Recommendation Systems in the Big Data Era

Neelam Singh, Shivanshi Tripathi, Devesh Pratap Singh, Bhasker Pant, Vijay Kumar

Abstract—Rapid progression in technology and increasing use of social media platforms like Facebook, Instagram and Twitter has altered the way of articulating people's judgment, observation and sentiments about specific product, services, and more. This leads to the production and accumulation of massive amount of data. Recommendation systems are getting impetus when it comes to find insights from this data to make decisions that can be represented in various statistical and graphical forms. They have proven useful in predicting or recommending products ranging from food, movies, restaurants etc. This paper presents an overview about recommendation systems and a review of generation of recommendation methods based on categories like content-based, collaborative, and hybrid approaches. The paper will enlist the limitations which the present recommendation system faces and the possible improvements required in their capabilities to fit into a wider range of application areas.

Index Terms:Recommender system, Recommendation, Models, Data analytics.

I. INTRODUCTION

Recommendation has become the need of the hour and has drastically changed the interaction among user and web sites. Felds like education, finance and scientific research, etc are some application areas of the recommender systems. Rapid advancement in the field of information technology has increase the volume of digital information available and being generated.

Search Engines like Google and Bing are often used in this field of machine learning to search and filter the information based on movies, music, or articles by deploying big data analysis tools and techniques. By implementing recommender systems on websites various e-commerce based organizations and retail companies leveraging the advantage of data and advancing their sales. Hence, these systems intend to forecast users' interests and also, prescribe items that are much off their interests for them. User ratings provided after watching a movie or purchase of an article, from search engine queries or feedback becomes a key source of data needed for recommender systems. Applications like Netflix or YouTube suggests playlist or video recommendations by utilizing the data.

1990's marked the beginning of research on recommender systems and from that time onwards research in this area got expanded with introduction to new approaches with more better and accurate recommendation result. The task of information filtering in recommender systems can be performed by the following three algorithms, namely "collaborative filtering", "content based filtering" and "hybrid filtering"[1].

Eventually novel strategies progressed from these basic categories with improved recommendations with the inclusion of information from social networking platforms, Internet of Things, location information, and genetic algorithm based methods etc. Recommender has gained impetus in both industry as well as academia with more advance research undergoing in these sectors. Recommender systems offers a wide range of possibilities and challenges for research as well as implementations. A wide range of applications includes recommendations in web search, books, movies, music, restaurants, food, apparels, vehicles, targeted advertisements, medicines, news, probable consumers for companies and many more. Recommender systems are extensively employed by E-commerce platforms for improving user experience and thus increasing the usage. These system enables to convert a web surfer into a buyer by giving suggestions.

The paper reviews recommender algorithms, their classification based in methodology used and application area. The paper organization is given as section II deliberates several type of recommendation approaches with a comparison among the methods presented in section III. Section IV enumerates challenges and limitations addressed by recommender systems are listed and also how recommendation system works with big data is listed in section V and in section VI we conclude and future scope is discussed.

II. TYPES OF RECOMMENDATION SYSTEM

Recommender system deals in two types of information such as:

- *Characteristic information.* This category of information relates to values like keywords, categories, etc and users information about their favorites, profiles, etc.
- *User-item interactions.* It contains values like ratings, quantity of buying, likes, etc.

On the basis of information, recommender system are classified as [1] as shown in fig.1:
a. *content-based*, employing characteristic information,
b. *collaborative filtering*, which utilizes user-item interactions.

  Neelam Singh, Dept. of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.
  Shivanshi Tripathi, Dept. of Computer Application, Graphic Era deemed to be University, Dehradun, India.
  Devesh Pratap Singh, Dept. of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India
  Bhasker Pant, Dept. of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.
  Vijay Kumar, Dept. of Physics, Graphic Era Hill University, Dehradun, India.

c. *Hybrid systems* that merge both types of information in order to avoid conflicts that are incorporated when working with only one type of information [2].

A. Recommender system utilizing Collaborative filtering:

To create tailored recommendations on the Web, collaborative filtering (CF) method is frequently used. Popular websites example Amazon, Netflix, iTunes, etc. utilizes collaborative filtering technology. Algorithms for collaborative filtering are employed for automatic predictions about a user's interests by accumulating likings of various other users[3]. For example, a site such as Netflix may recommend that the users who watch movie A and B may watch movie C as well. This could be performed by comparing the past preferences of those who have watched the same movies. Thus, **Collaborative Filtering** doesn't require anything else other than users' past preferences on a set of items. The core assumption here is that the users past approval is likely to be considered for agreeing in future as well. Hence based on user preferences, **Collaborative Filtering** can be expressed into two categories as [3]:

The first category is the **explicit rating** in which user rates an item on a sliding scale, like 5 stars for Samsung Galaxy S10. This is the most straight away feedback from users that shows how much they liked an item.

Second category is **Implicit Rating** that recommends user preferences indirectly based on shopping history, page views, clicks etc[4].

The methods for performing collaborative filtering are nearest neighborhood and Matrix factorization.

➤ Benefits of Collaborative Filtering centered Recommender System[4]:
a. Collaborative filtering focused systems are more flexible as they can be applied to any domain, and when used in appropriate way can also give inter domain recommendations.
b. Collaborative filtering based systems perform best on a large user space.
c. Collaborative filtering engines can overcome "filter bubble" problem, as user finds and connects subspace in the item space.
➤ Limitations to Collaborative Filtering based Recommender System:
a. New rater problem: As a CF drawback, they suffer from the "new item" problem much more than CB engines (both the engines are constrained roughly equally by the "new user" difficulty, though their resolutions for initial profiling might vary).
b. Transparency: Collaborative filtering provides you the recommendations depending upon some unknown users those have the same taste like you, so they are not able to tell what features are responsible for the recommendation.

B. Content based recommender system:

In order to give suggestions to users this method of recommender system entails two significant categories of information. The primary type of information required is the supplementary information related to the item provided by a set of attributes assigned to them. Another information required is user profile to find out the user interactions with items with a specific set of attributes.

Attribute common among multiple objects are given higher preference as compared to others. These attribute weight together with the history plays an important role in generating user preference model. All objects in the database are compared with the model and values are allotted based on its likeness with the user profile and then scores recommendations are made.

After examining the User-Based and Item-Based Collaborative methods that utilizes the interactions of the users with the various types of items in order to build recommendations, we recognizes that these systems faces some problems like:

- *Cold-start problem for new users.*
- *New-item problem.*
- *Scarcity problem.*
- *Transparency issues*

Content-Based Recommender Systems are being generated by taking into consideration the content of each item for recommendation purpose, and trying to resolve the problems that are illustrated above. A content based system requires the data given by the user, either explicitly or implicitly. Thus depending upon the data provided, a user profile is produced, which can be further used to generate recommendations[5].

➤ Merits of content based filtering:
a. User independence: Unlike collaborative filtering, content based method only examines the items and user profile for suggestion.
b. Transparency: Collaborative Filtering provides you with the suggestions based on some anonymous users, who have the similar choices, but content based filtering suggest the items based on the available features.
c. No Cold start problem: In case of Content Based method, before grading by a given number of users new items can be advised [6].
➤ Challenges with content based filtering[6]:
a. Content Based method requires items Meta data, thus entail domain modeling and it asserts problem if their extension is applied to different domains.
b. CB engines are relatively impervious to user size.
c. CB engines are focused on providing items relevant and similar to user tastes as there can be counter actions to increase variety.
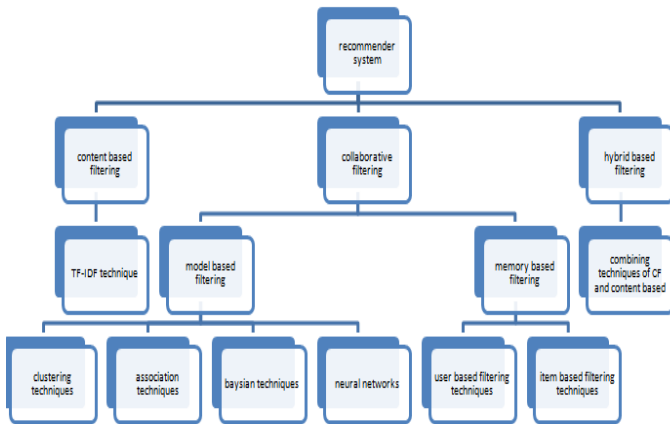
Fig. 1 Classification of Recommendation Sytem

C. Hybrid based recommender system:

These systems takes the advantage of both collaborative and content based filtering by evading their distinct restrictions. Joining collaborative and content based filtering methods to define a fusion system can be performed in many ways as classified below [7]:

1. Execution of both Collaborative and content based filtering is done separately and afterwards their results are combined as shown in fig.2.
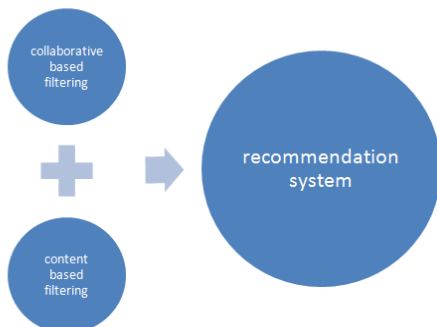

Fig. 2 Hybrid type 1

2. By applying collaborative filtering properties for content based recommendation as shown in fig.3.
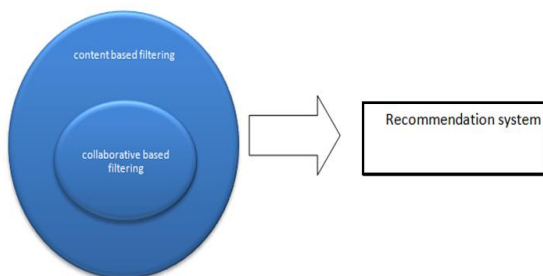

Fig. 3 Hybrid type 2

3. Fig. 4 discuss the use of content based filtering properties in collaborative based method.
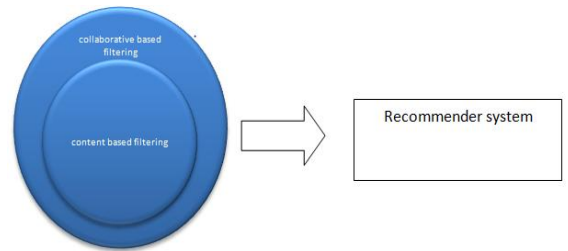

Fig. 4 Hybrid type 3

4. Attributes of both collaborative filtering and content based filtering are merged to develop Combination model as displayed in fig. 5.
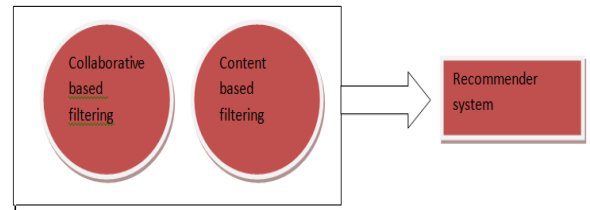

Fig. 5 Hybrid type 4

Hybrid Recommendation Systems classification:

Taxonomy for the hybrid recommendation systems presented by Burke categorizes Hybrid Recommendation Systems into seven classes as following [7] [8]:

1. Weighted hybrid: Hybrid class in which additive formula is taken to find cumulative scores of each class.
2. Switching hybrid: In this approach a section from the available recommendation components system is selected and applied.
3. Mixed hybrid: Suggestion of different recommender system are merged and multiple rated lists are transformed into single rated list.
4. Feature Combination hybrid: Two sections namely contributing and actual recommender are used. Actual recommender is dependent on altered data through contributing method where it throws features of one source on to the other sections source.
5. Feature Augmentation hybrid: Analogous to feature combination method with more flexibility as in this case contributor gives novel characteristic.
6. Cascade hybrid: In this class plays tie breaker method is used such that for every recommender priority is assigned and according to that lower priority recommenders plays a tie breakers part over higher priority.
7. Meta-level hybrid: Both contributing and actual recommenders exist but the initial section entirely substitutes the data for the latter section.

82

III. COMPARISON AMONG METHODS

For comparing the techniques available for building the recommender system we have plotted a bar graph in python using pandas based on the Fig. 6 and table 1 below:
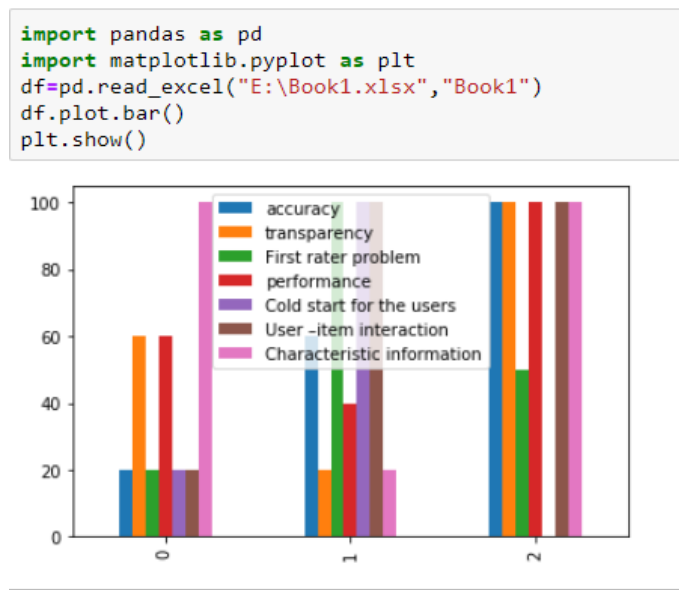
```python
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_excel("E:\Book1.xlsx","Book1")
df.plot.bar()
plt.show()
```


Fig. 6 Model comparison

Table 1 Comparison chart between among the models

Models\factors=>	accuracy	transparency	First rater problem	performance	Cold start for the users	User –item interaction	Characteristic information
Content based filtering technique	Less accurate	Yes	No	Better	No	No	Yes
Collaborative based filtering technique	More accurate	No	Yes	Good	Yes	Yes	No
Hybrid based filtering technique	More accurate	Yes	No	Best	No	Yes	Yes

IV. BIG DATA AND THE RECOMMENDER SYSTEM

Our day to day life has been majorly impacted by the existence of the Recommendation systems. Let's take an example for this as our online movie watching or say online book shopping experience is being redefined. The Recommendation system provides suggestions of items one person might be interested in as that person browses through the available items. Despite of the facts that business or consumer, Recommendation systems have been vastly valuable. Thus, the main concept behind the implementation of the Recommendation systems is big data. A job in a classic Recommendation system is totally based on the availability of sufficient data failing which it cannot perform. Thus, to suit this purpose big data supply excess amount of user's data in the form of past purchases, browsing patterns and history, and feedback to the Recommendation systems in order to generate significant and effective suggestions [9]. Without the use of big data, even the most advanced recommendations may not be effective.

Working of Big Data based Recommendation System

A Recommendation system with big data is divided into precise, rational phases as data acquisition, ratings, and filtering. And these phases are discussed as under [10]:

A. Data Acquisition phase

For example we consider that a user of YouTube or any other website is browsing a video or a movie for watching. Each time the user searches for a video or a movie, thus based on the search there comes n-number of results for that user. Depending upon the technology used, the event category could differ. Event is then entered into a NoSQL database. This entry is technical in nature but for a laymans it could be interpreted as something like "User a watched movie Z once or twice". This is how record are being captured and stored for upcoming suggestions.

For example, a customer who searched YouTube for Kabir Singh's "tujhe kitna chahne lage hai" was interestingly given several recommendations on other songs of this searched movie as shown in fig. 7:
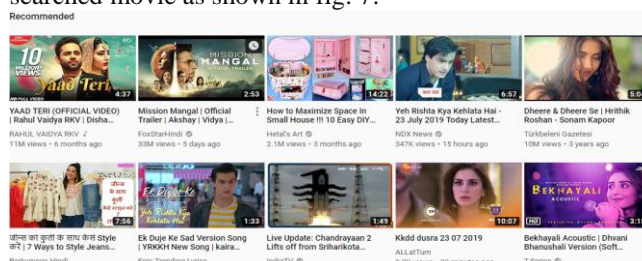

Fig. 7 Recommended list of songs

B. Ratings phase

Rating plays a significant role in telling about a user's perception for a product or an item. User's feelings for an item can be revealed to a level according to their actions such as likes, addition to watch list, watching or only searching. Based on user actions, Recommendation systems can provide implicit ratings. 5 are given as the highest rating which a user can give to an item. Let suppose, 4 rating can be assigned to watching, 3 to likes, searching can get 2 and so on. Recommender engines also consider the ratings and the feedback given by the users.

C. Filtering phase

Now this Filtering phase deals with the cleaning of the item details means based on ratings and other customer details provided with the help of big data techniques. Thus, based on the data provided, Recommender systems works on three classes of filtering techniques as: collaborative, user-based and hybrid methods. In collaborative filtering, commendations are provided on an assessment of user's preferences. Let suppose an example, if customer A likes products such as @, #, $, and % and customer B likes products such as @, #, $, % and &, then it is likely that customer A will be provided with the recommendation of the product '&' as and when seen between the similarities of both user A and B, there are lot more in common. Thus, depending upon similarities the recommendation has been done.

V. LIMITATIONS OF RECOMMENDATION SYSTEM

Recommendations generated by recommender system are based on data analysis which is susceptible to imperfection. The shortcomings of recommendation system are discussed in the subsequent section [11]-[12]:

A. Cold Start Problem

Cold start problem means to a condition in which the recommendation engine is not able to provide necessary and required predictions or suggestions because of the deficiency of primary ratings of a user or an article[13]. Also, this problem in recommender system can be found in two scenario i.e. when a new item or user is provided to the system. The problem of new item is found when there is an addition of an item into the system and that item does not have any primary rating before. Thus, the likelihood of suggesting an unrated product is minimum and thus they might become unobserved. In order to solve this problem the most likely way is to select a group of users who will give grating every new product. Major intention behind the new user problem is deficit of ratings for new user. Thus, in this condition, suggesting anything to a new user is impossible. When the user provide their initial ratings in the system they starts getting suggestions. But there will be a problem that the system will not be able to provision the required and necessary recommendation to the new user because of the redundancy or the lack of the data provided by new user [13].

B. Sparsity

Cold start is not the only problem with recommender system

Another problem faced is of Sparsity which is the situation in which the number of existing items to be rated are much more in comparison to already rated objects by the user. Thus, when a user article matrix is created based on the data available than there are very few entries that are to be marked which results the matrix to be sparse which results in reduced recommendations [14][15]. So, for this problem, one of the possible solutions is to provide suggestions to a user by checking user profiles and analyzing the similarities so that if two users share a common interest than one product may be recommended to other but this situation is not really necessary to conclude only based on the similarity of items that they have rated. This class of filtering method is also called as demographic filtering. To address the sparsity problem Singular Value Decomposition (SVD) to reduce the dimensionality of sparse rating matrix is also proposed [16].

C. Scalability

Due to the increase in no of users, products and ratings the issue of scalability is found in recommender systems. As and when the information about a product or an item increases than also the recommender systems are anticipated to answer quickly giving suggestions to the customers and this situation requires greater scalability[17]. But the execution of such systems becomes strenuous and exorbitant. Thus, the key challenge is to design an efficient and effective data model which can adapt itself to handle such massive datasets which keeps on growing. So for this problem, solution proposed by G. Takács et al. [17] to make use of an online learning algorithm which is capable of processing the updates associated to each user instantaneously and chronologically. Another method provided to address the scalability issue is to perform computation on multiple machines in parallel using distributed algorithm[18].

D. Overspecialization

Overspecialization in recommender system leads to suggesting items highly rated by the user thus can end up providing recommendation of items already purchased or experienced by the user. This can make user loose interest in the system because it does not work according to user preferences. Neighborhood collaborative filtering is proposed to handle overspecialization [19]. Randomness introduction using genetic algorithm or by removing similar items was also proposed.

E. Serendipity

A very crucial objective to be achieved by each recommender system is Serendipity, which focus on attaining user trust and loyalty. In contrast to user previous rating for items they will get significantly novel and relevant suggestions. It is challenging to apprehend the idea of serendipity completely due to its subjective nature and is rarely seen in real life scenarios. Solutions like re-ranking the accuracy results has been introduced to achieve serendipity [20].

VI. CONCLUSION

Recommender System research is on the path of improvement and standardization leading to better accuracy and relevance. New implementations and works came into existence which are promising but the approach needs to be more focused towards removing the underlying challenges faced by recommender systems. In this study we enlisted some of the significant implementations in this field with a comparative analysis between them using python. Also investigated challenges associated with recommender systems and its relationship with Big Data.

REFERENCES

1. Lu, Ji et al. "*Recommender system application developments: a survey.*" Decision Support Systems, vol. 74 pp. 12-32, 2015
2. Burke, Robin. "*Hybrid web recommender systems.*" The adaptive web. Springer, Berlin, Heidelberg. pp. 377-408, 2007
3. Babu, Maddali Surendra Prasad, and Boddu Raja Sarath Kumar. "*An Implementation of the User-based Collaborative Filtering Algorithm.*" IJCSIT) International Journal of Computer Science and Information Technologies vol. 2 no.3, pp. 1283-1286, 2011.
4. Chen, Rui, et al. "*A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks.*" *IEEE Access,* vol. 6, pp. 64301-64320, 2018.
5. Sorde, Roshni K., and Sachin N. Deshmukh. "Comparative study on approaches of recommendation system." *International Journal of Computer Applications,* vol. 118 no. 2, 2015.
6. P. Nagarnaik and A. Thomas, "Survey on recommendation system methods," *2015 2nd International Conference on Electronics and Communication Systems (ICECS),* Coimbatore, pp. 1603-1608, 2015. doi: 10.1109/ECS.2015. 7124857.

7. Choi, Keunho, et al. "A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis." *Electronic Commerce Research and Applications* vol.11, no.4, pp. 309-317, 2012.

8. Mu, Ruihui. "A survey of recommender systems based on deep learning." *IEEE Access,* vol. 6, pp. 69009-69022. 2018.

9. Zhang, Y., 2016. GroRec: a group-centric intelligent recommender system integrating social, mobile and big data technologies. *IEEE Transactions on Services Computing*, 9(5), pp.786-795.

10. Wang, Y., Wang, M. and Xu, W., 2018. A sentiment-enhanced hybrid recommender system for movie recommendation: a big data analytics framework. *Wireless Communications and Mobile Computing*, *2018*.

11. Bai, Xiaomei, et al. "Scientific paper recommendation: A survey." *IEEE Access* 7 (2019): 9324-9339.

12. Misale, Mohini, and Pankaj Vanwari. "A survey on recommendation system for technical paper reviewer assignment." *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*. Vol. 2. IEEE, 2017.

13. Chakraborty, Jayeeta, and Vijay Verma. "A survey of diversification techniques in Recommendation Systems." *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. IEEE, 2016.

14. Akhil, P. V., and Shelbi Joseph. "A Survey Of Recommender System Types And Its Classification." *International Journal of Advanced Research in Computer Science* 8.9 (2017).

15. Nagarnaik, Paritosh & Thomas, A. (2015). Survey on recommendation system methods. 2nd International Conference on Electronics and Communication Systems, ICECS 2015. 1603-1608. 10.1109/ECS.2015.7124857.

16. Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., 2002, December. Incremental singular value decomposition algorithms for highly scalable recommender systems. In *Fifth international conference on computer and information science* (Vol. 1, No. 012002, pp. 27-8).

17. Takács, G., Pilászy, I., Németh, B. and Tikk, D., 2009. Scalable collaborative filtering approaches for large recommender systems. *The Journal of Machine Learning Research*, *10*, pp.623-656.

18. Mohamed, Marwa Hussien, Mohamed Helmy Khafagy, and Mohamed Hasan Ibrahim. "Recommender systems challenges and solutions survey." In 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), pp. 149-155. IEEE, 2019

19. Sarwar, B.M., Karypis, G., Konstan, J. and Riedl, J., 2002, December. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the fifth international conference on computer and information technology* (Vol. 1, pp. 291-324). Almazro, Dhoha, et al. "A survey paper on recommender systems." arXiv preprint arXiv:1006.5278 (2010).

20. De Gemmis, M., Lops, P., Semeraro, G. and Musto, C., 2015. An investigation on the serendipity problem in recommender systems. *Information Processing & Management*, *51*(5), pp.695-717.

AUTHORS PROFILE

Neelam Singh, Assistant Professor, Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India. She has received M.Tech (CSE) degree in Computer Science and Engineering, GEU, Dehradun, India in 2014. Her research interests include Big Data, Machine Learning, Data Mining and Artificial Intelligence.

Shivanshi Tripathi, Student (M.Sc IT), Department of Computer Application, Graphic Era Deemed to be University, Dehradun, India. A meritorious student with research interest in Machine Learning, Big Data.

Dr. Devesh Pratap Singh, Professor and Head of Computer Science and Engineering department at Graphic Era Deemed to be University Dehradun India. He has received M. Tech degree in Computer Science and Engineering from Uttarakhand Technical University Dehradun India in 2009. He has also received Ph.D. in 2015. His research interests include Information Security, Wireless Sensor Networks, Internet of Things and Soft Computing. He has published more than fifty research papers in his area of expertise. He is the member of ACM.

Dr. Bhasker Pant, Dean Research & Development and Associate Professor in Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal.Has more than 15 years of experience in Research and Academics. He has till now guided as Supervisor 3 Ph.D. candidates (Awarded).and 5 candidates are in advance state of work. He has also guided 28 MTech. Students for dissertation. He has also supervised 2 foreign students for internship. He has more than 70 research publication in National and international Journals. He has also chaired a session in Robust Classification & Predictive Modelling for classification held at Huangshi, China.