Hybrid content and collaborative filtering based recommendation system for e-learning platforms

Venkata Bhanu Prasad Tolety[1], Evani Venkateswara Prasad[2]
[1]Department of Computer Science and Engineering, JNT University, Kakinada, India
[2]Department of Computer Science and Engineering, NMR College of Engineering, Telangana, India

Article Info

Article history:

Received Aug 30, 2021
Revised Mar 13, 2022
Accepted Apr 18, 2022

Keywords:

Collaborative filtering
Content-based filtering
E-learning system
Recommendation systems
User profiling

ABSTRACT

Recommendation systems, although a well-studied topic, experience several shortcomings when applied on e-learning platforms. While collaborative filtering methods have enjoyed great success in making recommendations on large scale e-commerce and social networking and observation, users of e-learning platforms have continually evolving preferences, which render collaborative filtering methods weak. On the other end of the spectrum are content-based filtering approaches. Although such methods are more suited for e-learning platforms, the primary concern is that these methods find it hard to generalize across content sources and content types. In this work, we present a hybrid recommendation system that combines the desirable characteristics of collaborative filtering, as well as from content-based filtering, for the task of recommending course content/curriculum to users of an e-learning system. Our recommendation easily incorporates changing user profiles (as learners step through course content) and also generalize across content sources (courses taught by various departments) and types. We apply our system on a real dataset comprising 111 students organized into interdisciplinary groups. Our results showcase the clear benefits that our hybrid recommendation system enjoys, showing more than 30 percentage points of improvement over conventional filtering techniques.

Corresponding Author:

Venkata Bhanu Prasad Tolety
Department of Computer Science and Engineering, JNT University
Kakinada, Andhra Pradesh, India
Email: tvbphyd@gmail.com

1. INTRODUCTION

Recent technological advancements have paved a path for enhanced learning via the use of electronic learning (e-learning) platforms [1], [2]. E-learning has become a popular platform not only for student groups, but also for the teaching community [2]. E-learning, web-enhanced instruction, and other pedagogical technologies offer the listener an ability to enrich their learning experience by (i) navigating material at their own pace [2], (ii) providing scope for self-assessment and introspection [3], and (iii) reaching a broader demographic [4] (learners with familial obJNligations that may otherwise be unable to attend structured classroom courses). Furthermore, in the context of the COVID-19 pandemic, e-learning tools have become an indispensable tool to both educational and corporate training programs [5]-[8].

As investigated by Truong [1], each learner has an optimal learning style, pace, and methodology. This is primarily due to the "composite of characteristic cognitive, effective and physiological factors" [1]. While some learners grasp concepts by active experimentation (e.g., learning new words by forming sentences), others prefer passive strategies (e.g., learning by rote), and others still by reflective observations (e.g., watching a co-learner step through the learning process).

Designers of e-learning platforms must therefore account for these various learning strategies, and importantly, allow users to determine their own strategy while respecting constraints set forth by the teachers (e.g., topic A is a prerequisite of topic B – so the learner must finish A before proceeding to B). However, due to the rapid rise in demand for e-learning systems (fueled even more so by the COVID-19 pandemic [5]-[8]), it has become near-impossible for teachers to design tailor-made curricula for individual student. Furthermore, the eccentric teacher-student ratios in most institutions prohibit each student to avail the teacher's complete attention [9].

In this work, we leverage machine learning to overcome this challenge and allow each student to have their own aI-designed curriculum tailored to their needs. Unlike a human teacher, an aI-enabled e-learning system can scale to an enormous number of students, making the lives of both the student and the teacher easier. In our framework, the teacher serves as an overall system designer. The teacher specifies a base curriculum, expected difficulty of each topic, and a dependency structure. The dependency structure, represented as a directed acyclic graph (DAG), denotes which topics must absolutely be covered before others. It also enlists topics that may be learned in parallel without conflict. This key piece of information allows us to introspect what modules a learner might need to cover, based on self-assessment tests. In particular, we look at a popular class of machine learning (ML) techniques, called recommender systems, which share a somewhat similar objective. In the context of ML problems, recommender systems recommend data points to a consumer based on an inferred behavior pattern of that consumer (and the general behavior pattern of other consumers). Recommender systems are ubiquitous in the digital era – they exist everywhere, ranging from e-commerce websites to video sharing/streaming platforms, to social media [10].

While there exist a rich set of techniques for recommender systems [10], they have so far only been applied to e-commerce or content sharing platforms. Applying traditional recommender systems algorithms to e-learning poses unique challenges that need addressal. Users of e-learning systems have continually evolving preferences, and need to respect the curriculum pattern stipulated by the course designer. This renders the typical classes of recommender systems (collaborative filtering, content-based filtering) inapplicable to e-learning. In this paper, we present a solution to applying recommender systems for e-learning. We design a hybrid content and collaborative filtering algorithm, which brings the best of both collaborative and content-based filtering approaches and designs a recommender system that shows strong performance on e-learning platforms. We present our results on a dataset comprising e-learning traces from more than 110 students, and across two courses. Our approach significantly outperforms both content and collaborative filtering approaches.

2. RELATED WORK

The current teaching learning process is completely teacher centric. The line teacher maintains a pace that is acceptable to all the types of students. According to Shri Adi Shankaracharya in his stotra named Guru Astakam a hymn with 8 verses in praise of the Guru says "Skill or knowledge acquired without guru doesn't shine". However, e-learning systems have been shown to be more effective, allowing for self-pace managed learning personalizing the learning materials and the learning content management [1]-[4]. Where classical teacher-based learning systems do not scale to the modern-day demands of personalized-attention and content design, e-learning systems aided by ML solutions offer an alternate mechanism by which education systems can cope with the ever-increasing demand for better pedagogy. This helps both teachers and students: teachers are now able to focus on the actual course content design and assessment, while students can craft their personalized paced study plan.

However, for students to benefit from such a system, these e-learning platforms must possess the ability to dynamically update curricula to cater to changing student needs. To build such an effective learning system it is essential have a good recommender system. A recommender in system is a program that attempts to recommend the most suitable in terms to specific users by predicting a user's interest [11]. The effectiveness of a recommender system lies in its ability to assess a user's preferences and interacts by analyzing the behavior of the user and/or the behavior of the users to generate a personalized recommended [12]. There are various technologies that were stated in earlier literature viz. Collaborative filtering (CF) [13], content-based (CB) [14] and knowledge-based [15] techniques.

However, there are few limitations for these techniques viz, for collaborative filtering has sparseness, scalability, and cold-start problems [13], [16]. For content-based technique has overspecialized recommendation [16]. To overcome these problems few advanced techniques were developed viz, social network-based recommender system [8], fuzzy recommender systems [17], [18], group recommender systems [11], hidden bayesian model [19], link prediction [20], deep prediction model [21] and many more. However, due to the technological advancements none of the techniques cater to the needs of the current learner. Hence in this paper we propose a new adaptive e-learning technique that combines the features of collaborative filtering and the content-based filtering techniques.

The techniques presented in this paper become even more topical due to the dire constraints imposed by the COVID-19 pandemic [5]-[8]. The pandemic has exacerbated the already-skewed teacher-student ratio. As outlined in [6], [8], [15], many schools have experienced funding cuts and had to let a number of teachers go. In this context, the remaining teachers at schools and other educational institutions have had to handle an alarmingly large number of students. Our solution may be used in such a situation to enable the pedagogical methods to scale by using ML-based recommendations to cater to the needs of a large student base where information overload is also a huge problem [22]. Where typical recommender systems [11]-[15] fail, our method may readily be applied to e-learning platforms without any fine-tuning.

The rest of the paper in organized as; section 3 provides an overview of our proposed hybrid content and collaborative filtering-based approach. Section 4 details our approach and presents the concrete algorithm. Section 5 details the experiments we conducted to empirically validate the superiority of our approach over state-of-the-art content and collaborative filtering techniques.

3. OVERVIEW OF THE PROPOSED APPROACH

We present a hybrid recommendation system that combines the desirable characteristics of collaborative filtering, as well as from content-based filtering, for the task of recommending course content/curriculum to users of an e-learning system. While collaborative filtering methods have enjoyed great success in making recommendations on large scale e-commerce and social networking platforms, we argue that the fundamental idea on which collaborative filtering is designed is not well-suited for an e-learning setting. The basic assumption in collaborative filtering is that people who agree at one point in time on a specific item, will agree in the future on similar items. Moreover, it is assumed that people will continue to keep liking items similar in nature to what they liked in the past. We reason why these assumptions are not very valid in the context of e-learning. In an e-learning platform, both the user and the content 'evolve' over time [23]-[25]. Users tend to acquire knowledge on various topics and hence the distribution of their preferences tends to change over time (e.g. a user who has learnt about a topic X may now no longer be interested in that topic, but may develop interest in another topic Y). Further, e-learning systems have content that changes across time. Course offerings may vary over time (e.g. topics covered this year may be different from what were covered the previous year, in the same course). In such a volatile system, collaborative filtering is hence not the most appropriate technique.

Another disadvantage of using plain old collaborative filtering methods is 'cold start'. We often need a lot of data to bootstrap recommendations, which is not available for a new course/new user. Moreover, these methods are item-agnostic; they do not make any assumptions about the item being recommended. For instance, these methods run the risk of predicting both beginner-level content and advanced-level content to a beginner-level student, which goes against the pedagogical model of an e-learning system.

On the other end of the spectrum lie content-based filtering approaches. These approaches explicitly model the items being recommended and ensure that such items suit a profile of the users' preferences. These methods build a model of the user's preferences by examining a history of the user's interaction with the system, and an item-profile (a set of attributes or feature vectors computed for an item). Although such methods are more suited for e-learning platforms, the primary concern is that these methods find it hard to generalize across content sources and content types. Hence, if a user develops interest in a newer subject, the performance of content-based filtering approaches degrades.

4. PROPOSED APPROACH

We take a middle-ground and attempt to fuse the best of both worlds, namely collaborative filtering, and content-based filtering, and design a recommender system for predicting course material on an e-learning platform. We consider a document-based e-learning platform (i.e., our items of interest are documents). Note that other kinds of content, such as videos and images, also form natural extensions of the proposed system. Because these days most e-learning platforms have transcriptions of video/other visual content stored as text files. Upon modelling user-preferences and profiles, we recommend documents to users, while considering the content of each document. To model the content of a document, for each item in the database D, we compute a 'content vector', which encodes the semantics of the item and can be used for retrieval/recommendation. We use a modified version of the classic 'bag of words' model to extract content vectors that can be compared for similarity. Given each document, we obtain content vectors in the following manner.

a. Tokenization: we first tokenize the document into a set of words/symbols

b. Stemming/stop-word removal: frequently used stop-words, that are common across all documents and do not benefit feature extraction, are removed.
c. TF-IDF similarity computation: For each document, we compute a tf-idf vector, and use this vector to compute a similarity score based on cosine similarity.

We then cluster these documents using soft/fuzzy K-means++, a technique that also handles the computation of initial seeds and obtain K partitions of the dataset. Since the clustering is fuzzy, each document can belong to multiple clusters, but with different strengths. For instance, a document X can belong to cluster A with probability p_A, cluster B with probability p_B, and so on. We concatenate all these probabilities to form a content-similarity vector. Note that we use fuzzy clustering to allow for interdisciplinary subjects that cannot be hard-assigned to a single cluster.

The first step in our hybrid recommendation system is to run and gather predictions from both collaborative filtering and content-based filtering schemes. Naturally, the predictions of both the approaches vary. Say we get two sets of predictions A and B, from collaborative filtering and content-based filtering respectively.

a. We first compute S=A U B (set union operation over both the recommendations)
b. For each item recommended, we examine the content vector to check if the clusters in the content vector align with the users preferences (again, cosine similarity is used). We obtain a set P.
c. For each item in set S, we also check the alignment of the collaborative filtering scores with a general model of access patterns of other users. We only accept an item if the access likelihood is greater than a threshold (usually 0.7). This results in a set Q.
d. We return a union of the two sets (P U Q) as the output of our hybrid recommendation system.

The Algorithm 1 is implemented in python, with the help of the scipy and scikit-learn open-source libraries.

Algorithm 1. Hybrid content and collaborative filtering
Input: set of documents (D)
1. Tokenize the set of documents (extract a vocabulary V of tokens)
2. Stem (remove) stop words (i.e., frequently used words)
3. For each document, compute tf-idf scores
4. Perform soft/fuzzy K-means clustering to partition documents into clusters
5. Initialize the content vector for each document, using the cluster membership computed in step 4 above
6. A = set of recommendations from content-based recommender system
7. B = set of recommendations from collaborative recommender system
8. S = A U B (union of both recommendations)
9. P = empty set
10. For each item s in S
11. Compute cosine similarity with user preferences
12. If cosine similarity > threshold, add s to P
13. Q = empty set
14. For each item s in S
15. If access likelihood of s > threshold, add s to Q
16. Return P U Q

Output: recommended items

5. EXPERIMENTS

5.1. Datasets

We evaluate our performance on the dataset presented in the paper "Student activity and profile datasets from an online video-based collaborative learning experience" [26]. This dataset has been collected from an e-learning platform over a period of 3 months. In all, the e-learning platform was trialled over a sample size of 111 students from two different courses. Of the 111 students, 29 students hailed from the computer engineering course (CE) and 82 students were from the media and communication course (M&C).

The students were organized into 9 groups, each comprising students from both CE and M&C (on average 3-4 CE students and 8-9 M&C students per group). A separate group (solely comprising M&C students) gathered this dataset, to eliminate bias. To ensure expert supervision and data quality assurance, 4 teachers supervised the trial. The dataset has 2984 meaningful events which include access patterns. This allows us to validate the performance of recommender systems based on post-hoc hit-rate (higher hit-rate implies better prediction). This dataset was the best applicable dataset to our task because it is the largest e-learning access pattern dataset released to date under a permissible license. Furthermore, the dataset exhibits ever-evolving access pattern changes in students, and across multiple disciplines, making it an ideal choice.

5.2. Results

We show results based on the top-10 hit rate as a criterion. Since the dataset is not an active e-learning setup, we use 33% of the dataset (1 month) for profile creation and the remaining for testing. The results are summarized in Table 1, Figure 1, and Figure 2. We measure the top-10 hit rate (%) (higher indicates better performance). Notice that collaborative or content-based approaches when used in isolation have lower hit rates because they fail to capture *both* user and content preferences. However, our hybrid method captures both due to the additional filtering steps listed in Algorithm 1. Furthermore, we evaluate the top-10 reciprocal hit-rank (lower indicates better performance). Here again we see that our proposed hybrid approach performs significantly better, and this is visible across both groups (CE and M&C). For hit rate, a higher number indicates superior performance. For hit-rank, a lower number indicates better performance. Note that, 0.4 implies, 40%.

Figure 1 plots the top-10 hit rates in a graphical from for easy analysis. Figure 2 plots the top-10 reciprocal hit ranks. In both cases, we note that our proposed approach significantly outperforms existing content-based and collaborative filtering techniques.

Table 1 evaluates our proposed hybrid recommender system against state-of-the-art collaborative filtering and content filtering approaches. We measure the top-10 hit rate (%) (higher indicates better performance). Notice that collaborative or content-based approaches when used in isolation have lower hit rates because they fail to capture both user and content preferences. However, our hybrid method captures both due to the additional filtering steps listed in Algorithm 1. Furthermore, we evaluate the top-10 reciprocal hit-rank (lower indicates better performance). Here again we see that our proposed hybrid approach performs significantly better, and this is visible across both groups (CE and M&C).

Table 1. Top-10 hit-rate (%) vs top-10 reciprocal hit-rank

Group	Top-10 hit-rate (%)			Top-10 reciprocal hit-rank		
	collaborative	content	hybrid	collaborative	content	hybrid
CE	0.40	0.61	0.702	0.33	0.27	0.13
M&C	0.43	0.71	0.751	0.36	0.20	0.10

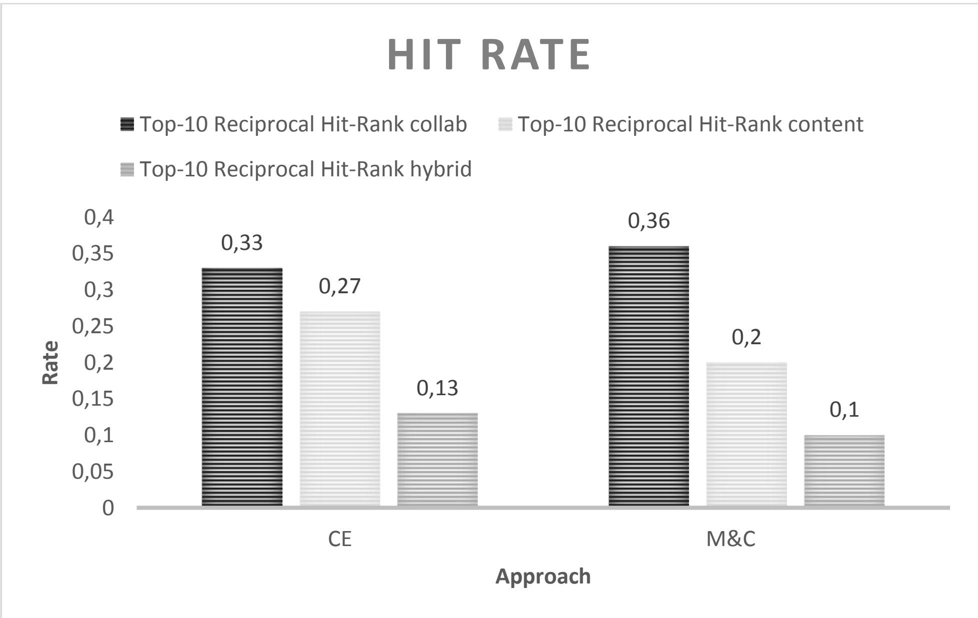

Figure 1. Graph comparing top-10 hit rates across various evaluated baselines. As can be seen, our hybrid recommender system outperforms both content and collaborative filtering approaches

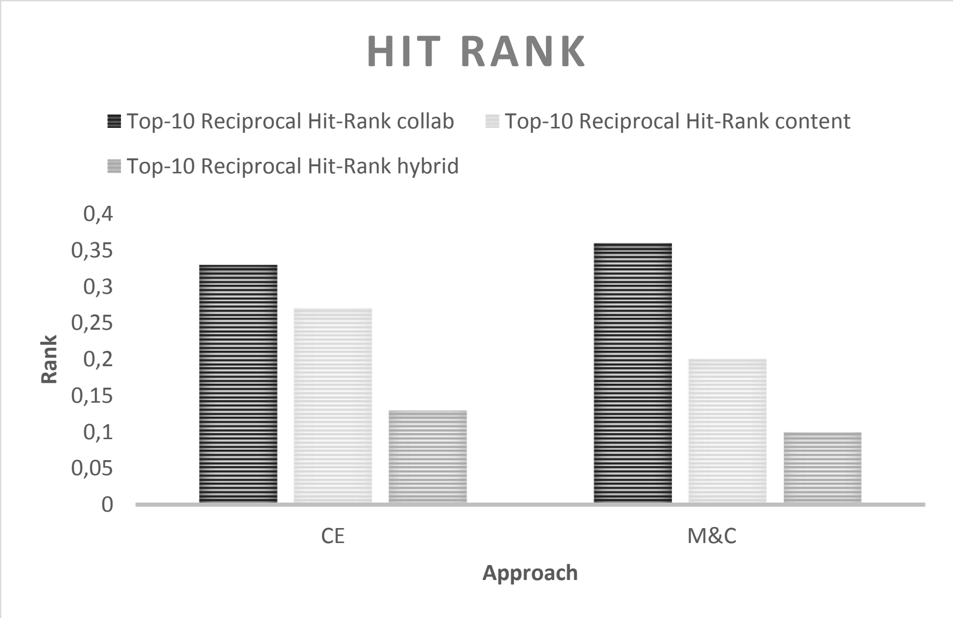

Figure 2. Graph comparing top-10 reciprocal hit ranks across various evaluated baselines. As can be seen, our hybrid recommender system outperforms both content and collaborative filtering approaches. Please note that a lower reciprocal hit rank indicates superior performance

6. CONCLUSION

In this work, we presented a novel, hybrid recommendation system tailored for e-learning platforms. We conlude that traditional content and collaborative filtering approaches are not well-suited to e-learning platforms (low hit rates). However, our hybrid approach achieves higher hit rates (and also lower reciprocal hit-ranks), denoting the benefits of using both user-preferences and content filtering and combining the best of both worlds for better performance. We obtain more than 30% improvements over state-of-the-art content and collaborative filtering approaches, enabling e-learning practitioners to design more accurate recommender systems for their platforms.

REFERENCES

[1] H. M. Truong, “Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities,” *Computers in Human Behavior*, vol. 55, pp. 1185-1193, 2016, doi: 10.1016/j.chb.2015.02.014.

[2] U. E. Franklin and A. A. Nahari, “The Impact of e-learning on academic performance: preliminary examination of King Khalid University,” *Development*, vol. 7, no. 1, pp. 83-96, 2018, doi: 10.6007/IJARPED/v7-i1/3903.

[3] H. Andrade and G. Brown, “Student self-assessment in the classroom,” Handbook of human and social conditions in assessment, pp. 319-334, 2016.

[4] A. A. Yahya, “E-learning contribution to the enhancement of higher education opportunities for women in Saudi Arabia (pilot study),” US-China Education Review A& B, USA, vol. 3, no. 9, pp. 637-648, 2013.

[5] H. Lee, “The Rise and Challenges of Postpandemic Online Education,” *IEEE Engineering Management Review*, vol. 49, no. 4, pp. 54-58, 2021, doi: 10.1109/EMR.2021.3105195.

[6] N. D. Vergara, and D. N. D. Perez, “Effects and consequences of covid-19 to higher education in Mexico,” *Journal of Higher Education Theory & Practice*, vol. 21, no. 3, pp. 1-11, 2021.

[7] D. R. Jessica, P. C. Marie and C. M. Jennifer, “How Implicit Assumptions about Engineering Impacted Teaching and Learning during COVID-19,” *Advances in Engineering Education*, vol. 8, no. 4, pp. 1-5, 2020.

[8] B. B. Lockee, “Online education in the post-COVID era,” Nature Electronics, vol. 4, pp. 5-6, 2021.

[9] S. Sangeeta, and J. Thakkar, “A comparative assessment of the performance of select higher education institutes in India,” *Quality Assurance in Education*, vol. 24, no. 2, pp. 278-302, 2016, doi: 10.1108/QAE-02-2015-0006.

[10] J. Bobadilla, F. Ortega, A. Hernando and A. Gutierrez, “Recommender systems survey,” *Knowledge-based systems*, vol. 46, pp. 109-132, 2013, doi: 10.1016/j.knosys.2013.03.012.

[11] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, “Recommender system application developments: A survey,” *Decision Support Systems*, vol. 74, pp. 12-32, 2015, doi: 10.1016/j.dss.2015.03.008.

[12] P. Resnick, and H. R. Varian, “Recommender systems,” *Communications of the ACM*, vol. 40, no. 3, pp. 56-58, 1997.

[13] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, “Collaborative filtering recommender systems,” The Adaptive Web, Springer, Berlin Heidelberg, pp. 291-324, 2007.

[14] M. Pazzani, and D. Billsus, “Content-based recommendation systems,” The Adaptive Web, Springer, Berlin Heidelberg, pp. 325-341, 2007.

[15] R. Burke, “Knowledge-based recommender systems,” *Encyclopedia of Library and Information Systems*, vol. 69, no. 32, pp. 175-186, 2000.
[16] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734-749, 2005, doi: 10.1109/TKDE.2005.99.
[17] Z. Zhang, H. Lin, K. Liu, D. Wu, G. Zhang, and J. Lu, “A hybrid fuzzy-based personalized recommender system for telecom products/services,” *Information Sciences*, vol. 235, pp. 117-129, 2013, doi: 10.1016/j.ins.2013.01.025.
[18] J. Lu, Q. Shambour, Y. Xu, Q. Lin, and G. Zhang, “A web-based personalized business partner recommendation system using fuzzy semantic techniques,” *Computational Intelligence*, vol. 29, no. 1, pp. 37-69, 2012, doi: 10.1111/j.1467-8640.2012.00427.x.
[19] Y. Xiao, X. Li, H. Wang, M. Xu, and Y. Liu,”3- HBP: A Three-Level Hidden Bayesian Link Prediction Model in Social Networks”, *IEEE Transaction on Computational Social Systems*, vol. 5, no. 2, June 2018.
[20] N. M. Ahmed, L. Chen, Y. Wang, B. Li, Y. Li and W. Liu, "DeepEye: Link prediction in dynamic networks based on non-negative matrix factorization," in *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 19-33, March 2018, doi: 10.26599/BDMA.2017.9020002.
[21] P. V. Kulkarni, S. Rai, and R. Kale, “Recommender System in eLearning: A Survey,” *Proceeding of International Conference on Computational Science and Applications*, Springer, Singapore, pp. 119-126, 2020.
[22] Ó. S. Martínez, C. P. G-Bustelo, R. González-Crespo, and E. T. Franco, “Using Recommendation System for E-learning Environments at degree level,” *Special Issue on Business Intelligence and Semantic Web, International Journal of Artificial Intelligence and Interactive Multimedia*, vol. 1, no. 2, pp. 67-70, 2009.
[23] J. He and W. Chu, “A social network-based recommender system (SNRS),” *Data Mining for Social Network Data*, Springer, Boston, MA, vol. 12, pp. 47-74, 2010, doi: 10.1007/978-1-4419-6287-4_4.
[24] J. Masthoff, “Group recommender systems: combining individual models,” Recommender Systems Handbook, pp. 677-702, 2010.
[25] D. Dagger, A. O'Connor, S. Lawless, E. Walsh and V. P. Wade, "Service-Oriented E-Learning Platforms: From Monolithic Systems to Flexible Services," in *IEEE Internet Computing*, vol. 11, no. 3, pp. 28-35, May-June 2007, doi: 10.1109/MIC.2007.70.
[26] E. Martin, M. Gertrundix, J. U. Fuantes and P. A. Haya, “Student activity and profile datasets from an online video- based collaborative learning experience,” *British Journal of Educational Technology*, vol. 46, no. 5, pp. 993-998, 2015, doi: 10.1111/bjet.12318.

BIOGRAPHIES OF AUTHORS

Venkata Bhanu Prasad Tolety is currently pursuing his PhD from JNTU kakinda. He has completed his masters from JNTU Hyderabad. He is currently serving in Accenture. He is member of IEEE and CSI. He is certified in Microsoft Azure. He can be contacted at email: tvbphyd@gmail.com.

Evani Venkateswara Prasad has thirty seven years of academic experience in technical education and is currently the Director, Lakireddy Bali Reddy College of Engineering, Mylavaram, Krishna Dt, Andhra Pradesh. He obtained Ph.D in Computer Science Engineering from the University of Roorkee in 1990, ME in Control Systems from the Madras University in 1978 and BE in Electronics and Communication from SV University, Tirupati in 1975. He has guided 11 students towards Ph.D degrees, co-authored six books and authored 102 research publications till date. He visited Malaysia on the invitation of Microsoft in 2004 and Blekinge Institute of Technology, Sweden, as part of MOU between the Universities, in 2010. He received State Award given to Meritorious Teachers from Government Andhra Pradesh, for the year 2008. He was adjudicated as the ‘best teacher’ by the students of CSE a number of times. He can be contacted at email: profevprasad@yahoo.com.