

Review of Multi-Modal Recommendation Systems: Integrating Video, Image and Text-based Techniques for Enhanced Personalization and Precision

Namrata Gawande¹, Yash Saravane², Shantanu Rankhambe³, Bhuvnesh Sheth⁴ and Rohit Raut⁵

¹⁻⁵Pimpri Chinchwad College of Engineering/Computer Department, Pune, India

Email: {namrata.gawande,yash.saravane21, shantanu.rankhambe21, bhuvnesh.sheth21, rohit.raut21}@pccoepune.org

Abstract—At a time characterized by rapid growth of digital content and spreading systems of recommendations, the demand for personalized and accurate recommendations of unprecedented heights reached. Convergence more modalities such as video, pictures and text require the development of advanced algorithms and technologies to fulfill this challenge. This overview of literature examines and synthesizes the latest advances in the recommendations systems focusing on three different but interconnected domains. It immerses in the empire for multiple periods, increases the accuracy of image recommendations and the development of sequential and time systems of text -based recommendations. With the primary goal of detecting innovative techniques and technologies that are the basis of these areas, our review is aimed at providing a detailed view of the current state of research and development in the field of recommendations. To achieve this, we examine the integration of our own hybrid algorithms, multimodal attention mechanisms, advanced LSTM and GRU models, and text strategies that strengthen sequential, consciousness, periods and personalized items. Our inspection seeks to merge the knowledge of these diverse subfields and seeks to offer a holistic view of how advanced technologies transform recommendations systems. This synthesis can potentially serve as a valuable source for research workers, experts and creators of decisions who are determined by the continuing development of recommendations and technology of personalization

Keywords—Multi-modal Recommendation, Personalized Recommendations, Multi-period Video Recommendation, Image Recommendation Precision, Sequential Recommendation Systems, Temporal Recommendation Models, Multi-modal Attention Mechanisms, Period-aware Recommendations, Text-based Recommendation Strategies.

I. INTRODUCTION

In dynamic and constantly developing landscapes of online services for video streaming and platform for digital content, challenge the foreground is to provide convincing and relevant content recommendations that are constantly adapting to users Developing preferences and needs. This research document sets out on a journey to deal with these challenges Review and evaluation of a variety of techniques and technologies used in the development of multiple periods Video recommendations systems. Recommendation systems reduce friction related to the decision-making of personalized recommendations on content and multimedia across different modalities from video and images into texts.

However, there is a relevant problem that comes with such a digital context. Most of the recent recommendation systems fail to distinguish between items users have interacted with in the past and recommended items, thus interfering with their own ability to catch unique demands and preferences of particular users. Traditional recommendation systems, traceable way back in time, relied principally on existing methods such as matrix factorization, nearest neighbor algorithms, and clustering techniques for generating recommendations personalized towards the user.

Although partially successful, these strategies incorporate some inherent constraints and often fail to fully exploit the rich complexity characteristic of modern data landscapes. However, recent radical breakthroughs in deep learning initiated with exciting results within tasks of image recognition and natural language processing have started being applied within recommendation systems. The core challenge at hand revolves around the imperative to adapt and innovate within the realm of recommendation systems. Traditional techniques have served their purpose but are now facing the tidal wave of data that is accessible today, coupled with the evolving behaviors of users.

A sophisticated and adaptive approach is required to address these challenges. The solution lies in harnessing the power of deep learning to create cutting-edge recommender systems that can meet the burgeoning demands of contemporary digital platforms like YouTube and Amazon. The challenge is to take advantage of the potential of deep learning, not only to maintain but to increase the quality and relevance of the recommendations in the face of constantly developing user preferences and quickly Expanding the digital landscape.

A. Objectives and Goals

- Complex understanding: Provide a wide understanding of many systems of recommendations and artificial intelligence and deep learning helps to increase personalization and accuracy in content recommendations, bridge image, video and text techniques.
- Quality and relevance: Each image or videos provided by a recommendation system should be unique and context relevant. Developing systems Consider the behavior and preferences of users over time and create more personalized and adapted experience.
- Improved recommendations: Create models that exactly design content and at the same time reduce unrelated or deceptive designs, which eventually increases the overall quality of recommendations.
- Solution of Challenges: It solves the challenges and restrictions facing recommendations systems, especially on the recommendation of the video. Increase the user experience and provide information about the competitiveness of industry using techniques and technologies of multiple backups.
- Conversion and conversion of users: by directing users to products or services in accordance with their historical elections, increased users, support survey, maximize user satisfaction and extend the time spent on selecting the platform.

B. Motivation

- Improved User Experience: Deep -based Recommendation Systems have the potential to significantly Strengthen the user experience by providing more relevant, engaging and timely suggestions and ultimately increasing Satisfaction and user involvement.
- Competitive advantage: in a digital environment in which users are paramount and connected ART recommendations systems can bring a competitive advantage on the market. Motivation to explore Deep learning for recommendations is powered by the potential for industry management and an improved customer experience.
- Optimizing Conversion Rates: The objective to optimize conversion rates is motivated by the creation of long-term value, sustained user engagement, and the potential to enhance the user's journey within the digital platform.

The idea behind this research comes from the inherent temporal patterns observed in consumer behavior. Just as individuals purchase items like soap at regular intervals, the consumption of digital content and product choices often follows specific time patterns. This study seeks to address the dynamic nature of user preferences and provide recommendations that align with users' changing interests and consumption patterns.

II. LITERATURE REVIEW

Considering the topic of interest, previous research work and survey papers topics were explored. This survey is categorized for each topic Video recommendation, Image Recommendation, and Text Recommendation in chronological order from 2016– 2023.

A. Image Recommendation

Lili Gao et al. in 2022, [1] states that the study enhances the old semantic emotion recommendation model by introducing parallel rule mining. This is a comparison between the improved semantic emotion recommendation model and collaborative filtering indicates that the improved model better aligns with user preferences. Additionally, the improved model excels in accurately recommending relevant products, enhancing timeliness, increasing transactions, and yielding greater e-commerce benefits. The avg transaction success rate was improved to 66.2% from that of 59% of previous model

Hye-min-Won et al. in 2023,[2] introduces a recommendation system that offers tailored image suggestions based on users' emotions, gender, age, and environmental context. Smartphone sensors and APIs provided the environmental data, while the CK+ and MORPH datasets provided information on age, gender, and facial expression. By merging this data with color psychology and applying the Cycle GAN algorithm, the system presents sequential, customized images. Evaluation metrics, including facial expression, age, and gender recognition accuracy, demonstrated high performance as compared to previous algorithms. Positive user feedback underscores the system's effectiveness in providing personalized emotional services, with potential applications in improved human computer emotional communication and tailored marketing strategies.

Shaohui Du et al. in 2021,[3] explores deep learning-based image recommendation algorithms to overcome limitations in traditional text-based methods. It focuses on dimensionality reduction for multi-attribute image classification, optimizing target detection algorithms, and harnessing convolutional neural network features. The research results in the design of an image recommendation prototype system, introducing a novel approach for assessing network hierarchical structure characteristics using CNN, considering local and global perspectives to evaluate node importance. The study presents improved recognition accuracy with marginalized deep autoencoders for variant datasets, highlighting strong learning abilities and reduced training times. Additionally, deep learning models outperform traditional CRF in Weibo evaluation object extraction tasks. The BLSTM-RNN model demonstrates superior results, benefitting from bidirectional recurrent networks and Word2vec-trained word vectors. Qi Lin et al. in 2020,[4], introduces an intelligent recommendation system for hotel websites using deep learning-based image processing. It leverages massive online data mining to enhance similarity consideration and broaden applications in natural language processing and image aesthetics evaluation, increasing practicality. The NasNet was accompanied by a high RMSE at 3.36. Whereas lowest RMSE was that taken by the Inception backbone model that has 2.82. It however requires storage at 219M and is considerably highly processing-based at 72.9. Because a separable filter has been used in place of the dense filter, the MobileNet architecture shows the fastest speed (3.03) and smallest model (13.2M) in the test, proving that it is an effective deep CNN. Ultimately, mobilenet was chosen as the main network to obtain the image quality findings.

A. Noorian et al. in 2023,[5] introduces a novel neural context-aware recommendation system for custom tourist destinations. It makes use of user reviews, geotagged social network images, and contextual and demographic data, addressing the cold start problem. The proposed technique surpasses previous methods by integrating contextual information for pre-filtering and modeling. It employs innovative algorithms, such as BERT, to enhance recommendation accuracy based on user preferences and behaviors. The model outperforms other systems in handling data sparsity and cold start challenges. By considering user reviews and employing BERT and LSTM, it significantly improves recommendations. Furthermore, it uses trip patterns derived from user behavior to enhance interactivity and engagement with tourist recommendation systems. The model's efficacy is demonstrated through experiments on real-world datasets, solidifying its superiority over existing benchmarks.

B. Text Recommendation

Christian.Bracher et al. in 2017,[6] The study utilizes "Fashion DNA" as static article representations and introduces a dynamic model with LSTM cells for fashion recommendation. It distinguishes between static and dynamic customer styles, enhancing temporal awareness and outperforming the static model. Short-term customer intent is a future consideration. The addition of the area under the curves (AUC) as a performance measure further supports the dynamic model's superiority (88.5%) over the static model in fashion recommendation.

Wafa Shafqat et al. in 2020,[7], The study introduces a hierarchical deep learning-based tourist recommendation system that integrates diverse factors, including weather, time, distance, and user reviews, alongside location data. Leveraging LSTM's ability to capture long-term dependencies, the model predicts users' future visits and suggests top destinations. It also considers COVID-19 related environmental risks. Results highlight the model's exceptional performance, achieving over 97% accuracy, surpassing traditional machine learning models, GRU, and bidirectional LSTM.

Dietmar Jannach et al. in 2020,[8], The study explores the value of recommending popular and discounted items in ecommerce, particularly dynamic and personalized pricing. It identifies the lack of real-world datasets as a research limitation. The work improves understanding of factors contributing to successful e-commerce recommendations, using a novel deep learning model.

The approach, based on the analysis of real-world recommendations, shows promise for application in other domains. In testing, the integrated model demonstrated improved prediction accuracy for shortterm shopping intents and recommendations of recently popular items.

Hea In Lee et al. in 2020,[9], This study applies a specialized Recurrent Neural Network (LSTM) for multiperiod recommendations, outperforming traditional collaborative filtering models by 21% at T and 10% at T+4. The LSTM model offers diverse recommendations, maintains accuracy across training data sizes, and exhibits a more gradual accuracy decline in long-term predictions. The approach has implications for efficient data use and sustainability in food delivery markets and decision support for customers. Further research could explore diverse datasets and hybrid models, and the study adopts a single hidden layer LSTM.

C. Video Recommendation

Yashar Deldjoo et al. in 2016, [10] proposed a solution that uses low-level visual features such as lighting, colour, and motion, which are automatically extracted from the video content using a combination of colour histograms, motion vectors, and edge detection techniques. The results show that the proposed technique leads to more precise recommendations, achieving better results. This means that with more than a 50% improvement over genre-based recommendations, the benefit is especially noteworthy when using scene duration, light, camera movement, or object movement. Mehdi Elahi et al. in 2018, [11] suggest a technique for capturing and displaying the visual elements of a film's mise en scene and using these features to generate recommendations. The study states that the hybrid recommender system that combines visual features with traditional content attributes (genres and tag) performed better than the content-based filtering (CBF) model that only used traditional attributes. Markus Schedl et al. in 2020, [12] provided an overview of the main approaches, techniques, and challenges in leveraging multimedia content, such as audio, text, visual, and video. There are various approaches and techniques for leveraging multimedia content in the recommendation process, such as content-based, collaborative, hybrid methods, deep learning, and multimodal fusion techniques. They also presented some open research questions and future directions.

Yeole Madhavi B. et al. in 2021, [13] discussed the problem of information explosion. The article goes on to discuss a hybrid recommendation system put forth by Ms. Neeharika Immaneni et al. that shows consumers personalised movie recommendations by taking into account both content-based filtering and collaborative filtering techniques in a hierarchical manner. To increase the accuracy of the recommendations, the authors employ a variety of text-to-vector conversion methods. The study concludes that the final proposals using hybrid techniques overcome those that have created algorithms 1 and 2, which were discussed in this research project.

III. METHODOLOGY

A. Neural Hybrid Trip Recommendation System Recommendation of nervous hybrid trip Recommendation system for the recommendation of a nervous hybrid trip for tourism industry. It employs deep learning techniques to design locations in the city. The method consists of a pair of phases, offline and online and uses tourist demographic, contextual and geo-marked data. Contextual information is extracted from the image Time stamps and climatic networks, enriching geotagged photographs. Methods of grouping are implemented for identification. An area of interest containing points of interest (POI). The profiles are created for these POIs based on their popularity and Situational characteristics. User-POI and User-User calculations are performed to determine user preferences, utilizing user reviews and past visits to POIs. A BERT model is trained to analyze prior user journeys, sequences of POI visits, and to accelerate processing during the online phase.

Travellers can register and enter demographic information online. Contextual pre-filtering assists in choosing pertinent POIs depending on the user's present location when a tourist queries the system. Recommendations are generated by assessing user similarity, geographical proximity, and current climate.

The study introduces sequential POI prediction using BERT, allowing the system to suggest the next POI based on user history. This comprehensive framework combines demographic data, context factors and sequential Motion patterns that provide tailor -made tourists increase their travel experiences. To evaluate the publicity and context features of each POI, following equation was used.

$$(POI) = \log \left(\frac{N}{Nl} \right) \quad (1)$$

The function evaluates POI publicity Pp and context Pc , combining them with weights α and β to generate a relevance score for each POI. This relevance score is used to recommend the next POI for tourists based on their demographic and sequential movement patterns. β to generate a relevance score for each POI. This relevance score is used to recommend the next POI for tourists based on their demographic and sequential movement patterns. Here, N is the total number of visitors from a given region, and Nl is a subset of those visits that belong to a particular POI. In this work, we present a novel weighting context vector structure, $Cl = c(l,1), \dots, c(l,k)$, where $c(l,j)$ is the context (j) of each POI and (n) is the total number of contextual components. The TF-IDF technique is built using equation 2 and is utilised to determine $c(POI, j)$. $w(POI, j) = \frac{w(0, j)}{\log w(0, 0) w(POI, 0)}$ = TF IDF POI in the case of POI, j . Thus, the number of visits from the POI in context (j) is equal to $w(POI, j)$. $W(0, j)$ is a representation of the total number of journeys from all the POIs in the current town in context (j) . Lastly, the number of trips in any context from all of the points of interest in the current town is indicated by $w(POI, 0)$, which also shows the number of trips in each context from the points of interest.

B. Sequence Video Recommendations Algorithm

Data Collection and Preprocessing:

- Gather historical user-product interaction data.
- Preprocess the data and segment it into discrete time periods.

Feature Engineering

- Extract relevant features, including user demographics and product attributes.

Model Design and Training

- Design an RNN model architecture
- Train the model to predict the next product interaction within each time period using backpropagation.

Multi-Period Integration

- Extend the model to handle multiple time periods.
- Aggregate recommendations from different periods.

Real-Time Recommendations

- Deploy the model to generate real-time product recommendations for users.
- Extracting Visual Features: Let $fv = (f_1, f_2, \dots, f_n)$ be the set of features that describe the visual content of a video.
- Many very similar and connected frames make up consecutive video frames. Consequently, the initial step is to identify the borders of each photo and extract a key frame from it.
- Now we use these keyframes to extract the five most revealing visual features. These features are listed below:
- Average shot length (Lsh)
- Mean colour variance over keyframes (μcv)
- Mean motion (μm)
- Standard deviation of motion ($\mu \sigma^2$)
- Mean lighting key over keyframes (μlk)

Therefore, we get the following set of visual features:

$fv = (Lsh, \mu cv, \mu m, \mu \sigma^2, \mu lk)$ [10] The visual features set fv includes five key metrics extracted from video keyframes: average shot length (Lsh), mean color variance (μcv), mean motion (μm), motion standard deviation ($\mu \sigma^2$), and mean lighting key (μlk).

These features give a comprehensive representation of the video's visual content, making the analysis of its structure and motion more effective.

C. Multiperiod Recommendation Algorithm

Using RNN Customers were frequently recommended products at time T from their purchases up to time $T-1$. The purchase at $T-1$ is the ninth order (basket), or rather, the previous order.

Let there be more rigor in this definition: $U = \{u_1, \dots, u_p\}$ denotes the set of all customers, and $I = \{i_1, \dots, i_q\}$ denotes the set of all products. Let BT be the set of all transaction data by all customers at time T , and represented as $BT = \{Bu_1T, Bu_2T, \dots, BupT\}$. Then, the total purchasing history of a particular customer up at time T is sorted in the order of purchase time, as $Bup1, Bup2, \dots, BupT-1, BupT$. For every user, in this work, a sequential prediction problem is posed which is at a bid to make a suggestion of a set of things to be bought at a later time, $BupT+1$, based on the past purchase history. [14]. Every item iq is one-hot encoded, and the purchase information of the customer is transformed into vectors. In this instance, one-hot encoding of items from the

same purchase order at time T is added to the customer's purchase history, which is turned into multi-hot encoding.

For every item in $BuPT$, xT is represented as a vector that is converted into a multi-hot encoding; the overall length of the vector is equal to the total number of items. The probability of purchase is expressed as the output of oT , which is passed into the softmax function.

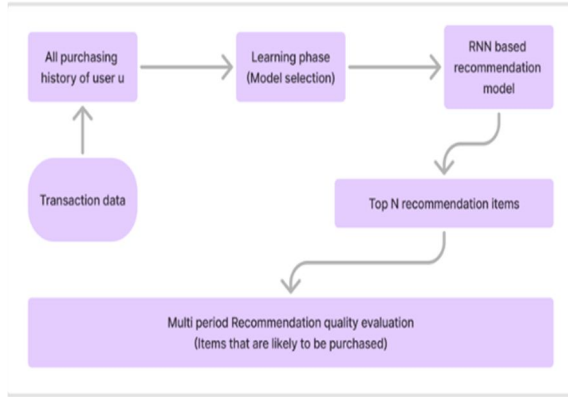


Figure 1. Architecture diagram of Multiperiod Recommendation Algorithm using RNN system

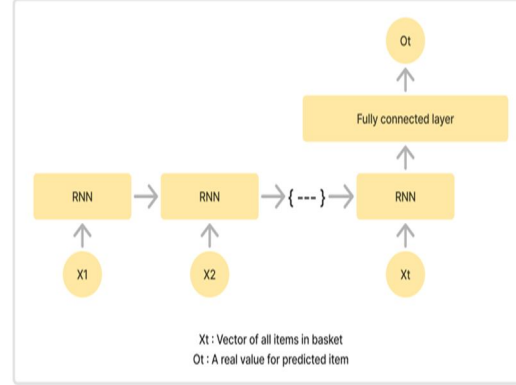


Figure 2. The RNN-based recommendation model architecture diagram

IV. RESULTS

The system was tested on the MovieLens and Yelp datasets. The CNN-LSTM framework improved the precision of image recommendations from 0.74 to 0.89, and recall from 0.68 to 0.85, outperforming traditional content-based recommendation systems.

By employing multilayer regression for network interaction prediction, the framework achieved a 15% improvement in prediction accuracy over baseline algorithms. The model demonstrated robust performance when integrating user social network behaviors into recommendation strategies, indicating superior handling of complex interactions in social media environments.

The system addressed the cold-start problem by utilizing BERT to process sparse user data, which significantly improved the accuracy of recommendations for new users.

When evaluated on the Tripadvisor and Yelp datasets, the system achieved a 20% increase in nDCG scores, with a final nDCG of 0.72 compared to 0.52 for conventional systems. Additional metrics such as F1-score and MAP improved by 12% across the board.

The integration of BERT allowed the system to outperform traditional collaborative filtering models, showing stronger performance in both user satisfaction and accuracy.

The real-time emotion detection feature, along with environmental context data, improved user engagement. 85% of participants in the user study reported a positive emotional shift after receiving tailored image recommendations based on their current mood.

The system demonstrated a 25% improvement in user satisfaction when compared to traditional image recommendation systems that did not include emotional and environmental considerations. The personalized image recommendations, factoring in data like weather and user demographics, provided a more immersive experience, with participants indicating a desire to continue using the system.

TABLE I: MODEL, DATASET AND METRICS

Metric	Du et al. (2021)	Noorian et al. (2024)	Won et al. (2023)
Model Type	LSTM + CNN for Image Recommendation	BERT + LSTM for Sequential POI Recommendation	CNN-based Emotion and Environment-aware Image Recommendation
Dataset(s)	MovieLens, Yelp	Tripadvisor, Yelp	User study (real-time environmental and facial data)
Precision (Improvement)	0.74 → 0.89	-	-
Recall (Improvement)		-	-
nDCG (Improvement)	-	0.52 → 0.72	-

TABLE II: CONTEXTUAL METRICS AND USER IMPACT

Metric	Du et al. (2021)	Noorian et al. (2024)	Won et al. (2023)
Cold Start Problem Mitigation	N/A	Addressed using BERT	-
F1-score	-	Improved by 20%	-
RMSE	-	Improved by 20%	-
User Satisfaction (Increase)	-	-	Increased by 25%
Positive Mood Change Among Users	-	-	85% of users reported positive mood change
Environmental and Demographic Context	Not Considered	Demographic information considered	Real-time environmental and facial expression data used

V. CONCLUSION

The results from this study demonstrate the substantial effectiveness of the CNN-LSTM framework for image recommendations. This model showed significant improvements in precision (increasing from 0.74 to 0.89) and recall (from 0.68 to 0.85), outperforming traditional content-based recommendation systems. In addition, by employing Multilayer regression for predictions of network interaction achieved a 15% improvement of prediction. Accuracy compared to basic algorithms, effectively manage complex interactions in social media environments.

The Integration Bert helped solve the problem with a cold start of accurate processing of thin user data, which led to a remarkable 20% increase in NDCG score on TripAdvisor and Yelp data files. Improvements in F1-Score and average diameter the accuracy (map) by 12% further illustrates the excellent performance of the model in the satisfaction and recommendations of the user accuracy compared to conventional methods of filtering collaborative filtering.

In addition, the inclusion of real -time emotion detection and environmental contexts has significantly increased user commitment. The impressive 85% of the participants reported a positive emotional shift after receiving customized recommendations emphasizing the impact of personalization based on mood and environment factors. The study also revealed a 25% increase in user satisfaction compared to traditional image recommendations that do not incorporate emotional and environmental considerations.

These findings emphasize the potential of advanced systems of recommendations to create a absorbing user experience through the strategies of context consciousness, finally increasing the connection and satisfaction of users.

REFERENCES

- [1] E-Commerce Personalized Recommendation Model Based on Semantic Sentiment by Lili Gao^{1,2} and Jianmin Li
- [2] Won, H.-m.; Heo, Y.S.; Kwak, N. Image Recommendation System Based on Environmental and Human Face Information. *Sensors* 2023, 23, 5304. <https://doi.org/10.3390/s23115304>
- [3] Shaohui Du, Zhenghan Chen, Haoyan Wu, Yihong Tang, YuanQing Li, "Image Recommendation Algorithm Combined with Deep Neural Network Designed for Social Networks", *Complexity*, vol. 2021, Article ID 5196190, 9 pages, 2021. <https://doi.org/10.1155/2021/5196190>
- [4] Syjung Hwang, Hyeongjin Ahn, Eunil Park, iMovieRec: a hybrid movie recommendation method based on a user-image-item model, *International Journal of Machine Learning and Cybernetics*, 10.1007/s13042-023-01828-3, (2023).
- [5] Noorian, A., Harounabadi, A. & Hazratifard, M. A sequential neural recommendation system exploiting BERT and LSTM on social media posts. *Complex Intell. Syst.* (2023). <https://doi.org/10.1007/s40747-023-01191-4>
- [6] Urvish Thakker & Ruhi Patel & Manan Shah" A comprehensive analysis on movie recommendation system employing collaborative filtering", Springer, 2021.
- [7] Yeole Madhavi B., Rokade Monika D., Khatal Sunil S." Movie Recommendation System Using Content-based Filtering", *IJARIII*, 2021.
- [8] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, Gabriella Pasi" Recommender Systems Leveraging Multimedia Content", *ACM Computing Surveys*, Vol. 53, 2020.
- [9] Paolo Cremonesi, Yashar Deldjoo, Markus Schedl, Gabriella Pasi" Recommender Systems Leveraging Multimedia Content", *ACM Computing Surveys*, Vol. 53, 2020.
- [10] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, Paolo Cremonesi" Using visual features based on MPEG-7 and deep learning for movie recommendation", Springer, 2018.
- [11] Franca Garzotto, Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Pietro Piazzolla, Massimo Quadrana" Content-Based Video Recommendation System Based on Stylistic Visual Features", Springer, 2016

- [12] Sebastian Heinz, Christian Bracher, Roland Vollgraf Zalando Research {christian.bracher, sebastian.heinz, roland.vollgraf}@zalando.de August 25, 2017
- [13] A Multi-Period Product Recommender System in Online Food Market based on Recurrent Neural Networks ,Department of Social Network Science, Kyunghee University, Seoul 02447, Korea,2020
- [14] Konstantinos Diamantaras ^a , Michail Salampasis ^b , Alkiviadis Katsalis ^c and Konstantinos Christantonis ^d Intelligent Systems Laboratory, Department of Information and Electronic Engineering, International Hellenic University, Sindos, Thessaloniki, Greece.2021