Computer Science and Information Systems 11(1):241–249 DOI: 10.2298/CSIS130226006L

Black-Box Testing of Practical Movie Recommendation Systems: a Comparative Study*

Namhee Lee[1], Jason J. Jung[2,4], Ali Selamat[3], and Dosam Hwang[2]

[1] School of Business Administration
Sogang University, Seoul, Korea
namhee.lee80@gmail.com
[2] Department of Computer Engineering
Yeungnam University, Gyeongsan, Korea
{j2jung, dshwang}@ynu.ac.kr
[3] Software Engineering Research Group (SERG)
Knowledge Economy Research Alliance and Faculty of Computing
Universiti Teknologi Malaysia
81310 Johor, Malaysia
aselamat@utm.my
[4] Universiti Malaya
Kuala Lumpur, Malaysia

Abstract. Many practical recommendation systems have been studied, and also the services based on such recommendation systems have been opened in real world. The main research questions of this work are i) how these recommendation services provide users with useful information, and ii) how different the results from the systems are from each other. In this paper, we propose a black-box evaluation framework of the practical recommendation services. Thus, we have designed user modeling process for generating synthesized user models as the inputs for the recommendation services. User models (i.e., a set of user ratings) have been synthesized to discriminate the recommendation results. Given a set of practical recommendation systems, the proposed black-box testing scheme has been applied by comparing recommendation results. Particularly, we focus on investigating whether the services consider attribute selection.

Keywords: Social networks; Recommendation systems; Black-box testing; Comparative study.

1. Introduction

Recommendation systems have been studied for a long time. There have been a lot of recommendation schemes to provide users with the most relevant information. In the real world, we have been able to access recommendation services in various domains. Particularly, movie is the most popular domain targeted by the recommendation services.

However, even though users are eager to get the recommendations from the practical systems, they have not been satisfied with the results. For example, users are consistently

* This paper is significantly revised from an earlier version presented at the 5th International Conference on Computational Collective Intelligence (ICCCI 2013) held in Craiova, Romania in September 2013.

acting with the same preferences (i.e., the same input ratings), the results from such recommendation systems are completely difference from each other [3].

Users (at least system developers) want to know what kinds of recommendation mechanisms are behind the systems. Somehow, depending on their situation, more appropriate system can be selected. Thereby, in this paper, we focus on comparing the recommendation results provided from the practical recommendation systems, and investigating what kinds of recommendation schemes have been exploited in these systems [6]. Especially, we want to show that the proposed *black-box testing strategy* can precisely reveal the recommendation schemes behind those practical recommendation systems [4]. To do this, we have synthesized a number of user models by generating user ratings. Consequently, target systems returns a set of recommendations, and the systems will be differentiated with each other [12].

The outline of this paper is as follows. In the following Sect. 2, we will address backgrounds on recommendation systems in the literature. Sect. 3 gives a research method on comparing the results obtained from a set of selected recommendation systems. Sect. 3.2 will give the description about the recommendation systems that we have selected in this work. Most importantly, Sect. 4 show how to conduct the experiments for collecting the results from the recommendation systems. In Sect. 5, we will analyze the collected recommendation results, and draw a conclusion of this work, respectively.

2. Backgrounds and Related Work

User modeling process in the recommendation systems is commonly based on analyzing various information collected by users in explicit or implicit ways [13]. Depending on the following two issues, we need to consider to build a taxonomy of user modeling in recommendation systems.

1. Recommendation systems usually ask users to explicitly input various information. First issue is what kinds of information is employed to build user models in the recommendation systems.
   - Demographic information
   - User ratings

   Given equivalently synthesized users (e.g., same age, same gender, and so on), we can compare the recommendation results to realize whether they are same or different.
2. Recommendation systems somehow exploit the information collected from the users to discover useful patterns about the users. Second issue is which recommendation strategies are applied to provide users with relevant information.
   - Personalization
   - Collaborative filtering

   Assuming that a user is interested in a certain value (e.g., “Steven Spielberg”), the recommendation results can be compared to reversely find out what kind of schemes are applied.

Hence, as shown in Table 1, the recommendation systems can be simply categorized into four different types. Of course, in practice, we are sure that the practical systems are employing a hybrid scheme [8] combining several recommendation approaches (e.g., collaborative filtering and personalization) [2,10].

Table 1: Taxonomy of recommendation systems w.r.t. user modeling processes

	Personalization	Collaborative filtering
Demographics	$\text{Rec}_{\text{P,D}}$	$\text{Rec}_{\text{C,D}}$
User ratings	$\text{Rec}_{\text{P,U}}$	$\text{Rec}_{\text{C,U}}$

Especially, in the context of attribute selection-based user modeling [9], the users can be synthesized to differentiate the results from recommendation systems. Table 2 shows an example with three users (i.e., U_1, U_2, and U_3). With two users (i.e., U_1 and U_2), we can discriminate $\text{Rec}_{\text{P,D}}$ and $\text{Rec}_{\text{P,U}}$ from $\text{Rec}_{\text{C,D}}$ and $\text{Rec}_{\text{C,U}}$ Also, with two users (i.e., U_1 and U_3), we can discriminate $\text{Rec}_{\text{P,D}}$ and $\text{Rec}_{\text{C,D}}$ from $\text{Rec}_{\text{P,U}}$ and $\text{Rec}_{\text{C,U}}$.

Table 2: Example with three synthesized users who have rated the same movies

		U_1	U_2	U_3
Age/Gender/Country		20/Male/Korea	20/Male/Korea	60/Female/France
Rating	Lincoln	5	1	Very good
	War Horse	5	1	Very good
	The Terminal	5	1	Very good

3. Research method

In this paper, we focus on black-box testing scheme [1], since we have no information about internal strategies of the practical recommendation systems [7]. As shown in Fig. 1, a set of users are synthesized to collect recommendations from the practical services.

3.1. Comparison of recommendation results

Once we select a set of recommendation service, recommendation results are compared with each other.

Definition 1 (Recommendation). *Given a recommendation service RS_i, the recommendation result M_i is composed of a set of movies which are regarded as the most relevant movies to user contexts. It is represented as*

$$M_i = \{m_1, m_2, \dots, m_N\} \tag{1}$$

where N is the number of movies recommended by the system.

The recommendation results can be matched to quantify the similarity between the corresponding recommendation schemes.

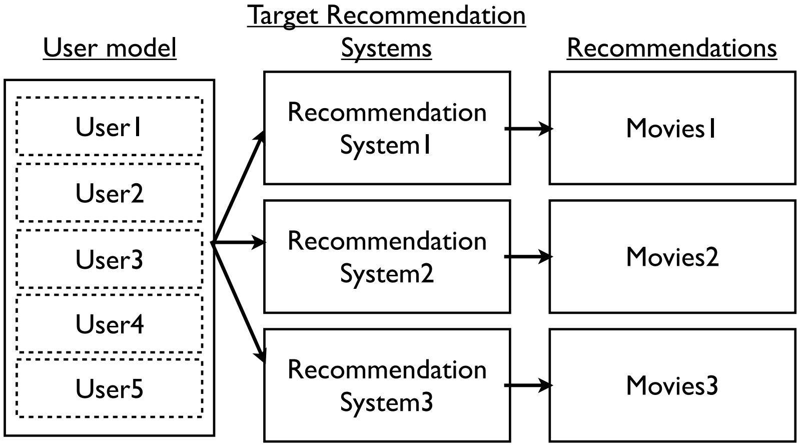

Fig. 1: Research model based on black box testing

Definition 2 (Similarity). *Two recommendation results M_i and M_j from recommendation services RS_i and RS_j, similarity between RS_i and RS_j can be measured by*

$$Sim(RS_i, RS_j) = \frac{M_i \cap M_j}{\max(M_i, M_j)} \tag{2}$$

where denominator can choose the maximum size of recommendation.

3.2. Selected practical recommendation systems

Initially, as shown in Table 3, we have tried to select 10 movie recommendation systems [14]. Out of them, 2 recommendation systems are not available for the moment.

User registration Practical recommendation systems usually ask users to input various personal information during registration, depending on their recommendation schemes. Table 4 shows the list of demographic information requested by the recommendation systems.

All systems ask Email in common. It is regarded as an unique identifier for each user. Since only two kinds of personal information (e.g., Email and username) are asked in Taste Kid and Nanocrowd, these two systems are not asking any personal information. They seem to be more focused on ratings from users. In contrast, Jinni, Criticker, IMDB and Rotten Tomatoes are asking more than five types of personal information.

Representation of user ratings Recommendation systems ask users to rate movies. Depending on the systems, the ratings are represented in several different ways. Table 5 shows how the user ratings are represented in the recommendation systems. Three of the systems (i.e., Flixster, Movielens, and Rotten Tomatoes) are allowing users to rate the

Table 3: Selected recommendation systems

Recommendation systems	URLs	
Jinni	`www.jinni.com`	Not available
Taste Kid	`www.tastekid.com`	
Nanocrowd	`www.nanocrowd.com`	
Clerkdogs	`www.clerksblog.com`	
Criticker	`www.criticker.com`	
IMDB	`www.imdb.com`	
Flixster	`www.flixster.com`	
Movielens	`www.movielens.org`	
Rotten Tomatoes	`www.rottentomatoes.com`	
Netflix	`www.netflix.com`	Not available in Korea

Table 4: User profiling during registration (○: required, △: optional)

Recommendation systems	Email	User name	Real name	Date of Birth	Gender	Country	City/ State	Postal Code	Marital Status
Jinni (RS_1)	○	○	△	△	△	○		△	
Taste Kid (RS_2)	○	○							
Nanocrowd (RS_3)	○	○							
Criticker (RS_4)	○	○	△	△	△	△	△		△
IMDB (RS_5)	○		○	○	○			○	
Flixster (RS_6)	○		○	○					
Movielens (RS_7)	○	○	△					△	
Rotten Tomatoes (RS_8)	○		○	○	○	○			

Table 5: Representation of user ratings

Recommendation systems	Data type	Range	Cardinality
Jinni	ordinal/ discrete	{ awful, bad, poor, disappointing, so so, ok, good, great, amazing, must see }	10
Taste Kid	ordinal/discrete	{like, dislike}	2
Nanocrowd	enumerate/discrete	{watched, not watched}	2
Criticker	numeric/integer/ discrete	{1, 2, . . . , 100 }	100
IMDB	ordinal/discrete	{1, 2, . . . , 10}	10
Flixster	ordinal/discrete	{1, 2, . . . , 5}	5
Movielens	ordinal/discrete	{1, 2, . . . , 5}	5
Rotten Tomatoes	ordinal/discrete	{1, 2, . . . , 5}	5

movies between 1 and 5. Jinni and IMDB are more diversified to between 1 and 10. Particularly, in Criticker, users can rate the movies from 0 to 100. On the other hand, Taste Kid and Nanocrowd are making the user ratings most simplified. Interestingly, Nanocrowd is simply asking users to record the list of movies (i.e., "watched").

Additionally, we have to consider the requirement of initial rating step, as shown in Table 6. Jinni, Criticker, and Movielens are collecting initial ratings from users, before they provide the users with recommendations.

Table 6: Initial requirement for recommendations; X indicates 'Not required'

Recommendation systems	Number of initial ratings	Rating scores
Jinni	More than 10 items	0 to 10
Taste Kid	X	0 to 1
Nanocrowd	X	0 to 1
Criticker	More than 10 items	0 to 100
IMDB	X	0 to 10
Flixster	X	1 to 5
Movielens	More than 15 items	1 to 5
Rotten Tomatoes	X	1 to 5

4. Experiments and evaluation

In this section, we want to describe how to synthesize user models and how to collect recommendation results from the practical services.

User models have been synthesized by assuming a user is interested in a certain attribute, as follows.

- U_1: Genre "Sci-Fi"
- U_2: Director "Steven Spielberg"
- U_3: Actor "Leonardo DiCaprio"
- U_4: Actress "Angelina Jolie"

Also, the movies can be rated in two difference ways.

- Unified rating: We need to express that a user's interest is consistent in each attribute.
- Random rating: We need to express that a user's interest is not consistent in each attribute.

For example, a user is assumed to be consistently interested in a genre "Sci-Fi". As shown in Table 2, this user can be synthesized in $\{\langle m_1, 5\rangle, \langle m_2, 5\rangle, \langle m_3, 5\rangle\}$. In opposite, his random ratings can be synthesized as $\{\langle m_1, 5\rangle, \langle m_2, 1\rangle, \langle m_3, 3\rangle\}$.

Since user ratings are differently represented in recommendation services (shown in Table 5), the user ratings in different recommendation services should be normalized to be comparable.

We have collected the recommendations from the practical services. Table 7 and Table 8 show evaluational results on personalization-based recommendation services. In both cases, only U_1 (Genre "Sci-Fi") has shown high matching ratio. The other three users (i.e., U_2, U_3, and U_4) are in the very low level. We found that the practical recommendation services are not considering attribute-based personalization.

Table 7: Evaluation on attribute-based personalization with uniform ratings

	RS_1	RS_2	RS_3	RS_4	RS_5	RS_6	RS_7	RS_8
U_1	$\frac{17}{27}$	$\frac{14}{17}$	$\frac{42}{75}$	$\frac{0}{8}$	$\frac{51}{84}$	$\frac{27}{50}$	-	$\frac{71}{135}$
U_2	$\frac{0}{26}$	$\frac{3}{17}$	$\frac{3}{75}$	$\frac{0}{8}$	$\frac{2}{60}$	$\frac{4}{40}$	-	$\frac{6}{191}$
U_3	$\frac{0}{27}$	$\frac{0}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{0}{90}$	$\frac{0}{48}$	-	$\frac{0}{171}$
U_4	$\frac{0}{26}$	$\frac{3}{17}$	$\frac{1}{75}$	$\frac{0}{8}$	$\frac{0}{54}$	$\frac{0}{15}$	-	$\frac{0}{103}$

Table 8: Evaluation on attribute-based personalization with random ratings

	RS_1	RS_2	RS_3	RS_4	RS_5	RS_6	RS_7	RS_8
U_1	$\frac{11}{20}$	$\frac{13}{17}$	$\frac{42}{75}$	$\frac{2}{8}$	$\frac{35}{65}$	$\frac{24}{50}$	-	$\frac{53}{92}$
U_2	$\frac{2}{20}$	$\frac{2}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{1}{66}$	$\frac{6}{50}$	-	$\frac{4}{98}$
U_3	$\frac{0}{20}$	$\frac{0}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{0}{66}$	$\frac{0}{11}$	-	$\frac{0}{82}$
U_4	$\frac{0}{20}$	$\frac{0}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{0}{66}$	$\frac{0}{50}$	-	$\frac{0}{78}$

Also, Table 9 shows comparison between uniform ratings and random ratings. RS_2 and RS_8 are showing high ratio between uniform and random ratings. It means that these recommendation services are not considering user ratings as importantly as the other services are.

5. Concluding Remark and future work

In this work, we have investigated what kinds of recommendation schemes have been exploited in the practical recommendation services. As a conclusion, this paper has evaluated user modeling process in several practical recommendation systems. Black-box testing scheme has been applied by comparing recommendation results. User models (i.e., a set of user ratings) have been synthesized to discriminate the recommendation results.

In future work, we are planning to extend our proposal by improving the model to track on multi properties of information propagation pattern other than the max value of coverage rate and its time; clarify further more about the relationship between positive/negative emotional words and efficiency of information propagation on Social Network Service [5]. Besides that, we have to consider to semantic-based user modeling [11] to evaluate the recommendation systems.

Table 9: Comparison of uniform and random ratings

		RS$_1$	RS$_2$	RS$_3$	RS$_4$	RS$_5$	RS$_6$	RS$_7$	RS$_8$
Common	U$_1$	18.5	35.3	-	0.0	39.7	20	-	67.4
	U$_2$	42.3	64.7	-	0.0	3.3	80	-	51.3
	U$_3$	14.8	29.4	-	0.0	14.4	6.3	-	47.9
	U$_4$	11.5	52.9	-	0.0	9.1	10	-	74.7
Similar	U$_1$	63.0	76.5	-	12.5	51.3	54	-	57.6
	U$_2$	10 .0	17.6	-	0.0	39.7	12	-	4.1
	U$_3$	0.0	0.0	-	0.0	0.0	0.0	-	0.0
	U$_4$	0.0	17.6	-	0.0	0.0	0.0	-	0.0

Acknowledgments. This work was supported by the 2013 Yeungnam University Research Grant.

References

1. Beizer, B.: Black-Box Testing: Techniques for Functional Testing of Software and Systems. John Wiley & Sons (1995)
2. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1), 5–53 (January 2004)
3. Hwang, D., Nguyen, N.T., Jung, J.J., Sadeghi-Niaraki, A., Baek, K.H., Han, Y.S.: A semantic wiki framework for reconciling conflict collaborations based on selecting consensus choice. Journal of Universal Computer Science 16(7), 1024–1035 (2010)
4. Jung, J.J.: Knowledge distribution via shared context between blog-based knowledge management systems: a case study of collaborative tagging. Expert Systems with Applications 36(7), 10627–10633 (2009)
5. Jung, J.J.: Reusing ontology mappings for query segmentation and routing in semantic peer-to-peer environment. Information Sciences 180(17), 3248–3257 (2010)
6. Jung, J.J.: Boosting social collaborations based on contextual synchronization: An empirical study. Expert Systems with Applications 38(5), 4809–4815 (2011)
7. Jung, J.J.: Exploiting multi-agent platform for indirect alignment between multilingual ontologies: a case study on tourism business. Expert Systems with Applications 38(5), 5774–5780 (2011)
8. Jung, J.J.: Service chain-based business alliance formation in service-oriented architecture. Expert Systems with Applications 38(3), 2206–2211 (2011)
9. Jung, J.J.: Attribute selection-based recommendation framework for short-head user group: An empirical study by movielens and imdb. Expert Systems with Applications 39(4), 4049–4054 (2012)
10. Jung, J.J.: Collaborative browsing system based on semantic mashup with open apis. Expert Systems with Applications 39(8), 6897–6902 (2012)
11. Jung, J.J.: Computational reputation model based on selecting consensus choices: an empirical study on semantic wiki platform. Expert Systems with Applications 39(10), 9002–9007 (2012)
12. Jung, J.J.: Evolutionary approach for semantic-based query sampling in large-scale information sources. Information Sciences 182(1), 30–39 (2012)
13. Pu, P., Chen, L., Hu, R.: Evaluating recommender systems from the user's perspective: survey of the state of the art. User Modeling and User-Adapted Interaction 22(4-5), 317–355 (2012)
14. Reisinger, D.: Top 10 movie recommendation engines (March 2009), `http://news.cnet.com/8301-17939_109-10200031-2.html`

Namhee Lee is currently a postdoctoral researcher in Sogang University, Korea. She received her B.A. degree in Business Administration from Hongik University, Korea in 2003. She received her M.S. degree in Management Information Systems from Sogang University, Korea in 2007. She earned her Ph.D in Management Information Systems at Sogang Business School, Korea in 2013. Her primary research interests are including Service Science and Management and Information Technology Innovation.

Jason J. Jung is a visiting research professor in Universiti Malaya, Malaysia. He is an associate professor in Yeungnam University in Korea. He has been a lot of industrial experiences on web-based information systems. His main research areas include recommendation systems, mashup services, and social network-based services.

Ali Selamat has received a B.Sc. (Hons.) in IT from Teesside University, U.K. and M.Sc. in Distributed Multimedia Interactive Systems from Lancaster University, U.K. in 1997 and 1998, respectively. He has received a Dr. Eng. degree from Osaka Prefecture University, Japan in 2003. Currently, he is the Dean of Research Alliance in Knowledge Economy (K-Economy RA) UTM. He is a professor at Faculty of Computer Science & IS UTM. Previously he was an IT Manager at School of Graduate Studies (SPS), UTM. He is also a head of Software Engineering Research Group (SERG), K-Economy Research Alliance, UTM. He is the editors of International Journal of Digital Content Technology and its Applications (JDCTA) andInternational Journal of Advancements in Computing Technology (IJACT). He was the conference chair of Asian conference in Intelligent Information and Database Systems (ACIIDS2013) and IEEE Malaysia Software Engineering Conference 2011 (MySEC2011). His research interests include software engineering, software agents, web engineering, information retrievals, pattern recognitions, genetic algorithms, neural networks and soft-computing.

Dosam Hwang is a corresponding author of this paper. He received his Ph.D. degrees in natural language processing at Kyoto University. In 1987, he was chosen the best researcher at Korea Institute of Science and Technology. Since 2005 he has been the Professor at Yeungnam University in Korea. Currently working on natural language processing, machine translation, ontology, semantic web, and information retrieval. He has served for a number of international conferences and a technical committee for ISO.

Received: April 10, 2013; Accepted: October 10, 2013.