Amalgamating Social Media Data and Movie Recommendation

Maria R. Lee[1(✉)], Tsung Teng Chen[2], and Ying Shun Cai[1]

[1] Department of Information Technology and Management,
Shih Chien University, Taipei, Taiwan
Maria.lee@g2.usc.edu.tw
[2] Graduate Institute of Information Management,
National Taipei University, New Taipei City, Taiwan
timchen.ntpu@msa.hinet.net

Abstract. Recommender systems (RSs) have become very common recently. However, RS techniques need large amounts of user and product data, which hinders RS usage for businesses with insufficient data. The RS cold-start problem may be mitigated by leveraging external data sources. We demonstrate the feasibility of solving the cold-start problem by implementing a hybrid RS that integrates the Facebook Fan Page data and the genre-classifications data from Yahoo! Movies. Our study amalgamates social media data and machine learning to build a hybrid-filtering RS. We also compared our system with three existing movie RSs—those used by Netflix, YouTube, and Amazon. Within the framework of a hybrid-filtering RS, content-based filtering was used to extract data from Yahoo! Movies and Facebook Fan Pages. The proposed RS overcame the cold-start problem and achieved a satisfactory level of accuracy.

Keywords: Recommender system · Social media data · Machine learning · Netflix · Youtube · Amazon

1 Introduction

With the increasing ubiquity of the Internet, social networking sites (SNSs) have also become increasingly sophisticated (Sarwar et al. 2000; Schafer et al. 2001; Lee et al. 2002; Kaplan and Haenlein 2010; Liu et al. 2013; Lu et al. 2014). Moreover, Facebook provides an application program interface (API) for retrieving user and Fan Pages data. Therefore, there is a promising opportunity for value creation if these data can be utilized by businesses and be collated with industry knowledge and analysis.

Recommender systems (RSs) play an indispensable role in e-commerce and increase the profits of businesses intending to sell merchandise (Hirji 1999; Jannach et al. 2010; Bobadilla et al. 2013). However, consumers demand different products, and users who are not familiar with the product will often make poor purchasing decisions. Hence, businesses need to determine the consumers' preferences and recommend suitable products.

The most commonly applied RSs in the mainstream use collaborative filtering and content-based filtering (Goldberg et al. 1992; Burke 2002; Ahn 2008; Al-Shamri and

Bharadwaj 2008). However, the data for both collaborative filtering and content-based filtering are obtained from the businesses' internal members and product databases (Said et al. 2011). Hence, because of the insufficient volume of product types and member data, the implementation of the aforementioned RSs proves to be challenging for small and medium-sized enterprises (SMEs) or creative industries, and the accuracy of the recommendations is below expectations. In view of this problem, this study combined content-based filtering and collaborative filtering to develop a movie RS which incorporates data from SNSs.

Currently, three major film and television service providers are available—Netflix, YouTube, and Amazon—all of which employ RSs (Szomszor et al. 2007; Debnath et al. 2008; Lekakos and Caravelas 2008, Korenl et al. 2009; McSherry and Mironov 2009; Adomavicius et al. 2010). This is a key service offered by these firms to provide more precise and individualized recommendations for each user. Such a service is helpful to users and may spark their interest, which brings in more profit for the businesses. However, the types of content provided by the three companies differ from each other. Netflix is mainly a video-streaming service for TV series or movies; YouTube is a video-sharing website that allows streaming of short video clips; Amazon is an e-commerce website that supplies digital goods. This article will explore the three major movie RSs and discuss their pros and cons. Subsequently, we will introduce our proposed system, and compare the differences between the four systems.

Creating an RS for SMEs that can fulfill client demands and give accurate recommendations will certainly enhance their competitiveness. Therefore, we developed an RS that does not require building an internal membership database. We demonstrate the feasibility of implementing a hybrid RS that integrates the Facebook Fan Page data and the genre-classifications data from Yahoo! Movies.

RSs for the premiere movies still employ traditional methods, which use texts or videos in Taiwan marketing practices. Before deciding whether to watch a movie, one must spend considerable time searching related websites (e.g., Yahoo! Movies, PTT Movie Board, and @ Movies) for reviews posted by fellow users. However, each person has their own subjective perception, and reviews on the same movie could be very different, which is not very helpful for the user. Therefore, building an SNS-based movie RS could facilitate public decision making, thereby reducing the time spent searching for movie information and increasing the satisfaction of movie viewers (Cai and Lee 2015). This study proposed an SNS-based, machine-learning, hybrid-filtering RS. Within the framework of a hybrid-filtering RS, content-based filtering was applied to extract data from Yahoo! Movies and Facebook Fan Pages.

2   Literature Review

Facebook includes three main features: the Facebook Wall on the personal page, Groups, and Fan Pages (Lu et al. 2014). Fan Pages are completely open to the public with no limit on the number of users who can join, and users are free to choose whether to join a Fan Page. Facebook also provides the corresponding API for users and Fan Page data; thus, allowing third parties to develop Facebook-based application services while also providing developers and researchers with the means to retrieve Facebook

data. This has led to the blossoming of Facebook third-party applications and the increasingly widespread application of value-added services, thereby consistently boosting the number of users and enabling Facebook to become the world's fastest-growing website.

An RS's main aim is to provide recommendations to users. It is an intelligent network system that assists users to search for items that they are interested in or that fulfill their desires' work, from among the wealth of products, services, and information available on the Internet. An RSs survey conducted by Bobadilla (2013) shows that recent research has been predominantly content-based, collaborative, demographic, and hybrid filtered. However, because demographic filtering is computed using general statistical methods and, hence, can be directly integrated with the content-based and collaborative filtering RSs, it is not discussed in this article.

A content-based filtering RS employs information-extraction and information-filtering techniques. It is a text-based system that uses keywords or tags to filter meaningful product descriptions, and refers to the users' browsing or purchase histories to establish related features between the users and the products. Machine-learning models are needed to identify and learn about users' interests or behavior patterns. Recommendations are further based on the degree of similarity with the training model (Balabanović and Shoham 1997).

Collaborative-filtering RSs process a vast amount of data through e-mail classification. Collaborative filtering mainly involves using the ratings on products and items from similar users within the same group as the basis for providing recommendations to other members (Sarwar et al. 2001; Davidson et al. 2010).

Hybrid filtering was primarily derived from the need to resolve the disadvantages of individual RSs. It uses the advantages of two or even three techniques to compensate for the limitations of the original techniques; the most common combination is collaborative and content-based filtering. This combination enhances the accuracy and solves the cold-start problem caused by new users and new products in content-based and collaborative filtering.

In terms of the recommended products, Netflix mainly recommends the online streaming of TV series or movies; YouTube recommends its platform for users to upload short video clips; and Amazon recommends multiple categories of digital goods (Zimmerman et al. 2004; Amatriain and Basilico 2012). For the recommendation techniques, Netflix uses a hybrid recommender system (collaborative filtering and content-based filtering) (Gower 2014); YouTube utilizes content-based filtering and computes the similarity of each clip (Davidson et al. 2010); and Amazon employs collaborative filtering and uses its products as the basis for recommendation (Linden et al. 2003). All three systems must establish their own member system before they can provide recommendations and are unable to provide accurate recommendations in the beginning, due to the lack of historical datasets on new users and products.

Computation Time. Although Netflix uses rapid matrix operations, the vast data volume that has grown over time implies that the computation will need to be replaced by parallel or distributed computing. YouTube calculates the similarity between each video clip, which requires pre-processing to provide rapid recommendations when users are using the service. Amazon uses cosine similarity combined with their

algorithms to significantly reduce the computation time. All three systems have problems in terms of data sparsity, and their quality depends on the historical datasets, but they do not require a knowledge of the domain to obtain the recommendation results.

Complexity. Netflix is simple, as it converts the products and users into matrices. YouTube uses the similarity computation of video clips uploaded by users within the last 24 h, and those clicked by users. Due to the sheer number of users on YouTube, the similarity computation of each clip is extremely complicated. Since Amazon has an extensive products portfolio, the data required by the RS are solely based on their own database for the computation of product similarity, which has been automated and is relatively simple. Table 1 shows a summary of the three recommendation systems.

Table 1. Comparison of recommending techniques

	Netflix	Youtube	Amazon
Recommended item	Video & Movie streaming	User video	Digital items (Movie & Music..)
Recommended techniques	A hybrid recommender system	Content-based filtering	Hybrid-based filtering (product-oriented)
Need membership	Yes	Yes	Yes
New user problem?	Yes	Yes	Yes
Computation time	Slightly fast	Slow	Fast
Data sparsity	Depends on the historical datasets	Depends on the historical datasets	Depends on the historical datasets
Domain knowledge required	No	No	No
Complexity	Simple	Similarity computation required	Simple

3   Research Method

This study combined external data from Facebook and Yahoo! Movies to create an RS, using premiere movies in Taiwan as the recommended items, and employing PHP to extract data. Subsequently, in data pre-processing, R was used to convert the data into a format compatible with collaborative-filtering techniques. Thus, a model was created for each movie genre using the collaborative-filtering technique in the R language environment to construct a hybrid-filtering RS (Ihaka and Gentleman 1996).

Once the prototype was completed, historical data were introduced into the system for cross-validation, to evaluate the accuracy level, as well as to adjust the parameters of the system; e.g., the number of Fan Pages, screening for historical data testing, dimensions to be reduced in singular-value decomposition (SVD), and selection of regression models. After adjustments were made, the accuracy level showed some improvements and the prototype could be used to conduct experiments. Thereafter, an RS was created using PHP and R, based on the API provided by Facebook, followed by a survey questionnaire. After processing and analyzing the returned questionnaires, the data were used to verify the effectiveness of the RS, as well as evaluate the differences between past well-known RSs and our proposed RS.

This study proposed an SNS-based, machine-learning, hybrid-filtering RS (Fig. 1). Within the framework of a hybrid-filtering RS, content-based filtering was applied for data extraction, including data crawling of Yahoo! Movies and Facebook Fan Pages, as well as data extraction of user information.
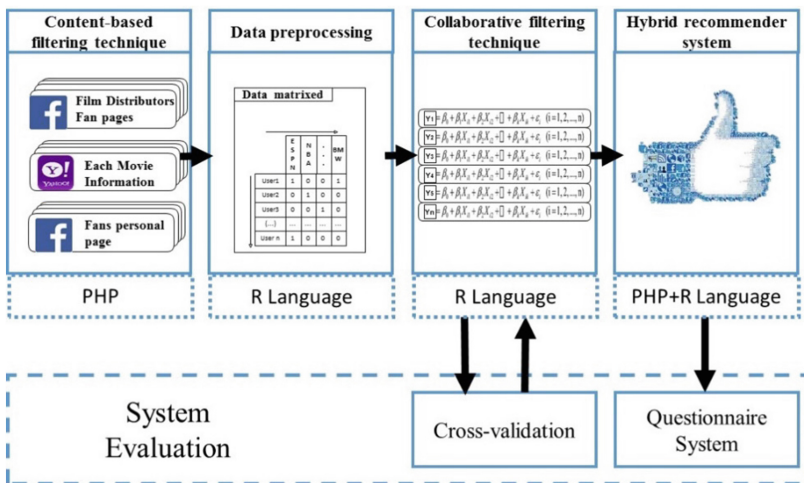

Fig. 1. Proposed hybrid movie recommender system

At the pre-processing stage, data conversion involved transforming data into matrix format, which enabled processing for collaborative filtering (Koren et al. 2009). Fan Pages were pre-sorted and used as related factors during the determination of recommendations. SVD was applied for dimension reduction to enable more-efficient computations. Finally, the data were imported into the collaborative filtering RS, and a machine-learning-based multinomial logistic regression model was constructed. The accuracy of the model was evaluated using the test dataset, and served as a reference for subsequent research.

4   Proposed System Demonstration

Figure 2 shows the proposed system architecture. Content-based filtering RSs stem mainly from two fields: information extraction and information filtering. This study applied the same principles: first, extracting individual movie attributes from the Facebook Fan Pages, Facebook personal pages, and Yahoo! Movies, followed by information filtering utilizing posts on Facebook Fan Pages. The primary aim of this study is to construct an RS for premiere movies, which links the movies with the content of Facebook Fan Page posts, thereby identifying which movies were "Liked" by each Facebook user.
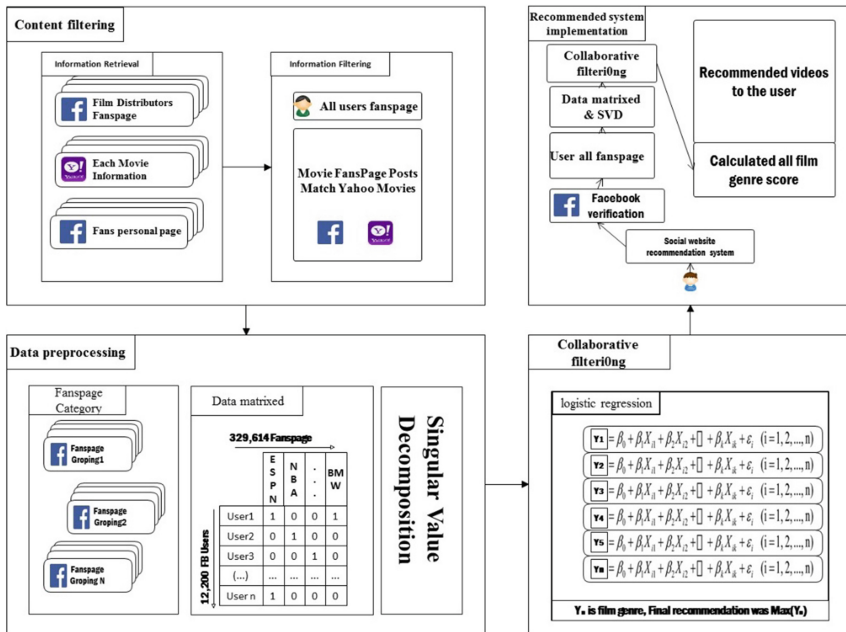

Fig. 2. Proposed system demonstration architecture

A Fan Page post might mention the movie title keywords; however, some posts might not be an exact match for the movie titles in Yahoo! Movies, because of differences in the input methods, e.g., differences in full-width and half-width forms in Chinese characters, and differences in punctuation. To enhance the accuracy level, punctuation was removed from each post, and then a PHP function was applied to implement string matching.

If a movie match was found after the program execution, the movie title and the ID of the Fan Page post were stored in the database. If no matches were found, the search continued sequentially until the end. The collated data from Facebook Fan Pages and Yahoo! Movies are shown in Table 2.

Table 2. Collated data from Facebook fan pages and Yahoo! Movies (Cai and Lee 2015)

ID	Film Name	Film Genre	Posts	Likes	ID	Film Name	Film Genre	Posts	Likes
1	Annabelle	Mystery	116	151,832	12	But Always	Love	69	24,512
2	The Maze Runner	Mystery	118	328,522	13	Turning Tide	Drama	36	7,001
3	The Frogville	Comedy	223	34,432	14	Blind	Drama	71	35,725
4	This Is Where I Leave You	Drama	93	45,537	15	Words and Pictures	Love	44	2,906
5	Paradise In Service	Drama	220	124,142	16	Deliver Us from Evil	Mystery	72	105,850
6	Lucy	Action	135	473,283	17	The Expendables 3	Action	155	59,850
7	The World of Kanako	Mystery	44	46,081	18	Before I Go to Sleep	Mystery	139	45,352
8	As Above, So Below	Terror	55	67,312	19	Chef	Inspiration	165	37,464
9	The Boxtrolls	Animation	60	59,994	20	The Purge: Anarchy	Mystery	35	35,534
10	The Sacrament	Drama	26	21,579	21	Partners in Crime	Mystery	237	144,450
11	Sex Tape	Comedy	83	74,305	22	Wood Job!	Action	106	6,446

The main aims of the data pre-processing stage include: (1) grouping the Fan Pages to explain the influential factors after analyzing the recommendation results; (2) data conversion to matrix format so that the data can be read and processed by the collaborative-filtering model; and (3) data-dimension reduction. Up to 329,614 items were collected from Facebook Fan Pages, which involved 12,200 users. Thus, to optimize the computation efficiency while also preventing the loss of data, SVD was applied to reduce the dimensionality and to extract the eigenvalues of the original data.

Collaborative filtering involves searching whether past users have eigenvalues similar to the recipient of recommendations, and is based on the fact that similar users might share similar preferences. As for the collaborative-filtering component, this study used a multinomial logistic regression technique for investigation. This is a supervised machine-learning algorithm that requires the input of eigenvalues and target values for training. The input values were from the user Fan Page matrix obtained after SVD processing, and the target values were from whether a Facebook user Liked a particular movie. If yes, the target value was denoted by 1, and if not, it was denoted by 0. The parameter values of each movie could be obtained after training. In the multinomial regression of collaborative filtering, a linear regression model was established, whereby the dependent variable **y** was the movie genre, which included suspense/thriller, comedy, action, drama, horror, animation, romance, adventure, and inspirational.

The target values for these nine items depended on whether a user was interested in this genre, denoted by 1 if interested, and 0 if not. Training was then conducted through a machine-learning model. The equation model is as follows:

$$Y_i = W_{g0} + W_{g1}X_{ug1} + W_{g2}X_{ug2} + \ldots\ldots + W_{g99}X_{ug99} + W_{g100}X_{ug100} \quad (1)$$

Y is the movie genre; W_{g0}: the constant value; W_{g1} to W_{g100}: movie genre independent variables (weights); X_{ug1} to X_{ug100}: user Fan Page independent variables (parameter).

The algorithm amalgamates social media data and machine learning to create a hybrid-filtering RS. SVD was used to reduce the dimensionality, and multi-logistic regression was used to train and establish the *Film genre* module.

|	
1	**Algorithm for movie recommender**
2	**For each** *Post in Fanspage* **in** *Film Genre*
3	**For each** *Post Like user* **in** *Post in Fanspage*
4	collect *Post Like user 's Like Fanspage*
5	set the *Post Like user - Fanspage* Matrix, M_1
6	use Singular Value Decomposition to dimensionality reduction M_1
7	get I_1
8	
9	**For each I_1 in** *Film genre*
10	Use Muti-logistic regression training and establishing of
11	*Film genre module.*
12	
13	**IF** new user to be recommended
14	**Then**
15	collect *Post Like new user 's Like Fanspage*
16	set the *Post Like new user - Fanspage* Matrix, M_2
17	use Singular Value Decomposition to dimensionality reduction M_2,
18	get I_2
19	
20	**For each** *Film genre module*
21	put I_2 to *Film Genre module*
22	get the each *Film genre score*
23	
24	**For each** *Film genre score*
25	**IF** *score* is the highest
26	**Then**
	Recommend that *Film genre*

5 System Evaluation

The system evaluation of our proposed RS was divided into two parts. The first part was cross-validation to investigate the accuracy of the logistic regression model for each movie genre. K = 10 indicated that 90 % of the overall fan population was used as the training dataset, while the remaining 10 % was used as the test sample. The accuracy for each genre is shown in Fig. 3. After validation, the overall mean judgment accuracy was more than 65 %, but lower than 50 % for the suspense/thriller genre.
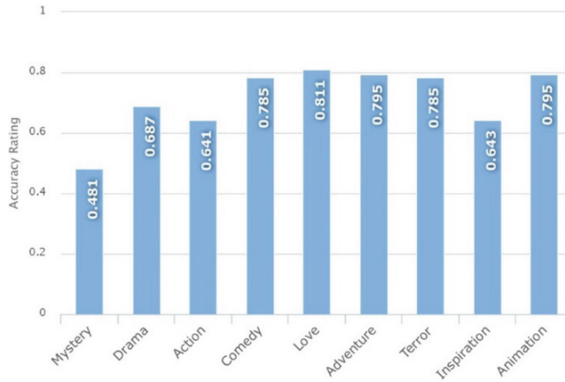
**Fig. 3.** Results of cross validation evaluation (Cai and Lee 2015)

The second part was the online system questionnaire, and our research subjects were selected from the Facebook user population. The online system-questionnaire evaluation process is shown in Fig. 4. During the testing period, eight premiere movies were recommended, each in the 50th, 51st, and 52nd weeks, giving recommendations for a total of 24 movies. The movie genres included 10 dramas, 4 actions, 3 comedies, 3 horrors, 2 romances, 1 adventure, and 1 animation.
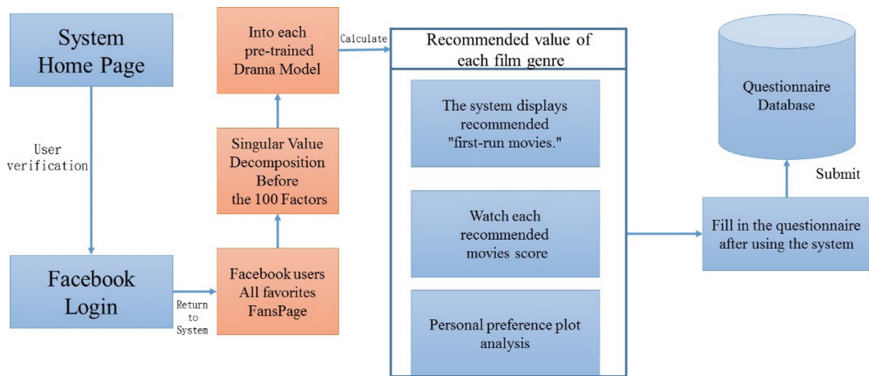
Fig. 4. On-line system questionnaire evaluation process

The questionnaire asked users to rate the individual items recommended by our RS and their overall perception. For all experiences, the maximum score was 10 points, and the minimum was 1 point. The overall mean score for all users was 8.52 points.

The overall results of the online system questionnaire evaluation are shown in Fig. 5. Forty-eight out of 102 people gave 10 points (highest score), 13 people gave 9 points, 15 people gave 8 points, 12 people gave 7 points, 5 people gave 6 points, 6 people gave 5 points, 2 people gave 3 points, and 1 person gave 2 points. No one gave the lowest score (1 point).
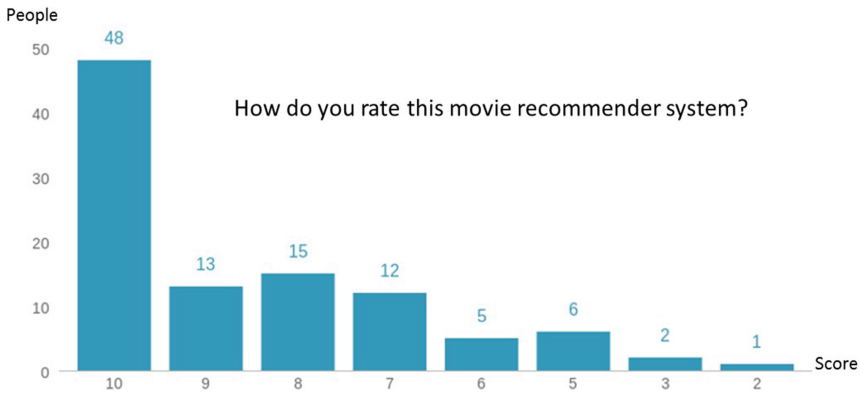
Fig. 5. Results of on-line system questionnaire evaluation

6 Discussion and Conclusion

With regard to the RS techniques applied in our proposed system, content-based and collaborative filtering were combined to form a hybrid-filtering RS. No initial member database is required since we can utilize the membership information from Facebook. To avoid the problems caused by new users, the user's Fan Page combined with SVD could be used to compute the scores of the recommended products, and provide recommendations for the users.

The creation of Facebook Fan Pages is required for all products, new and old. The computation time only involves performing SVD on all Likes by Fan Page users. Due to the data-sparsity problem, only Fan Pages with sufficient Likes by users will provide accurate predictions. Our RS is extremely dependent on the knowledge-classification methods of each domain. Only good classifications will allow us to achieve a more comprehensive model for training, and to reach a certain level of accuracy. Sufficient background knowledge and techniques are needed because of complexity; e.g., using the Facebook data concatenation method and SVD, as well as the application of logistic regression to construct this SNS-based hybrid RS.

This study developed a movie RS combined with SNS data, and used the Facebook historical dataset for validation. We also compared the three major movie RSs to address the accuracy issues caused by the cold-start problem; thus, enabling our RS to achieve a satisfactory level of accuracy immediately after it was introduced online, while also solving the new-user problem. Furthermore, our RS also achieved rapid computation of the customized recommendation items.

Text analysis was involved in the recommendation process, and was based mainly on traditional Chinese; hence, experiments and investigations in other languages were not performed. Since premiere movies are limited by the duration of their release in cinemas, the products were renewed every week; hence, movies by genre were used to supplement the recommendation mechanism in this study.

Facebook Graph API version 2.5 was used to construct the RS in this study. If Facebook updates or terminates the provision of its user information, e.g., user ID,

all Fan Pages Liked by users, Fan Page information, and the number of Likes on Fan Pages, our method will be unable to run. Furthermore, if users using this RS do not have a Facebook account or if the number of Fan Pages Liked is less than 300, this RS is not appropriate.

In terms of the machine-learning algorithm, future research should consider using multidimensional algorithms, such as neural networks or fuzzy logic, to replace multinomial logistic regression to compare the different algorithms and make modifications to enhance the accuracy.

References

Adomavicius, G., Tuzhilin, A., Berkovsky, S., De Luca, E.W., Said, A.: Context-aware recommender systems: research workshop and movie recommendation challenge. In: ACM RecSys2010, pp. 26–30 (2010)

Ahn, H.J.: A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Inf. Sci. **178**(1), 37–51 (2008)

Al-Shamri, M.Y.H., Bharadwaj, K.K.: Fuzzy-genetic approach to recommender systems based on a novel hybrid user model. Expert Syst. Appl. **35**(3), 1386–1399 (2008)

Amatriain, X., Basilico, J.: Netflix recommendations: beyond the 5 stars (part 1). Netflix Tech Blog (2012)

Balabanović, M., Shoham, Y.: Fab: content-based, collaborative recommendation. Commun. ACM **40**(3), 66–72 (1997)

Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowl.-Based Syst. **46**, 109–132 (2013)

Burke, R.: Hybrid recommender systems: Survey and experiments. User Model. User-Adap. Inter. **12**(4), 331–370 (2002)

Cai, Y.S., Lee, M.R.: Research on social network film genre recommender system. In: The 21th Cross Strait Conference on Information Management development and Strategy, Macao, China (2015)

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B.: The YouTube video recommendation system. In: Proceedings of the Fourth ACM Conference on Recommender systems, pp. 293–296. ACM, New York (2010)

Debnath, S., Ganguly, N., Mitra, P.: Feature weighting in content based recommendation system using social network analysis. In: Proceedings of the 17th International Conference on World Wide Web, pp. 1041–1042. ACM, New York (2008)

Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Commun. ACM **35**(12), 61–70 (1992)

Gower, S.: Netflix Prize and SVD (2014). http://buzzard.ups.edu/courses/2014spring/420projects/math420-UPS-spring-2014-gower-netflix-SVD.pdf. Accessed 8 Jan 2016

Hirji, K.K.: Discovering data mining: From concept to implementation. ACM SIGKDD Explor. Newsl. **1**(1), 44–45 (1999)

Ihaka, R., Gentleman, R.: R: a language for data analysis and graphics. J. Comput. Graph. Stat. **5**(3), 299–314 (1996)

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. Cambridge University Press, New York (2010)

Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horiz. **53**(1), 59–68 (2010)

Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)

Lee, W.-P., Liu, C.-H., Lu, C.-C.: Intelligent agent-based systems for personalized recommendations in Internet commerce. Expert Syst. Appl. **22**(4), 275–284 (2002)

Lekakos, G., Caravelas, P.: A hybrid approach for movie recommendation. Multimedia Tools Appl. **36**(1–2), 55–70 (2008)

Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Comput. **7**(1), 76–80 (2003)

Liu, N.N., He, L., Zhao, M.: Social temporal collaborative ranking for context aware movie recommendation. ACM Trans. Intell. Syst. Technol. (TIST) **4**(1), 15 (2013)

Lu, Y., Wang, F., Maciejewski, R.: Business intelligence from social media: A study from the vast box office challenge. IEEE Comput. Graph. Appl. **34**(5), 58–69 (2014)

McSherry, F., Mironov, I.: Differentially private recommender systems: building privacy into the net. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 627–636. ACM, New York (2009)

Said, A., Berkovsky, S., De Luca, E.W., Hermanns, J.: Challenge on context-aware movie recommendation: CAMRa2011. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 385–386. ACM, New York (2011)

Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of the 2nd ACM Conference on Electronic Commerce. ACM (2000)

Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, pp. 285–295. ACM, New York (2001)

Schafer, J.B., Konstan, J.A., Riedl, J.: E-commerce recommendation applications. Applications of Data Mining to Electronic Commerce, pp. 115–153. Springer (2001)

Szomszor, M., Cattuto, C., Alani, H., O'Hara, K., Baldassarri, A., Loreto, V., Servedio, V.D.: Folksonomies, the semantic web, and movie recommendation. In: 4th European Semantic Web Conference, Bridging the Gap between Semantic Web and Web 2.0, Innsbruck, Austria, pp. 1–14 (2007)

Zimmerman, J., Kauapati, K., Buczak, A.L., Schaffer, D., Gutta, S., Martino, J.: TV personalization system. In: Personalized Digital Television, pp. 27–51. Springer (2004)