrometry data analysis," *eScience*, 2011.
[19] L. D. D. Babu and P V. Krishna, "Applying operations management models for facility location problem in cloud computing environments," *International Journal of Services and Operations Management*, vol.15, no.1, pp. 1 – 27, 2013.

Jyothi Bellary is a pusing Ph.D from Jawaharlal Nehru Technological University Anantapur, Anantapuramu. She is a life member of ISTE, a member of IAENG, a member of IACSIT. She is working as an associate professor at the Department of CSE, Aitya College of Engineering, Madanapalle, Chittoor District, Andhra Pradesh, India. She had published 3 research papers in Nationl and Interntionl Journals and had presented papers in 4 National and International Conferences .

E Keshava Reddy is a presently professor of the Department of Mathematics, Jawaharlal Nehru Technological University Anantapur, Anantapuramu. He guided two Ph.D students and one M.Phil student. He adjudicates 9 Ph.D theses and 8 M.Phil theses. He published in many research papers national and international journals. He presented many papers in National and International Conferences. Delivered many guest lectures at various colleges and universities in the country. Life member of ISTE, IACSIT, Andhra Pradesh Society for Mathematical Sciences, Indian Mathematical Society, Calcutta Mathematical Society, Allahabad Mathematical Society, Indian Science Congress Assocition, Marathwada Mathematical Society and Indian Science Congress. Authored 8 books for undergraduate courses.