Recommendation Systems: A Deep Learning Oriented Perspective

Igor Luiz Lampa, Vitoria Zanon Gomes and Geraldo Francisco Donegá Zafalon

Department of Computer Science and Statistics, Universidade Estadual Paulista (UNESP), Rua Cristóvão Colombo, 2265, Jardim Nazareth, São José do Rio Preto - SP, 15054-000, Brazil

Keywords: Recommendation Systems, Deep Learning, Collaborative Filtering, Content-Based, Hybrid Approach.

Abstract: The massive use of the digital platforms has provided an exponential increase at the amount of data consumed and daily generated. Thus, there is a data overload which directly affects the consume experience of digital products, whether at find a news, consume an e-commerce product or to choose a movie in a streaming platform. In this context, emerge the recommendation systems, which have the finality of provide an efficient way to comprehend the user predilections and to recommend direct items. Thus, this work brings the classical concepts and techniques already used, as well as analyzes their use along with deep learning, which through evaluated results has a grater capability to obtain implicit relationships between users and items, providing recommendations with better quality and accuracy. Furthermore, considering the review of the literature and analysis provided, an architectural model for recommendation system based on deep learning is proposed, which is defined as a hybrid system.

1 INTRODUCTION

The high volume of digital data currently available, whether these generated on social media, e-commerce platforms, search, news agencies and other applications, caused an overload of information. In this context, users when consuming services available on the web, are faced with an excess of options, which in many cases are divergent to their interests and not useful for their personal profile. So it is necessary to employ some methodology that is efficient to provide the information tailored to the profile and interests of the person requesting it.

According to Sarker and Matin (2021), with the increase in web development, around 2.5 quintillion of bytes are produced daily. In this sense, it is stated this high amount of data harms the taking of decisions. Thus, this situation has contributed to the development of a segment of computing called recommendation systems, which using algorithmic approaches enables the customization of content targeted at users and aligned with their expectations.

The main objective of a recommendation system is to indicate to its users what they are most willing to be interested in, providing a personalized experience and avoiding excessive and unnecessary information (Negi and Patil, 2021). Recommendation systems is vital to improve access to information in order to support the decision-making process (Zhang et al., 2019; Petter and Jablonski, 2023). The recommendation system primarily ability is to understand users behaviors and habits in relation to the items they are interacting to (Zhou, 2020). That is, recognizing which trends of user preferences and then indicate the items that match their expectations.

As specified by Da’u and Salim (2019), the development of systems recommendation systems based on deep learning has become a growing trend in present. This condition is justified by the capacity of this new technique to provide better representation learning of the interaction between users and items, when compared to traditional methods previously established in literature. In this scenario, the hybrid use of traditional methods combined with deep learning, proved to be innovative in discovering non-linear and implicit relationships between users and items. Therefore, high-quality and non-trivial recommendations are generated.

For this reason, considering the literature review and analysis provided in this work, we propose an architectural model by applying innovative machine learning techniques associated with the traditional recommendation systems, with the aim of developing...
a hybrid approach. In this scenario, we seek to reduce individual deficiencies of these methods while improving the quality of the results by using them together. So, this work focuses on providing an optimized strategy, which favors the recommendation process in terms of quality and error reduction.

This work is organized as follows: in section 2, we present a review concerning recommendation systems and deep learning; in section 3, related works are presented; in section 4, a hybrid architectural model is proposed supported by the review; and, finally, in section 5, the conclusions are showed.

2 RECOMMENDATION SYSTEMS AND DEEP LEARNING

Traditional recommendation systems are classified into three groups: content based, collaborative filtering and hybrid approaches. These classification are explained at sections below.

2.1 Content Based

The content based approach directly depends on meta-data regarding to users and items, that is, it requires detailed information related to the user profile, as well as the attributes that describe the item. In the case of users, data such as age, gender, geolocation, among others can be considered registration data. For items, assuming the example of a movie, they can be observing the attributes of gender, title, summary, classification tags, running time, year of release, actors and other data deemed necessary to describe them (Sarker and Matin, 2021). Therefore, this data is used at comparisons between the user profile and the items, to obtain the recommendations (Bhanuse and Mal, 2021).

To establish the user’s preference profile, it is necessary to use of some learning technique, that is, algorithms that reveal the relationship implicit in the direct interaction between users and items. In this sense, artificial neural networks, Support Vector Machine – SVM and Bayesian classifiers are frequently used to obtain the user profile (Da’u and Salim, 2019).

One of the positive factors of content-based recommendation is user independence, since indications are carried out in isolation, that is, disconnected from the preferred profile of other users. Another factor is transparency, denoted by the clarity of the process of obtaining items recommended. Also it is able to avoid the absence of first-rater problem, because when new items are included and it is capable to make recommendations, even without prior assessments of them. On the other hand, one of the negative factors is the occurrence of obviousness in recommendations and this condition can result in predictable indications. According to Saat et al. (2018), this phenomenon is classified as the filter bubble problem recommendation, due to the mechanism of this strategy being specific per user and not consider the rest of the community’s preferences, that is, a cycle of recommendations self-referenced by the user’s own predilections. This fact along the time provide a reduction at variations and diversity that are considered fundamental for the recommendations, moreover this point was target of critics, because it can make user to lose opportunities and newer possibilities (Grossetti et al., 2019).

Although the content based approach is capable of handling the insertion of new items, when new users are included it becomes an obstacle. This condition occurs due to the lack of interaction history of new users with the available items, which would be used for training and learning your preferences.

As described by Bhanuse and Mal (2021), there are two types of content-based approaches:

- **Case Dependent Reasoning Method:** in this model, items that are highly related are associated with those that have already been previously appreciated by the user. There is a tendency to increasing the quality of indications as the user interacts with the items available, as your history will be increased and new connections will be able to be established.

- **Attribute Dependent Method:** Recommendations are made based on the list of item attributes and its adherence in relation to the user profile. In this sense, there is no dependence in relation to user interaction with items, therefore the inclusion of new users do not create difficulties, as the recommendations will use your profile as a basis previously declared. For example, if a user is registered at a computer science area of interest, then the recommendation algorithm will use this explicitly defined attribute to make indications.

2.2 Collaborative Filtering

Collaborative Filtering is one of the most widely used methods (Da’u and Salim, 2019) and as opposed to content based, this approach does not depend of the attributes of users and items, but only of the relationship between them. In general, the mechanism is based on extracting a user’s interaction history with
the items and establish comparisons with the historical records of interaction of the other users with the items, with the purpose of obtaining new classifications (Sarker and Matin, 2021).

The fundamental concept employed by this approach is to consider the assumption that similar users have similar interests (Bhanuse and Mal, 2021). The recommendation is based on the assessment of the behavior of the community of users with the items, or even the links between the items themselves (Da’u and Salim, 2019). One of the most used algorithms and considered the best calculation methods for this purpose is called cosine similarity (Guo and Liu, 2019).

In real system applications, a lack of data about users or items is common, either due to their non-existence or difficulty in processing to obtain it (Peng et al., 2020). Therefore, the ability of this method not to depend on these attributes for its operation, is considered a positive advantage. Furthermore, it has the ability to provide indications with diversity, since does not consider the user exclusively and encompasses the analysis of all user-item relationships.

One of the factors that degrade the performance of this approach is called cold start, which refers to the condition in which there is a new item or new user. In this context, there is an absence of a user or item history, which causes a lack of resources to recognize preferences. Therefore, the system will be unable to generate assertive indications, a fact that will reduce system performance of recommendation. (Fayyaz et al., 2020; Liu et al., 2020)

Another factor that negatively impacts this method is the scarcity of data, which occurs when the volume of interaction between user and items is low. That condition generates a sparse user-item matrix (Isinkaye et al., 2015), which causes the inability to locate similar neighbors and therefore a recommendation process with low assertiveness.

The constant increase in the volume of data used in the recommendation is also considered a problem for this method. The scalability problem, is related to the difficulty that the method at deal with constant data increase. A technique called Singular Value Decomposition - SVD which is based on dimensionality reduction, also others strategies based on clustering process can be applied to mitigate this problem (Fayyaz et al., 2020).

The collaborative filtering strategy is subdivided into two segments as follows:

- **Memory Based:**
  Memory based is also categorized into two types, user based and item based. In the first approach, the recommendation process focuses on understanding similar interests among users (Bhanuse and Mal, 2021). In practical terms, items are recommended to the target user considering that they have been rated or acquired by users who resemble him. At the item based method, the objective is to establish the relationship between the items themselves and the interest in them, that is, based on the item a particular user chose to make recommendations for items that are similar to the same (Anil et al., 2018).

  In item and user-based techniques, there are factors that influence the choose between them. Firstly, the similarity of items is considered with greater importance stability, that is, the relationship established between them is very likely to stay. However, when considering the similarity between users it is the opposite, since that users’ interests evolve and transform over time, as a result, new similarity calculations will be required more frequently. Furthermore, another positive factor of the item based technique is that, in general, there are more users than items, therefore, the item-item matrix is smaller in dimensions than the user-user; that is, it can mean a competitive advantage in environments with resource limitation, whether in terms of time or even hardware.

- **Model Based:**
  On the other hand, model based is a category of collaborative filtering algorithms, that use statistical and machine learning techniques to perform model training and classifications. In this context, the core of the strategy is detect how likely a given user is to rate an item according to your interest or not, based on the classifications previously carried out (Bhanuse and Mal, 2021). Matrix factorization is a classic algorithm of this area, whose basic function is based on a sparse user-item matrix, produce probabilistic values for unfilled gaps, that is, items without prior evaluation, which will be used for recommendations.

  In the meantime and given the scientific contributions in the area of machine learning and primarily advances at the sub-area of deep learning, provided a fundamental ally for recommendation systems. Since the use of models, such as auto-encoders, Convolution Neural Network - CNN, Recurrent Neural Network – RNN, among others, aim to improve the performance of traditional recommendation algorithms by providing greater efficiency and lower error rate.
2.3 Hybrid Approach

Hybrid recommendation systems arise from the combination of positive aspects of one or more recommendation system strategies, with the aim of providing an optimized methodology, but also to mitigate individual deficiencies inherent to them when used in isolation (Da’u and Salim, 2019). The hybridization of the recommendation process can be achieved by combination of different techniques, for example, there is the classic union of algorithms content based with collaborative filtering, which has the function of providing greater accuracy of the prediction results. Several current researches reveal the importance and increase of quality of results obtained when recommendation strategies are used together. In this context, one of the fastest growing areas in joint application with traditional recommendation systems algorithms is called deep learning (Zhang et al., 2019), whose definition and use cases will be discussed in the following section of this work.

2.4 Deep Learning

This concept is considered the new generation of artificial neural networks, which traditionally has been one of the pillars of artificial intelligence and machine learning (Da’u and Salim, 2019). The objective of its application is to improve the representation of learning through multiple layers and stages of data processing, being able to learn various levels of representations and data abstraction (Zhang et al., 2019), therefore, discover implicit relationships when moving between layers.

Currently, the application of deep learning is growing and expanding, standing out in satisfactory results in areas such as computer vision, natural language processing, image processing and in recommendation (Xu et al., 2021). As Da’u and Salim (2019) highlighted, since the first publications involving deep learning with recommendation systems, there was an accelerated growth of related studies and most has been published in the last six years.

In the meantime, there is a massive demand from both academia and in industry to use deep learning for a wide range of applications, with in order to take advantage of their intrinsic ability to deal with complex tasks processes and obtain improved results. Furthermore, this methodology surprises not only due to the increase in performance in several aspects in these areas as well as its ability to learn representations from the beginning of application, that is, from zero stage (Zhang et al., 2019).

The effectiveness demonstrated in several scientific studies regarding the application of this approach with recommendation systems, enabled the development and progress in the use of this method. In this context, several algorithms that combine these two approaches, created an important scientific field with a range of emerging applications capable to produce results with greater performance than compared to traditional algorithms (Peng et al., 2020; Sarker and Matin, 2021).

3 RELATED WORKS

Some of the state of the art researches that consider the recommendation systems based on deep learning are presented as follows.

Zhou (2020) proposed a recommendation system approach with the objective of reducing the problem of high computational cost inherent to traditional algorithms when new items or users are included. In this sense, was used deep neural networks to perform similar content search for new users or new products, and then dynamically include them in the original system. Through this strategy, it reduced the number of recalculations that would be necessary and in comparison to other algorithms, there were better results for the mean squared error parameters (RMSE) and mean absolute error (MAE).

A collaborative filtering methodology applied with deep learning was proposed by Negi and Patil (2021), whose objective is to unite these two areas and obtain better results in recommendations. Two models were built, one based on stacked autoencoder – AE, with the function of identifying multiple compressed representations of the same data, and another based on Restricted Boltzmann Machine – RBM, used in the stage of building recommendations.

Sarker and Matin (2021) defined a recommendation system hybrid based on matrix factorization and deep neural networks, in addition used auxiliary information about users and items. The proposed model aims to obtain internal and implicit relationships between users and items, for this it uses factorization matrix and the Multilayer Perceptron – MLP model.

4 HYBRID MODEL OF RECOMMENDATION SYSTEM

This section aims to describe the proposed model for the system of recommendation. In this sense, it is explored the concepts of the hybrid model proposed for recommendation based on collaborative filtering and
According to Dellal-Hedjazi and Alimazighi (2020) the most of the works that use recommendation systems based on deep learning uses collaborative filtering strategy and these approaches provide high diversification are capable to deal with the constant evolution of volume of data. However, they are impacted by the cold start problem and data scarcity. On the other hand, although the content-based approach is less explored through the implementation of algorithms based on deep learning, it is essential to address the problem of cold start and data scarcity. In that sense, it is a technique capable of providing predictions, through descriptive attributes individuals, regardless of the historical interaction relationships between users and items.

In this context, the basic idea is to use a hybrid model that is capable to unify the qualities of collaborative filtering strategies with the content-based type, and that is implemented with fundamental deep learning algorithms. Furthermore, this is confirmed by Huang et al. (2019), that the hybrid recommendation algorithms are still little explored by deep learning approaches, and therefore it is an area with scope to be explored for innovative contributions.

The architecture of the proposed is organized into two main modules, namely Content-Based Deep Learning and Collaborative-Filtering Deep Learning. The first uses as input data the user attributes and items, and from them generates a model that is used to generation of recommendations. The second uses historical data of the relationship between users and items to train the model and generate preferences. Finally, the results are combined and a recommendation list is presented.

### 4.1 Content-Based Deep Learning Module

This module has an architecture based on the recommendation system developed by Dellal-Hedjazi and Alimazighi (2020) and has two concepts which are the demographic and content-based approach. That is, consider both the attributes that describe and qualify the items as well as those that are referring to users, in order to be the basis for training and generation of recommendations. In Figure 1 are showed the sequence of steps that are performed to obtain the recommendation.

The steps that make up this module are described below:

**Pre-Processing:**
Phase of receiving raw input data, in which the attributes of interest are selected. Additionally, a data cleaning process is applied for removing discrepant (outliers) or unstructured data, and filling of missing data. There is data coding, a stage in which the conversion of data from the original format to digital format so that it can be used in mathematical calculations. Finally, the application of normalization called Min-Max to reduce the data range between 0 and 1, this operation being carried out with the in order to facilitate the learning process of the neural network.

**Learning Module:**
One of the most suitable architectures for classification problems based on the attributes and characteristics of the input data is called Multilayer Perceptron – MLP, which is a deep neural network. In this case, it was dimensioned to be composed of an input layer, twelve hidden layers and a about to leave. The input layer will be fed with the characteristics of the items, for example, in the MovieLens database the attributes of title, year of release and gender, and also user attributes, as age and occupation. Regarding the intermediate part, it will be composed of twelve hidden layers, which may be resized to adapt the network in relation to the results obtained and thus allow their optimization. In observation, the choice of number of hidden layers is a relevant decision point for the algorithm, since which can generate two types of problems called overfitting and missfitting, the first occurs when the number of hidden layers is high and this generates a very large in relation to the input data, that is, the model in practice becomes specialized and is not able to generalize to new evaluated data (Sabiri. et al., 2022). The second problem is related to the inability of the model to adhere to the behavior of input data and becomes too much generic.

Finally, the output layer is responsible for generating the result of the recommendation, whose neuron with the best probability is considered the network prediction. Furthermore, it has a variable number of neurons, grouped functionally and fully connected. Basically,
the learning process starts with the pre-processed data, of which 80% will be for training, 10% validation and 10% testing, so learning occurs by passing data through the deep neural network and calculating the error between the expected result and the found result. In this way, at each interaction the error is minimized through backpropagation and the weights for each parameter matrix are saved.

**Recommendation Phase:**
Step that considers the previously trained and learned model, thus, loads the saved weights and uses them under the new pre-processed data. The result obtained is a list in descending order containing the largest values that were predicted.

### 4.2 Collaborative Filtering Deep Learning Module

The main objective of this module is to enable the extraction of predilections that are intrinsic to the history of user and item relationships. According to literary review research, there is a model that stands out for its capacity extract the relationship between users and items through multiple perspectives and that has proven efficiency in obtaining results with greater performance and quality. Therefore, this model proposed by Huang et al. (2019) is called Deep Matrix Factorization Learning – DMFL, and is used as a basis for reference so that it can be adapted and coupled to the hybrid algorithm proposed in this work. This module is visually represented at Figure 2.

![Figure 2: Collaborative Filtering Deep Learning Module.](image)

The basic operating structure is segmented into two stages, namely, feature learning and preference generation, which are described in follow:

**Feature Learning:**
Composed of two parallel neural networks that are responsible for extract the latent feature vectors of items and users. In this context, Multilayer Perceptron – MLP structures are used with the aim of extract deep and hidden characteristics regarding users and items. Due to the characteristics of the items that will be recommended are considered stable because they do not change frequently, a method is established static for training and extracting them. Therefore, the latent vector data of item characteristics are extracted directly from the static descriptive data of the items. The characteristics are considered: id, name, year, gender and other data that can describe it and that are available in the MovieLens database.

On the other hand, data relating to user characteristics can change more frequently according to changes in preferences. Therefore, this method is a process of dynamic learning which considers both basic and static data, for example, age, gender, occupation and region, as well as using the history of interactions with items. Through this history, the vectors of characteristics of the k items that the user “liked” most recently and thus allows consider the natural dynamics of changes in user preferences.

The user’s preference history vector is defined through the equation 1, whose parameter k represents the number of items that the user recently liked and y₁ is the preferred history item of user u₁:

\[
x_i = \frac{1}{k} \sum_{t=1}^{k} y_{it}
\]

(1)

Furthermore, in order to obtain the final vector of user characteristics, the vectors \(x_i\) and \(y_i\), which refer respectively to the basic data vector and static images, and the preference history vector, are inserted into the input layer of a Multilayer Perceptron. This final vector is described by equation 2:

\[
x_i = f(W_a(x_i : y_i) + b_a)
\]

(2)

Where \(f\) is the activation function of the neural network layer, \(W_a\) represents the weights and \(b_a\) is the bias of the network layer.

**Generation of Preferences:**
The result of the step described previously, which are the characteristics of items and users, is used as input to the algorithm responsible for generating user preferences. This process is based on the concept of simultaneously combining several models, so that it is possible to extract characteristics of data from multiple perspectives. In this way, in addition of being able to combine the advantages of these different models, also minimizes their deficiencies, therefore it makes it possible to provide results with greater precision.

In this context, there are three sub-modules that run in parallel and their results are merged to obtain user preferences. The modules are called SDAE-FM,
Deep Neural Network – DNN, Metric Learning as presented in Figure 3.

The sub-modules of the preference generation process are detailed below:

- **SDAE-FM:**
  - Based on two Matrix Factorization strategies – FM and Stacked Denoising AutoEncoder – SDAE, which together are responsible for reducing the dimension of features and extracting deep latent features. In this way, it is used to learn the importance of each characteristic, as well as the relationship that occurs between them and, consequently, obtains the user’s preferences.

- **Deep Neural Network:**
  - The main objective of this module is to explore the relationships deep and non-linear relationships arising from the relationship between items and users, and finally, generate user preferences with greater precision. Consider a neural network with four layers, with the calculation in each of them being carried out as equation 3:

\[
a^{(l+1)} = f(W^{l}a^{l} + b^{l})
\]  
(3)

Where \( a \) represents the input, \( W \) is the weights, \( b \) is the bias, with all parameters referring to the \( l \)-th layer. The function \( f \) corresponds to the function of activation, in which in this case the Rectified Linear Units – ReLUs method was used. In observation, at this stage the use of the Batch normalization with the aim of standardizing inputs at each layer of the neural network and prevent the model overfitting problem.

- **Metric Learning:** At this stage the main objective is to consider the users and items, and also measure them from a distance perspective. Uses two parallel convolutional neural network models with different parameters. The computational process of these networks is represented by equation 4:

\[
h_{t} = f(x_{t} + k_{t} + b_{t})
\]  
(4)

Where \( x_{t} \) refers to the input data, \( k_{t} \) represents the \( t \)-th filter, \( b_{t} \) is the bias, \( f \) is the ReLUs activation function and, finally, the symbol \( * \) is the convolution.

In summary, each of the sub-modules aims to obtain user preference for through specific perspectives: linearity, non-linearity and distance. Thus, at the end of executing the three steps there are three user preference vectors which will be processed through the sigmoid operation. In this context, when executing In this operation, the main objective is to obtain a combination of results in order to effectively optimize recommendation accuracy. This operation is defined by equation 5:

\[
z_{ij} = sigmoid(y_{ij} + y_{ij}^{d} + y_{ij}^{m})
\]  
(5)

In summary, the individual results of each of the modules will be unified and then the recommendation will be presented to the user. This process of joining the results will initially be the simple union of the final prediction vectors individual issues arising from each of the modules, however, it is a decision point for the algorithm design and that it can be evolved and used some specific operation when it is subjected to evaluation criteria and tests.

5 CONCLUSIONS

The amount of resources that are offered and consumed digitally through the Internet is constantly increasing, involving information from work, leisure, studies or simply for communication. In this context, it is noted an overload of information that is considered harmful in relation to experience of users of these resources and of the enterprises that are not able to target the content that really matters to their customers. In this way, it was verified how fundamental it is to use recommendation to support the resolution of the problem of information overload and thus employ useful strategies for it. Furthermore, it was found that the use of lonely classical recommendation techniques is not such efficient when compared to its hybrid use or supported by other methodologies.

The study and analysis of approaches that adopt deep learning in conjunction with traditional recommendation system techniques demonstrate greater quality and accuracy than classical algorithms. In addition, they provide the discovery of implicit relationships between users, items and user-items, therefore, providing better understanding of user prefer-
ences and then making predictions with greater efficiency and degree of assertiveness.

Finally, this work was conducted with the objective of provide improvements and optimizations to recommendation systems based on development of the proposed hybrid architectural model, considering both aspects of content-based strategy and collaborative filtering supported by deep learning.

ACKNOWLEDGEMENTS

The authors would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), under grant 88887.686064/2022-00, and São Paulo Research Foundation (FAPESP), under grant 2020/08615-8, for the partial financial support.

REFERENCES

Anil, D., Vembar, A., Hiriyannaiah, S., G.M., S., and Srini-


