Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
2021, Vol. 18 No.1 (Suppl. March) **E-ISSN: 2411-7986**

DOI: http://dx.doi.org/10.21123/bsj.2021.18.1(Suppl.).0746

User-Oriented Preference Toward a Recommender System

Pei-Chun Lin [*1] ***Nureize Arbaiy*** [2]

[1]Department of Information Engineering and Computer Science, Feng Chia University, No. 100, Wenhwa Rd., Seatwen, Taichung 40724, Taiwan
[2]Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia
[*]Corresponding author: peiclin@fcu.edu.tw[*], nureize@uthm.edu.my
[*]ORCID ID: https://orcid.org/0000-0003-0735-2693[*], https://orcid.org/0000-0002-9535-8384

Received 11/12/2020, Accepted 15/3/2021, Published 30/3/2021

Abstract:

Nowadays, it is convenient for us to use a search engine to get our needed information. But sometimes it will misunderstand the information because of the different media reports. The Recommender System (RS) is popular to use for every business since it can provide information for users that will attract more revenues for companies. But also, sometimes the system will recommend unneeded information for users. Because of this, this paper provided an architecture of a recommender system that could base on user-oriented preference. This system is called UOP-RS. To make the UOP-RS significantly, this paper focused on movie theatre information and collect the movie database from the IMDb website that provides information related to movies, television programs, home videos, video games, and streaming content that also collects many ratings and reviews from users. This paper also analyzed individual user data to extract the user's features. Based on user characteristics, movie ratings/scores, and movie results, a UOP-RS model was built. In our experiment, 5000 IMDb movie datasets were used and 5 recommended movies for users. The results show that the system could return results on 3.86 s and has a 14% error on recommended goods when training data as $K = 50$. At the end of this paper concluded that the system could quickly recommend users of the goods which they needed. The proposed system will extend to connect with the Chatbot system that users can make queries faster and easier from their phones in the future.

Keywords: Correlation coefficient Analysis, K-Nearest neighbors (KNN) Algorithm, Recommendation system, Regression analysis.

Introduction:

The main purpose of the recommendation system is to get useful and user-friendly information from many messages or product orders so that users can make optimized choices. This saves time for consumers to find information and products, and makes products easier for others to buy, and creates business opportunities. If the recommended users are more accurate, they will be the exclusive customers of this recommendation system. The purpose of building a recommendation system is to reduce the time to search for goods, goods, good products, movies, music, etc. Typically, the system will recommend items that are in high ratings and level. If the recommended item is suitable, it will bring more business benefits to the retailer. Finding information about user information in the Internet era of Big Data is a very difficult process due to a large amount of information and inefficient search methods. therefore, this paper provided information filtering technology to help users find what they want. This paper made a process in using the K-NN algorithm, Euclidean distance, and Bayesian classification to design the proposed systems and evaluate the performance in this paper. By creating a user profile, the content of previous user information and course rating actions are compared and compared to the nature of the course to generate the recommended course. The technology also needs to use coach in-formation to make recommendations, so without user information, users cannot recommend the online courses they need.

The proposed system can be divided into four ways of doing collaborative (based on Cooperation) filtering. In terms of Popularity, Content-based, Hybrid Model, this paper used the

public network dataset IMDb. Collect content to suggest appropriate movies and change algorithms explore revenue differences and find recommended methods. The scope of this study is to use open datasets and data on the Internet, the IMDb movie dataset for analysis. This dataset was chosen because it requires a lot of rating parameters and complete movie information. The IMDb data set contains the most complete data and current file parameters and has full credibility. By 2018, IMDb has 83 million registered users and 5.3 million users. Movie and drama titles, and a selection of 9.3 million individual users.

However, IMDb dataset only has movie content parameters and no user information. This paper wants to simulate the real-time system provided by the movie theatre in the future. At the same time, there would not be too many movies in the release, so this paper adopted the movie's 5,000 IMDb dataset information to explore the relationships between parameters, equations between parameters, and use the content of the dataset to create constraints.

This paper contains five sections: Section 1 presents the Introduction, and Section 2 contains Literature Reviews. Section 3 introduces the methodology. The experimental process is presented in Section 4. Section 5 concludes the results and future works.

The four most commonly used systems of recommendation technology are presented (1). The first is the most common and most popular use of technology (2-5) using a Collaboration-based recommender system that uses long-term user information. For example, A and B might have the same preference for an item. If there is a new item A, the priority is higher than other random users' opinions. The second is Popularity (Average Population). This method is the most recommended way to recommend the system. The average score obtained using the item is used as the recommended order (6). There are methods based on user information and preferences in the system, but they are based on user requirements and cannot provide new user suggestions. The author uses this to create a recommendation system that uses popularity recommendations. The third is Content-based: A recommended way to find the equivalent of a user's favorite item and use the information to advise the user. The last technique is the Hybrid Model like (7-9). This is a way to improve collaboration filtering, based on content, as datasets may vary from time to time. If the data is always old, the data may lose its accuracy, and the user's association with the item may seem simple. However, it can also be a very complicated relationship.

Thus, the graphical model can be used to add time factor changes. Bayesian networks are used (10). Besides, matrix decomposition is used in (11, 12) to improve and solve collaborative filtration, a content-based suggestion method.

It is a very challenging method to find information about user-related information in the Big Data age on the Internet. This makes it important for information filtering technologies to help users find the information they need. Moreover, the program must also be able to locate the data correctly and rapidly. In this research, the authors used the K-NN algorithm, Euclidean distance, and Bayesian classifier to design and evaluate the performance of a course recommendation system. The user's prior knowledge content and the activities of the ranking course are contrasted by creating a user profile and contrasted with the course attributes to create a recommended course. This technology often needs to use information from the teacher to make suggestions, because students can not recommend the online courses they need if there is no user information.

Because of this, many researchers worked on combining different methods into a recommender system, such as Danil, B. et. al. use similarity measures to the recommender system (13). Also that Li, et al., consider the user's feature and interest in the movie recommender system (14). Desrosiers et al., used neighborhood-based recommendation methods (15). Considering the hyper uncertain environments and users' indefinite emotions, Lin, et al,. also provided many models to solve the uncertain problems on recommender systems (16, 17).

Research Methodology:

This study uses K-NN to perform type analysis, find similarity parameters to join algorithms, and experimental data sets using IMDb movie data sets. Parameters include budgets, movie types, movies, and more. Among the many parameters, comparing the parameters that have the most relevant and most influential influence on the film. To more clearly our process of data analysis, the architecture of the UOP-RS is given in Fig.1.

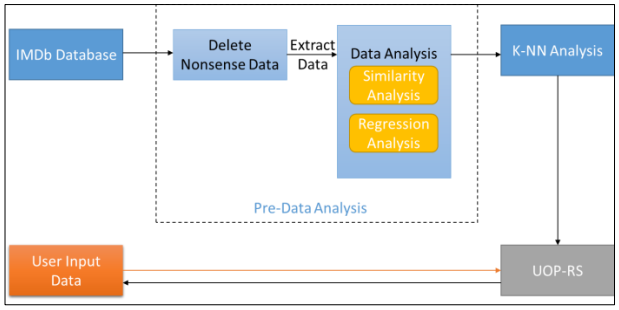

Figure 1. An Architecture of UOP-RS

Pre-arranged Data

First, data with missing values is identified, and the unnecessary ones are deleted first. For example, the official website does not affect the movie data itself and then extracts the keywords that appear in the data set and type of movie as a classification. The list of data set parameters is as in Table 1.

Table 1. Data set parameter

budget	language	production_countries	director_name	vote_average
genres	overview	release_date	actor_1_name	vote_count
id	popularity	revenue	actor_2_name	title_year
plot_keywords	production_companies	run_time	actor_3_name	country
status	cast	spoken_languages	title	original_title
tagline	crew	homepage	movie_title	

Then the Pearson correlation coefficient was used to find the correlations and equations between the parameters. The movie parameter correlation matrix was created, and then used category classification and movie correlation values, IMDb scores and netizens vote count as the recommended rating standard. Score equations for comparison are obtained. Pearson correlation coefficients are used because there are many parameters in the IMDb data set. Some parameter fields have many correlations and reductions, such as movie revenue and movie budget, and so on. So Pearson's correlation coefficient is used to help find the two results.

Finally, changing the value of K type analysis and algorithm parameters to optimize the recommended results and compare the withdrawal rate, mean absolute MAE error (Maximum Error), and RMSE means root error of different K (Root Quadratic Error). Implementation time is considered a performance evaluation for this system.

Pre-processing of data involves removing unnecessary parameters such as the movie's official website, original movie title (not the movie later releases name), movie language. Then, the missing values for each parameter are identified. If anything is missing from the movie list. The value continues to delete this item.

The official website of the movie was deleted because this parameter has nothing to do with the content. The original movie title does not match the title of the last movie, so this parameter was also deleted. The original language of this movie can be presented in a world-wide way, and will not be limited by language issues, so this parameter is not used to perform the recommended rating project.

The remaining data after filtering is 4803, and there are still 26 parameter forms. First, we define the keyword forms in the dataset, identify the type of video, and use K-NN to classify the film group. This helps us to better understand in this data set where keywords are, and then apply analytics to find better movies. There are more than one kind of film. For example, a movie (Science, Science Fiction) is included in a movie. The number of movie categories (10394) in the following pictures will exceed the number of movies (4803).

Similarity Analysis

The European similarity matrix is then used to calculate distance parameters to facilitate the design of the final proposed system algorithm. This distance lets us know the similarities between the two parameters. The Euclidean distance formula is as follows:

$$d_{M,N} = \sqrt{\sum_{i=1}^{N}\left(x_{M,i} - x_{N,i}\right)^2} \qquad (1)$$

Figure 2 uses the movie year (title_year), movie number (id), movie cost_budget, movie popularity, movie vote (vote_count), movie box office_revenue, the movie length of the film (run_time), and the voting average score (vote_average) of the movie to make similar comparisons, so the matrix can know that the budget, popularity, vote count, and revenue are the most similar among the parameters, so this paper uses these parameters to extend the final recommendation algorithm formula.

The similarity matrix provides information about the estimates, popularity, vote count, and results closest to the parameters. These parameters are then used to extend our final proposal algorithm formulas.

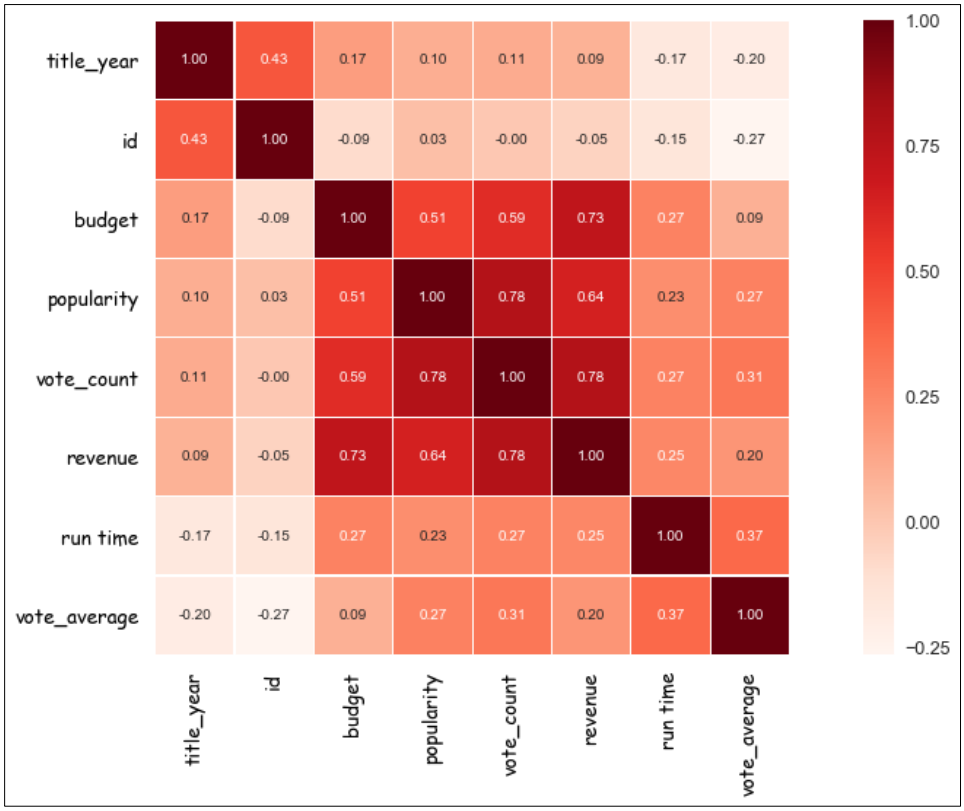

Figure 2. Parameter similarity matrix

Regression Analysis:

Regression analysis is widely used in statistical analysis data to understand whether there is a correlation, direction of correlation, and strength of correlation between two variables or multiple variables to form a mathematical formula model. The algorithm can obtain and adhere to certain parameters to predict interesting variables. In particular, regression analysis can help us to understand the variation of variables corresponding to parameter changes. In this section, the system will use the two-parameter budgets (budgets) and the box office (revenue) obtained from the similarity matrix for regression analysis. This is to find the correlation between them, and check whether one needs to take one as the standard for scoring.

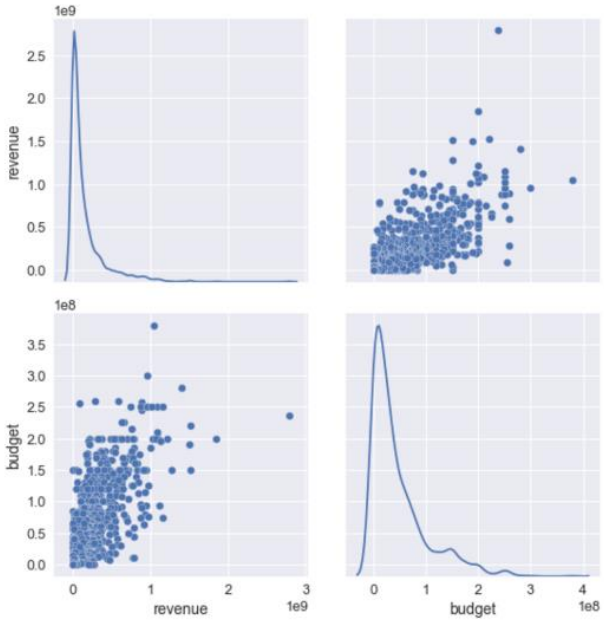

Figure 3. Movie budget and income regression analysis chart

From this regression analysis, the results show that the budget and income are close to each other (Fig. 3). In the movie recommendation, the income is chosen as the formula parameter for the recommended score. In the end, the system retained the original IMDb movie score and added the square to make the original score better and better as the scoring standard, plus the box office income obtained in this section and the same total number

of keywords in the movie as the recommended formula. The final recommendation is as follows:

$$\text{Result}_{\text{score}} = \text{IMDB\_score}^2 * \text{revenue} * \text{keyword} \quad (2)$$

➢ Note that the meaning of formula sentences is given as follows:

1. $\text{Result}_{\text{score}}$ denoted as Final recommendation movie score.
2. IMDB_score denoted as Original score of IMDb
3. revenue denoted as Movie revenue from cinema
4. keyword denoted as Total number of the same keywords

Experiment Result and Analysis:

In this section, the accuracy and performance of this proposed system are analyzed through experimental results. The system also tried different parameters and methods to find the best method. The development used in the experiment is discussed. Software, programming languages, and hardware devices. In the second section, the number of experimental movie categories are compared to let us know that more recommended categories should be recommended in the best way.

The experimental hardware used in this paper is the 2017 Apple MacBook Pro, the interpreter (IDE) used is the iPython (Jupyter notebook) development environment, written in Python, and evaluates performance, draws bar graphs, and predicts recommendations. and many more.

The film list produced by the film classification is a movie with related categories and a movie with no correlation category. The confusion matrix of the classification index helps us to conduct an experimental evaluation and the related movie in the confusion matrix. It is judged that the relevant quantity is True Positive (TP), the number of related movies judged to be unrelated is False Positive (FP), and the number of irrelevant movies judged as unrelated is True Negative (TN), and irrelevant movies are judged to be related The number is False Negative (FN), however, the exact value is our most common indicator of performance, but since comparing the same type of film, it is meaningless to explore Accuracy here. Therefore, using the recall rate (Recall) to detect the performance of this classifier. The recall rate is to observe how many related movies can be found under actual conditions and how many correct predicted answers can be recalled. The formula for the recall rate is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

For K-NN Training data, we used $K = 20, 30, 50, 100$. The number of these four categories is $0.75\ (K = 20), 0.83\ (K = 30), 0.9\ (K = 50)$, and $0.88\ (K = 100)$. This is because the $K > 50$ recall rate does not increase. So $K = 50$ is the best solution for this number of categories. Because the execution time of the recommended system is also a very important performance evaluation. In Fig.4. , the same number of classifications is listed in Fig.3. , and the execution time is $3.44\ s\ (K = 20), 3.55\ s\ (K = 30), 3.86\ s\ (K = 50), 3.87\ s\ (K = 100)$. However, the most important thing about the recommendation system is accuracy. So, using the mean absolute error MAE (Mean Absolute Error) and the root means square error RMSE (Root Mean Squared Error) as the evaluation index for our classification, the average absolute error. The formula for the mean absolute error and the root mean square error is as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |(y_i - y'_i)| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - y'_i)^2} \quad (5)$$

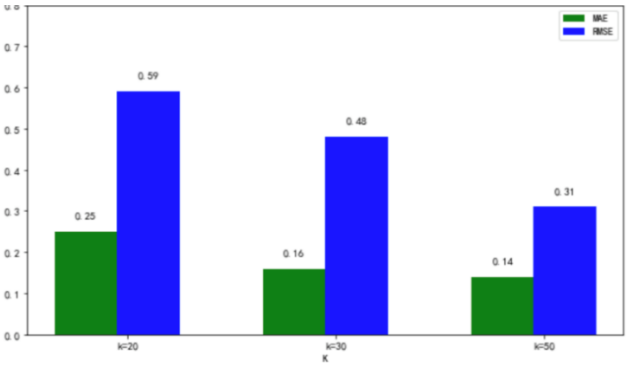

Figure 4. Evaluation index of different classification numbers (K)

Through the comparison shown in Fig.4 , the average absolute error and root mean square error of $k = 20, k = 30$, and $k = 50$ are provided. The lower MAE and RMSE values equal 0.14 and 0.31 respectively, of course, the better, hence, when k=50 is the smallest error. This also corresponds to the best solution for the Recall value. To classify the predicted movie into 50 types, that is, $k = 50$, and then make predictions. Such a prediction is the most accurate classification result. Finally, film number 1 movie "God and Wonders: The End of the World" is taken as an example. The recommended results are finally displayed.

Conclusions:

To be a good content-based suggestion system, the relationship between each dataset field parameter and the content of the dataset should be known. In the experiment, it is also known that the number of recommended system classifications affects the classification decision. Finally, the best

classification number to make the final proposal formula is given in this paper. In the future of society, much of the manual work will be replaced by automated machines. The ticket system can use a fully automated approach. The proposed system will also be a trend in the future, and in the current state of selective diversification, how important is it for consumers to find the product that works best for them. In the future, the proposed system will be able to connect with robots and add a Chatbot system that users can make queries faster and easier.

Acknowledgments

The authors would like to extend their appreciation to the Feng Chia University, Ministry of Higher Education (MOHE), and Universiti Tun Hussein Onn Malaysia (UTHM). This research work is supported by the Ministry of Education, R.O.C, under the grants of TEEP@AsiaPlus and the Ministry of Science and Technology under Grant No. MOST 109-2221-E-035-063-MY2. Moreover, it is also supported by the Fundamental Research Grant Scheme (FRGS) Vot K208. The authors thank the anonymous viewers for the feedback.

Authors' declaration:

- Conflicts of Interest: This research work is supported by the Ministry of Education, R.O.C, under the grants of TEEP@AsiaPlus and the Ministry of Science and Technology under Grant No. MOST 109-2221-E-035-063-MY2.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in Feng Chia University.

References:

1. Ricci F, Rokach L, Shapira B. Introduction to recommender systems handbook. In Recommender systems handbook 2011 (pp. 1-35). Springer, Boston, MA.
2. Pal A, Parhi P, Aggarwal M. An improved content-based collaborative filtering algorithm for movie recommendations. In2017 tenth international conference on contemporary computing (IC3) 2017 Aug 10 (pp. 1-3). IEEE.
3. Chen AY, McLeod D. Collaborative filtering for information recommendation systems. In Encyclopedia of E-Commerce, E-Government, and Mobile Commerce 2006 (pp. 118-123). IGI Global.
4. Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. Advances in artificial intelligence. 2009;2009.
5. Aiolli F. A Preliminary Study on a Recommender System for the Million Songs Dataset Challenge. In IIR 2013 Jan 16 (pp. 73-83).
6. Halder S, Sarkar AJ, Lee YK. Movie recommendation system based on movie swarm. In 2012 Second International Conference on Cloud and Green Computing 2012 Nov 1 (pp. 804-809). IEEE.
7. Kbaier ME, Masri H, Krichen S. A personalized hybrid tourism recommender system. In2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA) 2017 Oct 1 (pp. 244-250). IEEE.
8. Bobadilla J, Bojorque R, Esteban AH, Hurtado R. Recommender systems clustering using Bayesian nonnegative matrix factorization. IEEE Access. 2017 Dec 29;6:3549-64.
9. Neamah AA, El-Ameer AS. Design and Evaluation of a Course Recommender System Using Content-Based Approach. In 2018 International Conference on Advanced Science and Engineering (ICOASE) 2018 Oct 9 (pp. 1-6). IEEE.
10. Baltrunas L, Ludwig B, Ricci F. Matrix factorization techniques for context-aware recommendation. InProceedings of the fifth ACM conference on Recommender systems 2011 Oct 23 (pp. 301-304).
11. Zhang R, Mao Y. Movie Recommendation via Markovian Factorization of Matrix Processes. IEEE Access. 2019 Jan 11;7:13189-9
12. Walek B, Spackova P. Content-based recommender system for online stores using expert system. In 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) 2018 Sep 26 (pp. 164-165). IEEE.
13. Danil B, Elena Y, Ekaterina P. Similarity Measures and Models for Movie Series In Recommender System. International Conference on Internet Science 2018 Oct 24 (pp. 181-193). Springer, Cham.
14. Li J, Xu W, Wan W, Sun J. Movie recommendation based on bridging movie feature and user interest. J Comput Sci. 2018 May 1;26:128-34.
15. Desrosiers C, Karypis G. A comprehensive survey of neighborhood-based recommendation methods. In Recommender systems handbook 2011 (pp. 107-144). Springer, Boston, MA.
16. Lin PC, Arbaiy N. A Novel Classifier for a Kansei Recommender System. In2018 IEEE International Conference on Cognitive Computing (ICCC) 2018 Jul 2 (pp. 114-117). IEEE.
17. Lin PC, Arbaiy N. An Algorithm Design of Kansei Recommender System. InInternational Conference on Soft Computing and Data Mining 2018 Feb 6 (pp. 115-123). Springer, Cham.

التفضيل الموجه للمستخدم تجاه نظام التوصية

بي تشون لين نوريز أربيي

[1]قسم هندسة المعلومات وعلوم الكمبيوتر ، جامعة فنغ شيا ، رقم 100 ، طريق وينهوا ، سيتوين ، تايشونغ 40724 ، تايوان
[2]كلية علوم الكمبيوتر وتكنولوجيا المعلومات ، جامعة تون حسين أون ماليزيا ، 86400 باتو باهات ، جوهور ، ماليزيا

الخلاصة:

في الوقت الحاضر، من الملائم لنا استخدام محرك بحث للحصول على المعلومات المطلوبة. لكن في بعض الأحيان يسيء فهم المعلومات بسبب التقارير الإعلامية المختلفة. نظام التوصية (RS) شائع الاستخدام في كل الأعمال لأنه يمكن أن يوفر معلومات للمستخدمين التي ستجذب المزيد من الإيرادات للشركات. ولكن أيضًا ، في بعض الأحيان ، يوصي النظام المستخدمين بالمعلومات غير الضرورية. لهذا السبب ، قدم هذا البحث بنية لنظام التوصية التي يمكن أن تستند إلى التفضيل الموجه للمستخدم. هذا النظام يسمى .UOP-RS لجعل-UOP RSبشكل كبير، ركزهذا البحث على معلومات السينما وتجميع قاعدة بيانات الأفلام من موقع IMDb الذي يوفر معلومات متعلقة بالأفلام والبرامج التلفزيونية ومقاطع الفيديو المنزلية وألعاب الفيديو والمحتوى المتدفق الذي يجمع أيضًا العديد من التقييمات والمراجعات من المستخدمين. حلل البحث أيضًا بيانات المستخدم الفردي لاستخراج ميزات المستخدم. بناءً على خصائص المستخدم ، وتقييمات / درجات الفيلم ، ونتائج الأفلام ، تم بناء نموذج .UOP-RS في تجربتنا ، تم استخدام 5000 مجموعة بيانات أفلام IMDb و 5 أفلام موصى بها للمستخدمين. تظهر النتائج أن النظام يمكنه إرجاع النتائج في 3.86 ثانية ولديه خطأ 14٪ على السلع الموصى بها عند تدريب البيانات على أنها .K = 50في نهاية هذه الورقة خلص إلى أن النظام يمكن أن يوصي بسرعة مستخدمي السلع التي يحتاجون إليها. سوف يمتد النظام المقترح للاتصال بنظام Chatbot بحيث يمكن للمستخدمين جعل الاستعلامات أسرع وأسهل من هواتفهم في المستقبل.

الكلمات المفتاحية: تحليل معامل الارتباط ، خوارزمية-K أقرب جيران (KNN) ، نظام التوصية ، تحليل الانحدار.