IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Systematic Study and Application of Machine Learning Algorithms in Recommender System Design

Shweta Sharma[1], D. P. Sharma[2]

[1]ICFAI University, Hyderabad, India
[2]ICFAI University, Hyderabad, India

Available online at: www.ijcseonline.org

Accepted: 03/Jun/2018, Published: 30/Jun/2018

Abstract- To perform product or services' recommendations, the Recommender System (RS) is used by most of the social media, such as Twitter, LinkedIn, Netflix, etc. and potential e-marketers, to name, Amazon, Flipkart, Alibaba, eBay, Myntra, etc. including the famous search engine Google. All of these systems uses Machine Learning (ML) algorithms claimed from the field of Artificial Intelligence (AI). However, choosing an appropriate ML algorithm to fulfil this task of Recommender System (RS) is a critical issue, if not impossible, since a considerably large number of algorithms find place in the literature. Practitioners and researchers developing Recommender System leaves a very little information about their current approaches in algorithm usage, thus it is sufficient to create further confusion to perform the task of selecting appropriate algorithm.

The current paper presents a systematic insight in to the subject analysing the usage of machine learning algorithms for Recommender System (RS), and thereby identifies the research opportunities to bring further improvement into the system used. The study carried exposes that the Bayesian network and Decision Tree algorithms are widely adopted and used in the Recommender System (RS) due to their relative simplicity along with required performance. The software system requirements and the design phases adopted for the same also appears to have ample of further research opportunities. This paper presents a systematic analysis of the topic under consideration with recommendations of performance measures and evaluation procedures as per its suitability for designing an effective RS.

***Keywords*:** Machine Learning, Recommender System, Deep Learning, Artificial Intelligence, systematic study.

I. Introduction

Recommender System (RS) is usually employed to help user find new items or services of their interest, such as dress materials, books, electronic gadgets, music CDs, etc. It is done based on the information obtained about the user, or the recommended item [1]. Such systems plays a critical role in decision-making by helping user to minimize risk [2]. Now the Recommender System (RS) is used by most of the social media, such as Twitter, LinkedIn, Netflix, etc. and potential e-marketers, such as Amazon, Flipkart, Alibaba, eBay, Myntra, etc. including the famous search engine Google. The RS has its origin in the mid of 1990s with introduction to Tapestry [3], supposed to be the first RS. Along with evolution of this field, the researchers begin studying the utility of machine learning algorithms belong the area of Artificial Intelligence (AI). The advent of machine learning (ML) is being marked as late 1950s [4], that is, along with emergence of the field of Artificial Intelligence (AI).

Presently we find a plethora of ML algorithms, such as clustering [5], k-nearest neighbour [6], Bayes network [7], etc., to name a few, which are used in appliances ranging from the vacuum cleaning robots [8], assisting disabled people [9] to pattern recognition [10], or even a self-driving vehicle [11]. The potential application of ML is vast enough but the field looks as one of the most promising one.

As it is mentioned above, we use ML algorithms in RS to provide the users with better recommendations regarding goods and services both. However, the clear classification scheme is missing in ML regarding its algorithms. It is due to a number of approaches and variations in algorithms as present in a number of literatures [12]. As a consequence, it becomes literally difficult and a bit confusing too to select an appropriate ML which may fit into one's need while developing RS. Additionally, the researcher may also find it quiet challenging to track the use of ML algorithms in the given Recommender system (RS).

Considering the difficulty in choosing the appropriate algorithm for RS, a detail work is done to investigate how the actual ML algorithm is implemented to get the RS working. It is also taken into consideration to find the impacts of software engineering (SE) in development of the given RS. It is expected that the researchers as well as practitioners will be able to draw a lots of information out of the work done to implement it while developing an appropriate RS.

II. Theoretical Backdrop

This section comprises of a brief overview of the ML and RS. This section focuses on a short historical description along with certain definitions and classifications. The main objective of this section is to get the readers aware of historical perspectives of RS and ML for clarity of understanding as we proceed further to the higher concepts.

2.1 The Recommender System (RS)

It is quiet known that the RS uses ML algorithms to suggest or recommend the users with appropriately optimized product or services. For instance, an online bookstore may utilize ML (Machine Learning) algorithm for the purpose to classify the books by genre, and then after make a recommendation of appropriate books to the customers buying books of specific type.

Process of Recommender Systems (RSs) is also known as "Collaborative Filtering". The Recommender Systems are classified into three categories, e.g., i) Collaborative, ii) Content-based, and iii) Hybrid filtering [1]. Recommender System (RS), while processing information for specific recommendation using collaborative approach, considers *user's data*. For example, RS accesses all the user's data, such as – the name, age, city, country, the books purchased in past, etc., while accessing user's profile in an online book store. Getting this information, the RS can now identify all users that share the same book preference, and thus it suggests the books bought by the similar type of buyers.

Recommender (RS) base its recommendations on the *item data* it can access, with a content-based filtering approach. For example, let us consider a customer looking for a new gold set using an online store. When he browses a specific set, the RS collects information about that specific gold set and the searchers in the database allocated for gold sets having similar attributes, e.g., set type, price, number of sticks, etc. Thus, outcome of this search is then returned to the buyer as a recommendation.

The third classification describes recommender system (RS) that uses combination of the previous two classifications turning it into a hybrid filtering approach that is based on the *user's* as well as the *item's* data. For instance, considering a social network, the RS may recommend such profiles that are similar to the user (collaborative filtering), by comparing their choices or interests. At the same time, on the second step, the RS may consider the recommended profiles as "items", and thus access their data to search for the fresh but similar profiles (content-based filtering). At the end of the story, both sets of selected profiles are returned as recommendations.

This is mandatory that the Recommender System (RS) must collect information about the user in order to develop recommendations while using collaborative or a hybrid filtering approach. However, such activities can be performed explicitly or implicitly. The explicit data collection takes place while user is aware of providing his information to such systems. For instance, when one register for a new online service, he / she usually full in a form asking his / her name, age, sex, email ID, etc. The other forms of explicit (user's) data collection takes place when user express his / her preferences by rating items using a numerical value or preference such as Facebook "like". Implicit data collection accesses user's information indirectly. Example of such data collection is, while visiting an online store the server at the store exchange messages with user's computer, and based on that, the store's RS comes to know the browser being used by the user along with country of origin of the user. More advance application may even monitor users clicks and keystroke logs.

Apart from the common recommendation process in which users are presented with items or services that might be of interest to the user, the recommendations may even be provided using other process too.

In the trust-based recommendation process [13], the trust relationship prevailing between users is taken into consideration. A link in a given social network to a friend or a following connection is treated as the trust relationship. The recommendation based on trusted friends is treated as more worthy than those that do not have the trust links between them. Similarly, the context-aware recommendation [14] is based on the context of the user under consideration. Context is a set of information regardi9ng the current state of user, such as – the time at the user's location (e.g., morning, afternoon, evening), or their activities (e.g., sleeping, idle, running, etc.). The magnitude of context information to be processed is high enough, turning context-aware recommendation a challenging field of research. The risk-aware recommendation [2] is a subset of context-aware recommendations, and it takes into consideration a context in which the critical information exists, e.g., user's vital signs, etc. It is, of course, risk-aware since a wrong decision may become too critical to cause even user's potential damage. Instances may be taken of – recommendation of pills to be taken or the stock the user must buy or sell.

2.2 Machine Learning (ML) Algorithms

The Machine Learning (ML) employs electronic machines (computers) to simulate human learning and allows the machine to identify and acquire knowledge from the real-world, as well as improving its performance on certain tasks based on this new knowledge. More formally, we can define machine learning [21] as "a computer program that learn from the experience (E) with regard to certain class of tasks (T) and performance (P). Its performance (P) improves with experience (E) gained while performing the task (T)".

Learning is a process of knowledge acquisition. Human learn in a natural way from his experience by dint of their ability to reason out the experience being encountered. In contrast, the machine do not learn by reasoning rather it learns by the algorithm it happen to follow. Now, in literature, a vast number of machine learning algorithms are available to perform various purposes. These can be classified based on the approach it uses for the learning process. However, there are four main classifications of MLs, that is, i) Supervised learning algorithm, ii) Unsupervised learning algorithm, iii) Semi-supervised learning algorithm, and iv) Reinforced learning algorithm.

The supervised learning algorithm is used where algorithm is provided with training data along with its correct answers. The task of machine learning algorithm is to learn on the basis of training data and then apply the knowledge gained in real data. For instance, consider a machine learning algorithm used for book's classification in an online bookstore. A training set (training data plus answers) can be a table relating to information about each book in the correctly classified manner. Information about each book may contain book-title, author's name, publisher, etc. The algorithm learn it by the training set provided. Now, whenever a new book is arrived at the bookstore, the algorithm will classify the book on the basis of the knowledge it has acquired about classifying the book.

In un-supervised learning, the machine learning algorithm do not have a training set. It is presented with certain real world data and it is expected that the machine will learn from this data on its own. Unsupervised learning algorithms are mainly focused on finding the hidden patterns from the data presented to it. For instance, supposing the ML algorithm has access to user's profile information available in a social network. By using the unsupervised learning approach, the algorithm can classify user into various categories, e.g., social activist, artist, writer, etc., allowing the social network company to target advertisements in a more focused way to specific group of users.

The semi-supervised category of ML algorithm learn when t raining data set is provided with certain missing information, and still expected to learn from it. For instance, when ML is provided with books review report. Not all user reviewed every books, and thus there is certain missing information. Semi-supervised algorithm is able to learn and draw appropriate conclusions even if the information is incomplete, that is, even when data is missing.

Yet another class of ML algorithm has a reinforced learning approach. The reinforced learning takes place when algorithm learns based on external feedback is supplied by certain entity or the environment. This approach is very much analogous to teaching a child to read the language. When the child perform well (that is, correctly), he is rewarded by patting on his back (a positive feedback). No patting (negative feedback) is there if he performs wrongly. In machine learning too, the positive feedback is learnt and repeated, whereas the negative feedback is always avoided.

With the advent of memory size and increased processor's speed, the ML has become a popular process. As a consequence, profound work is done in this direction and a large number of learning algorithms have come out that uses various processes such as mathematical and statistical analysis, to learn, as well as draw conclusions, etc. A good number of variations of the measure algorithms are also published in scientific journals that propose variations or combinations of ML algorithms. However, they all share specific classes as defined above. However, proper standardization is still due.

III. Algorithm Used in RS

The foremost task is to select the adequate RS algorithm out of the range of available systems/algorithms. It must be selected to fulfil the task faithfully without flaw. This selection has a significant impact on the rational of the Recommender System, on the user's supplied data (that is, the recommendation items) to be required, and on very performance of the RS itself. This selection is a real challenge since there is a large number of algorithmic variations and combinations available in literature to be selected from.

A vast number of RS algorithms that appears to be continuously growing with sufficient alterations into it, poses a real and continuous challenge for the software engineering. Along with the growing market, the hurdles, challenges and trends keeps on changing and growing which need to be properly taken care of.

This paper considers the literature available along with selected case studies in the field of development of RS. The RSs that are implemented using ML algorithms and the real or simulated data are taken into consideration for a reliable study. Authors restricts the scope of software engineering areas to the five stages of the waterfall model for Software lifecycle, that is, requirements, design, implementation, verification, and maintenance [15]. This paper identifies research opportunities that need to be investigated for further development and fine tuning the RS algorithms.

Finally, the problem or research opportunity is classified into one of the five software lifecycle stages. Through this classification, it is expected to have a perspective on which areas of SE research would be valuable in supporting the implementation of RSs. There are several synonyms for RSs. Based on [16] this systematic review considers RS terms that replace "recommender" by "recommendation", and terms that replace "system" by "platform", or "engine". This systematic review does not consider any "machine learning" synonyms.

A survey of literatures suggests the following algorithms are normally used in RSs (given according to frequency of use, stated in percentage):

Table 1: Frequency of use of algorithms

Sl. No.	Algorithm	Percentage of use
1	Bayesian	70%
2	Decision Tree	50%
3	Matrix factorization-based	40%
4	Neighbour-based	40%
5	Neural Network	40%
6	Rule Learning	40%
7	Ensemble	30%
8	Gradient descent-based	30%

9	Kernel methods	30%
10	Clustering	20%
11	Associative classification	10%
12	Bandit	10%
13	Lazy learning	10%
14	Regularization methods	10%
15	Topic Independent Scoring Algorithm	2%

However, a variation in the selected algorithm can be brought according to need and stated objectives in the research undertaken by an individual.

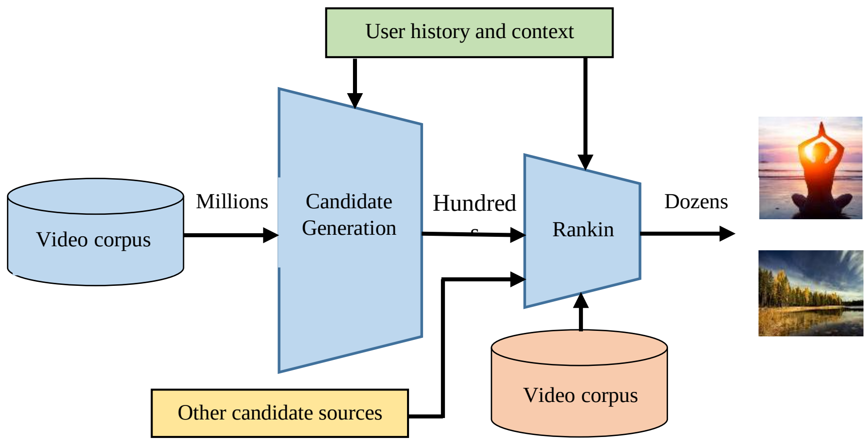

Figure 1: Model Recommender System used in YouTube

According to the study "*Deep Neural Networks for YouTube Recommendations*", the YouTube recommendation system algorithm comprises of two neural networks: one for candidate generation and one for ranking [17]. We have a quick summary here.

Taking events from a user's history as input, the candidate generation network significantly decreases the amount of videos and makes a group of the most relevant ones from a large corpus. The generated candidates are the most relevant to the user, whose grades we are predicting. The goal of this network is only to provide a broad personalization via collaborative filtering.

Undoubtedly, it's a very challenging task to make recommendations for such a service because of the big scale, dynamic corpus, and a variety of unobservable external factors.

At this step, we have a smaller amount of candidates that are similar to the user. Our goal now is to analyze all of them more carefully so that we can make the best decision. This task is accomplished by the ranking network, which can assign a score to each video according to a desired objective function that uses data describing the video and information about users' behaviour. Videos with the highest scores are presented to the user, ranked by their score.

While designing your own RS, you need to have the followings:

i. Clear objective (s) as to what do you want your RS to do?
ii. What design of RS do you propose, and why?
iii. Which algorithm do you choose to employ and what is its suitability?
iv. What is the search overhead the server/machine is expected to bear?

IV. Outcomes of the Current Study

Outcomes of the current study is presented in terms of performance measures, evaluation of the recommender system, and anticipatory design.

4.1 Performance Measures

To assess the effectiveness of recommender system, its evaluation is important. The commonly used metrics are MSE (Mean Squared Error) and RMSE (Root Mean Squared Error). The information retrieval metrics, such as precision and recall or DCG are used to assess the quality of a recommender system. In information retrieval system, the *precision* (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while *recall* (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.

DCG (Discounted Cumulative Gain) is a measure of ranking quality. In information retrieval, it is often used to measure effectiveness of web search engine algorithms or related applications like RSs. Currently, diversity, novelty, and coverage are also taken as important ingredient in evaluating an information retrieval system. It is always a prudent decision to take into consideration an evaluation system that has a proven track record particularly for Recommender Systems.

4.2 Evaluation of Recommender Systems

Evaluation of a Recommender System is carried to ascertain its effectiveness. To measure its effectiveness, we have three evaluation types available, that is, i) User Studies, ii) Online Evaluations, and iii) Offline Evaluations. The effectiveness is, however, measured depending upon “how well an RS can predict users’ ratings in the given dataset”.

User Studies: A user study is conducted by recruiting a set of test subject, and asking them to perform several tasks requiring an interaction with the RS (Recommendation System). While the subjects perform the tasks, we observe and record their behaviour, collecting any number of quantitative measurements, such as what portion of the task was completed, the accuracy of the task results, or the time taken to perform the task [18].

Online Evaluation: The experiment which provide the strongest evidence such as the true value of the system fall in the category of online evaluation, where the system is used by real users that perform real evaluation tasks. It is the most trustworthy process to compare a few systems online, obtaining a ranking of alternatives, rather than absolute numbers that are more difficult to interpret.

Offline Evaluation: Offline experiments are attractive because they require no interaction with real users, and thus allow us to compare a wide range of candidate algorithms at a low cost. The downside of offline experiments is that they can answer a very narrow set of questions, typically questions about the prediction power of an algorithm. In particular, we must assume that users’ behaviour when interacting with a system including the recommender system chosen will be modelled well by the users’ behaviour prior to that system’s deployment. Thus we cannot directly measure the recommender’s influence on user behaviour in this setting.

The selection of evaluation process, however, depend upon the actual requirements, and also the result of a pure evaluation type or hybridizing the evaluation results also depends on the actual objective (s) of the researcher on a particular project.

4.3 Anticipatory Design of RS

As per assumption, the anticipatory design differs largely from the conventional design. The goal, in anticipatory design, is to simplify the process along with minimizing the difficulty while making decision on users’ behalf. The instances can be sighted of Amazon and Netflix production recommendation which uses various levels of anticipatory design for recommending products based on previous behaviours recorded [19].

In its finest form, the anticipatory design goes way beyond simple personalization. For instance, Netflix recommend you movies to watch based upon your choice preferences and history, is called personalization. With anticipatory design, the UI (user interface) actually changes in the moment as you are interacting with specific app [20].

Anticipatory design would meanthat, in case of online shopping, for instance, “the system would know and personalize an experience to the degree that it would feel like a magic hand guiding your experience”. It would actually change the UI (user interface) on the fly, eliminating any extraneous information, and present only the most relevant options in a timely, simple, and efficient manner.

V. Conclusion & Future Work

At present, the RSs are widely utilized for e-commerce, social networks, and various other domains, rather it is slowly becoming an integral part of the e-commerce platforms for automatic showcasing a range of articles to various customers according to their buying inclination. Since its very introduction in the year 1990’s, the research in this area has been marked as continuously evolving. RS works by adopting machine learning algorithms that facilitates the machine to learn depending upon the user’s information and to personalize further recommendation. The machine learning (ML) is basically used to predict the outcome of data processing. It has made potential breakthrough in the field of SE (Search Engine), digital security, and image recognition.

The ML consists of a number of algorithms possessing various characteristics. However, the literature has a shortcoming that it lacks proper classification system for the ML algorithms indicating the environment they are most suited in. Thus, selecting an ML algorithm, to use it in an RS, is a challenging task. The researchers in the area of RS do not have clarity regarding trends and effectiveness in ML algorithm, thus it confuses them to focus their research efforts on a specific algorithm. This study, therefore, proposes a systematic and rigorous review to observe ML algorithms to be suitably used in RSs and even the SE (Software Engineering) effort to be chosen for the development of an effective RS.

This study concludes with the followings:

1. In most of the cases, in RS development, the ML algorithm used is Bayesian or decision tree approach. Their recent popularity and the low complexity in calculation and implementation contribute it to a larger extent. In the future, more studies on the use of Bayesian algorithms in RSs can be carried to observe the implications of their use, performance, and utility.
2. Some of the literatures also propose designing the RS working with MapReduce algorithm. This result shows that it is relevant in the field of handling Big Data, though

allowing even more personalized recommendations. An exhaustive study is recommended to ascertain its proper uses versus effectiveness before using it in specific RS.

3. Most of the approaches have mathematical or statistical description of algorithms, since the ML has origins in mathematics and statistics. Thus, the mathematical basis of the algorithm must be considered in a sacrosanct manner while designing the algorithms.
4. Algorithm, for various types of RS, varies largely in performance. Thus, it is the responsibility of researcher to justify the best possible algorithm applicable at various places in an RS.

References

[1]. Adomavicius, G., & Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(6), 734-749, 2005.

[2]. Bouneffouf, D., Bouzeghoub, A., & Ganarski, A. L. Risk-aware recommender systems. In *Neural Information Processing* (pp. 57-65). Springer Berlin Heidelberg, 2013, January.

[3]. Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61-70, 1992.

[4]. Martens, H. H. Two notes on machine "Learning". *Information and Control*, *2*(4), 364-379, 1959.

[5]. Jain, A. K., Murty, M. N., & Flynn, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323, 1999.

[6]. Patrick, E. A., & Fischer, F. P. A generalized k-nearest neighbour rule. *Information and control*, *16*(2), 128-152, 1970.

[7]. Friedman, N., Geiger, D., & Goldszmidt, M. Bayesian network classifiers. *Machine learning*, *29*(2-3), 131-163, 1997.

[8]. Burhams, D., & Kandefer, M. Dustbot: Bringing Vacuum-Cleaner Agent to Life. *Accessible Hands-on Artificial Intelligence and Robotics Education*, 22- 24, 2004.

[9]. Karimanzira, D., Otto, P., & Wernstedt, J. Application of machine learning methods to route planning and navigation for disabled people. In *MIC'06: Proceedings of the 25th IASTED international conference on Modeling, indentification, and control* (pp. 366-371), 2006, February.

[10]. Torralba, A., Fergus, R., & Weiss, Y. Small codes and large image databases for recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-8). IEEE, 2008, June.

[11]. Thrun, S. Self-Driving Cars-An AI-Robotics Challenge. In *FLAIRS Conference* (p. 12), 2007.

[12]. Lv, H., & Tang, H. Machine learning methods and their application research. In *2011 International Symposium on Intelligence Information Processing and Trusted Computing* (pp. 108-110). IEEE, 2011, October.

[13]. O'Donovan, J., & Smyth, B. Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces* (pp. 167-174). ACM, 2005, January.

[14]. Adomavicius, G., & Tuzhilin, A. Context-aware recommender systems. In *Recommender systems handbook* (pp. 217-253). Springer US, 2011.

[15]. Pressman, R. S. *Software engineering: a practitioner's approach*. Palgrave Macmillan, 2005.

[16]. Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[17]. Daniel Korbut, Recommendation System Algorithm, https://blog.statsbot.co/ recommendation-system-algorithms-ba67f39ac9a3, 2017.

[18]. Guy Shani and Asela Gunawardana, Evaluating Recommendation Systems, Microsoft Research, 2010

[19]. Recommender System - Wikipedia, https://en.wikipedia.org/wiki/Recommender_system

[20]. Miklos Philips, Anticipatory Design: The Secret of Magical User Experiences, https://uxdesign.cc/ anticipatory-design-that-magic-moment-a9f34fc908e1.

[21]. Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

