Evaluating Functional Requirements-Based Compound Critiquing on Conversational Recommender System

[1]Z.K. Abdurahman Baizal, [2]Yusza Reditya Murti, [3]Adiwijaya
School of Computing
Telkom University
Bandung, Indonesia
[1]baizal@telkomuniversity.ac.id, [2]yuszaa@ telkomuniversity.ac.id, [3]adiwijaya@telkomuniversity.ac.id

***Abstract*— Conversational recommender system (CRS) is a form of recommender system that can refine user preference through conversational mechanism. User preference refinement can be proceeded in reference to the user's feedback towards the products recommended, called as critiquing technique. Compound critiquing technique has been widely developed to ensure the interaction efficiency in CRS. However, the compound critiques offered refer to the product technical features. In terms of hi-tech products, not all consumers are familiar with technical features. A model for generating functional requirement-based compound critiques (instead of technical features-based) has been developed in our previous work. In this paper, we evaluate this model from the aspects of recommendation accuracy, query refinement, and user satisfaction. The user study involving 88 users (either familiar or unfamiliar with technical features) shows that the approach has successfully increased the users' positive perception compared to the recommender system commonly used in e-commerce. Besides, this approach has a high recommendation accuracy (89.77%) and has successfully refined the users' needs.**

Keywords—recommender system; conversational recommender system; product functional requirement; compound critiquing; ontology

I. INTRODUCTION

Recommender system is a system that can help and guide user in finding the products that meet the user preference [1]. Conversational recommender system (CRS) is a type of recommender system, where the system attains the user preference by explicitly and repeatedly giving several questions or asking for user's feedback. The user-system interaction will be over if the user has chosen a product that meets their needs.

Navigation by asking (NBA) and 2) Navigation by proposing (NBP) are navigation strategies used in CRS to refining user preference [2]. In this study, we adopt NBP strategy. In NBP, the system refines user preference based on the user's feedback towards the products recommended by the system. Designing an effective user's feedback and helping users in expressing their preference is one of the main problems in CRS. Critiquing is a form of user's feedback that has been proven to be success and popular in CRS [3]–[5]. Unit critiquing is the simplest form of user's feedback in critiquing where the system gives critiques (feedback) in the form of a single attribute or feature. By using unit critiques, it is possible for user to freely choose several critiques at a time [5], [6]. As a result, unit critiquing may cause the products not to be found as there are some contradictions between user preference (represented through the critiques) and the available products. To make the interaction between the user and system more efficient, it is wise to use compound critiques where the user can only choose a single critique offered. Critiques in compound critiquing are combination of multiple attributes or features that lead to a pattern [7]–[9].

Recently, in CRS, there are two perspectives in finding user preference: 1) CRS that uses questions which refer to product technical specification (e.g. i need smartphone with LTPS IPS LCD 1080 x 1920 pixels) [4], [8], [9] and 2) CRS that uses questions which refers to product functional requirement (e.g. need smartphone for online activities) [10], [11].

Functional requirements-based compound critiquing is one of the approaches in CRS, which uses compound critiquing feedback referring to functional requirement of a product. We have introduced this functional requirements-based compound critiquing in our preliminary research [12]. The approach is aimed at providing an interaction that may help users in interacting with the system, which does not require the user to be familiar with the technical features.

We choose smartphone as the domain of the study. Smartphone is one of the products that has many functional requirements, and many people are not familiar with the technical specification it has. Recently, most of people have smartphone. Even, it can be said that smartphone is one of the urban society's primary needs.

ISBN: 978-1-5090-4911-0 (c) 2017 IEEE

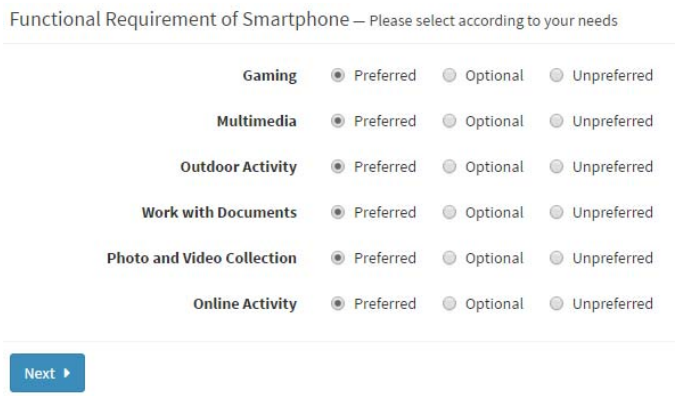


Fig. 1. An Example of Initial Interaction (CR1)

The CRS uses ontology as the knowledge base. The ontological structure has 3 main classes including functional requirement, specification, and product [11]. Data pertaining to the specifications of smartphones are taken from gsmarena.com and phonearena.com which are known for its database for smartphone. However, the functional requirement and mapping between functional requirement and product (the process of knowledge acquisition) are supported by smartphone experts. CRS we developed involves 259 smartphones, 51 functional requirements, and 18 specifications.

In this paper, we evaluate our proposed functional requirement-based compound critiquing. We apply empirical experiments to evaluate the recommendation accuracy [13], query refinement [14], and recommendation efficiency [15]. In addition, we also conduct a user study to evaluate user satisfaction [16]–[19]. It is because the user satisfaction does not always ensure a high recommendation accuracy [19].

The paper is organized in some sections. Section 2 discusses about overview of our funtional requirements-based compound critiquing. Section 3 outlines the evaluation framework. Section 4 and 5 present performance-based evaluation and user satisfaction evaluation, respectively. Finally, we provide conclusions and future work in the last section.

II. Functional Requirements-Based Compound Critiquing

During the interaction, CRS models the user preference in user profile. In this approach, there are two steps of interaction between the user and the system: 1) CR1, i.e. the initial step in interaction where the user has not given user preference (empty user profile), 2) CR2, i.e. the interaction step following CR1, where the system gives a question in the form of compound critiques to refine the user's needs. In CR1, system gives several common critiques functional requirement. In addition, the user also has a chance to give priority of preference. Fig. 1 shows the screenshoot of CR1. Priority of preference refers to several options including "*preferred*", "*optional*", and "*unpreferred*". Priority of preference will then be used as one of the parameters in proceeding generation of compound critiques in CR2.

In CR2, system will give n options in the form of compound critiques as well as m actual products that satisfy each compound critiques [20]. The user study we have conducted that involves 150 users shows that the number of compound critiques options (n) that may still make the user feel comfortable is four options. User can choose one of the products recommended. However, if there is no product the users like, the user can choose one of the compound critiques given by the system. After that, the system will recommend several products that fit with the compound critique chosen by the user, and the system will generate the more specific compound critiques for the user. The form of feedback in CR2 can be seen in Fig. 2.

The interaction between the user and the system will keep continue until the user chooses a product. The functional requirement-based compound critiques are combination of individual product functional requirements that lead to a pattern.

III. Evaluation Framework

To evaluate the functional requirements-based compound critiquing we proposed, we built two applications with two different interaction model. The first application (called as ***proposed system***) is conversational recommender system that implements functional requirement-ba sed compound critiques interaction model. The second one is common recommender system that is used in e-commerce (Interaction model based on the technical features of the product) such as amazon, bukalapak, bhinneka, etc, for baseline model (called as ***flat system***) [18].

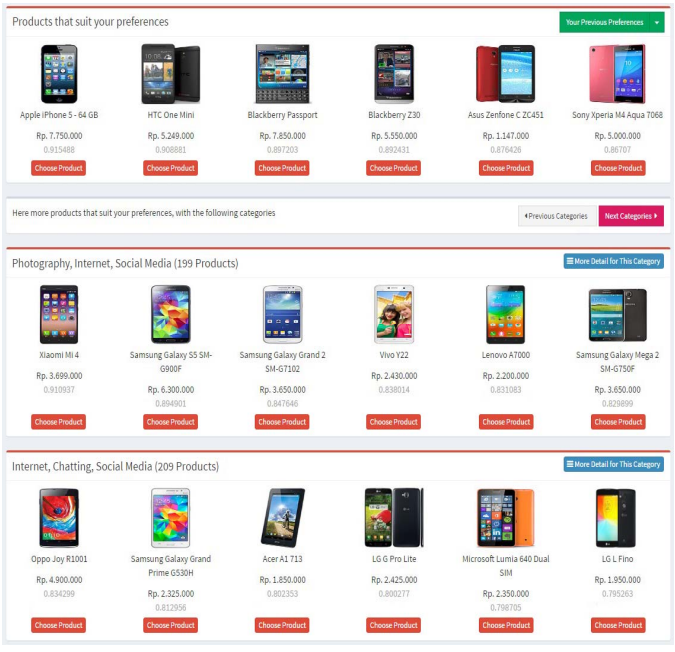


Fig. 2. An Example of Interaction CR2

The evaluation is divided into two categories: 1) performance-based evaluation and 2) user satisfaction evaluation. For performance-based evaluation, we take the aspects of: 1) recommendation accuracy, and 2) query

refinement. Meanwhile, for user satisfication evaluation we take: 1) user experience evaluation, and 2) user satisfaction evaluation.

The user study involves 88 users. All users participating in the test are smartphone users and familiar with web-based application and e-commerce. Users are then classified into two types, i.e. familiar with technical features (called as *expert user*) and unfamiliar with technical features (called as *novice users*) [19]. Besides that, users are also classified based on their age using cohort age group [21]. Furthermore, it also involves detail of demographic characteristics including: 1) type of users (34 expert users and 54 novice users), 2) gender (57 male users and 31 famale users), and 3) age (43 users with age between 18-24 and 45 users with age between 25-34).

Firstly, user is prompted to try both *proposed system* and *flat system*. Then each user is asked to fill out our questionnaire. System records all users' activities when they interact with the system. The data pertaining to the users' activities are used as the data for performance-based evaluation (recommendation accuracy dan query refinement). To evaluate performance-based evaluation, we only focus on *proposed system*.

IV. Performance-based Evaluation

A. Recommendation Accuracy

Recommendation accuracy is used to measure the success rate of a recommendation on a method in recommender system. One of the methods in measuring recommendation accuracy or the success rate of a recommendation is by taking a note on the percentage of successful interaction. The interaction is considered to be successful in CRS if the user can choose a product recommended [13]. In this study, we will see the recommendation accuracy based on user categories: 1) gender, 2) familiarity of product technical specification, and 3) age.

Fe´lix Herna´ndez del Olmo [13] suggests that recommendation accuracy can be formulated into Eq. 1.

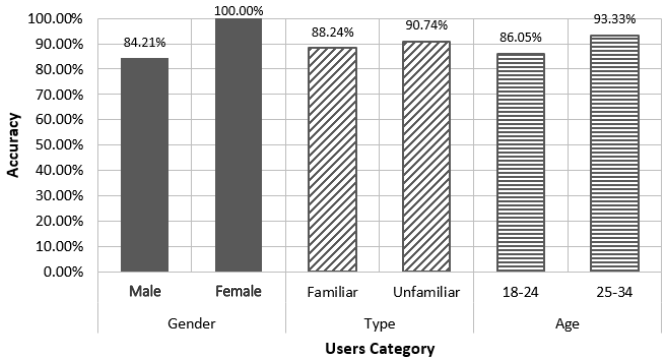


Fig. 3. Result of Recommendation Accuracy

$$accuracy = \frac{number\ of\ successful\ recommendations}{number\ of\ recommendations} \quad (1)$$

Based on the results depicted in Fig. 3, generally, the average percentage of recommendation accuracy for *proposed system* is as 89.77%. An interesting result can be seen in the category of gender. The system can reach the success rate of 100% in female category (10 out of 10 users find the product that meets their preference), while the percentage for male users can only reach the point of 84.21%. It is in line with the studies in the field of consumer behavior that basically, females are more detail in choosing a product but easy to be persuaded to choose products with various features supported by detail information [21]. It is because the *proposed system* is equipped with the explanation facility in depicting a product.

However, for novice user category, the system reaches a high success rate (90.74%), while for expert user category, the success rate reaches the point of 88.24%. It is in line with the our hypothesis that the functional requirements-based compound critique approach will be more helpful for the novice user. On the other hand, based on the test, expert users also feel satisfied with functional requirement-based compound critiquing. It can be seen from the success rate for expert user that reaches the point of 88.24% (8 out of 10 users).

B. Query Refinement Evaluation

Query refinement in CRS is a mechanism in generating an interaction that can refine the users' needs, so that it can ensure that the user-system interaction will be more efficient. The performance of query refinement can be evaluated based on the system ability in limiting the number of products recommended in each interaction session [14].

In this evaluation, we attempt to see the quality of query refinement of functional requirement-based compound critiquing method by comparing with the result of baizal, et.al [14] that uses same ontology (with similar number of product, number of functional requirements and number of specifications). The study uses functional requirement-based interaction, where each question is built in the form of single functional requirement with navigation-by-asking (NBA) strategy. We use a scenario similar to the study where the simulation is conducted through 40 interactions and the users (simulated users) are consistent with their "preferred" answers for each question given. However, it is slightly different in a case that we use a similar question in each interaction but the users' answers are random for each question given. Functional requirements-based compound critiquing we developed is one of navigation-by-proposing (NBP) strategies in CRS.

Fig. 4 shows the result of query refinement evaluation. It can be seen that functional requirements-based compound critiques in the initial interaction has already been able to limit the recommendation. It is because that compound critiques can refine the user preference more. Generally, the functional requirements-based compound critique approach is better in refining user preference compared to single functional requirements-based interaction. However, single functional requirements-based approach is more significant in refining users' need in each interaction session.

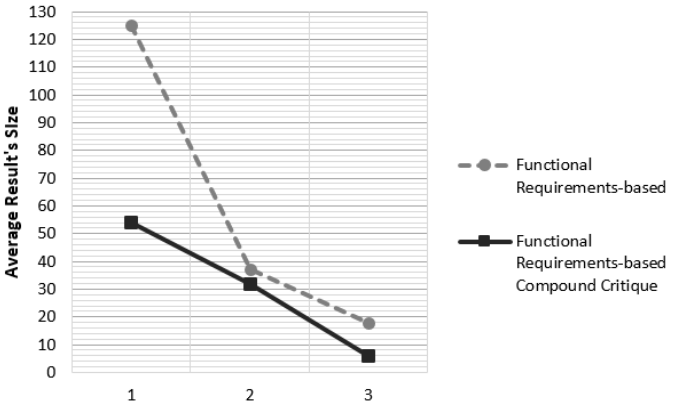


Fig. 4. Result of Query Refinement Evaluation

V. User Satisfaction Evaluation

A. First Stage Quesionare

Besides doing performance-based evaluation, we also evaluate user satisfaction by using questionnaire. The evaluation is inspired by the study from James Reilly et.al [16] and Bart P. Knijnenburg et.al [19]. Pearl Pu and Li Chen [17] argues that user is not always associated with a high recommendation accuracy, so that the user satisfaction evaluaton is then conducted to validate recommendation quality in terms of users' subjective evaluation.

In this test, we use two kinds of questionnaire: 1) first-stage questionnaire and 2) post-stage questionnaire. First stage questionnaire consists of 10 statements as can be seen in table 4.2. The statements are then classified into 6 factors or constructs as to help the analysis. The six factors are: 1) *perceived recommendation quality* (PRQ), 2) *perceived efficiency* (PE), 3) *trust* (TR), 4) *informative* (INF), 5) *easy to use / usability* (ETU), 6) *ease of understanding* (EOU). The users evaluate the system by putting a checklist to sign their agreement representing the score of 1 or cross to sign their disagreement representing the score of 0 for each statement. Statements in the questionare consist of positive and negative statements to control the respondents' answers not to be bias [6]. The average score that represents the agreement level for each statement is then calculated ranging from -1 to 1, where -1 refers to strongly disagree and 1 to strongly agree.

A summary of the responses is shown in Fig. 5. As we can see, *proposed system* in general gets positive responds for all statements. Based on the result, it can be seen that the *proposed system* gets a better agreement level compared to the *flat system*, in all factors. It is relevant to our hypothesis that the system with functional requirement-based compound critiquing helps users in interacting with system. Functional requirements-based compound critiquing method makes the presentation style of recommender system simpler and helps the users understand the questions given. It is in line with the result of the questionnaire that factors *easy to use* (S4, S9) and *ease of understanding* (S1, S2), *proposed system* has a better agreement level compared to *flat system*.

TABLE I. Statements of First Stage Questionnaire

ID	Factors	Statements
S1	EOU	Questions/options delivered are easy to understand
S2	EOU	I can understand very well, all the questions that were given to me
S3	INF	I can find information on products easy
S4	ETU	Overall, I found it difficult to find products that comply with my wishes
S5	PE	I can find a product that I prefer fast
S6	PRQ	I really like the product I selected
S7	TR	I really would buy the product that I choose in this system, someday
S8	PRQ	I do not like the interaction of this system
S9	ETU	I have no difficulty in using this system
S10	TR	I am interested to use this system again, if one day I want to buy a phone

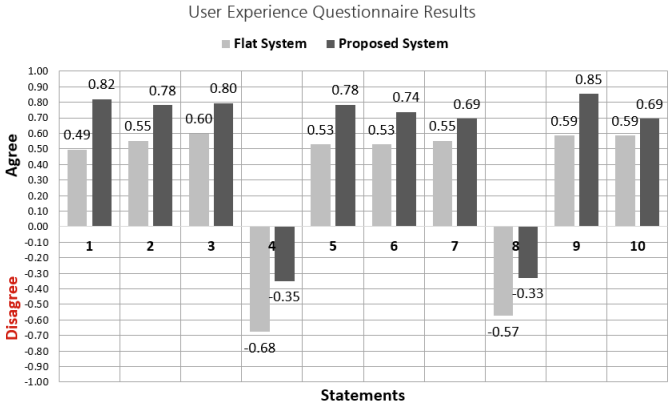


Fig. 5. First Stage Questionnaire Results

We then use T-test as to determine the significance of the agreement level average between *proposed system* and *flat system* on each factor. We involve novice user and expert user as the variables in the evaluation.

The T-test result on Table II shows that the agreement level of the novice and expert users is significantly different. For the novice users, *proposed system* receives positive feedback, and significantly different (significant level $\alpha = 0.001$ and $\alpha = 0.01$) for all factors. The results show that novice users are more satisfied with *proposed system*, compared to *flat system*. Therefore, we can conclude that the *proposed system* can increase the positive perception of the novice user for all factors.

What makes it interesting is that the agreement level of the expert user is inversely proportional to the one of the novice user. For the expert user, the different of agreement level between *proposed system* and *flat system* is not significant though the two systems get a positive result. Meanwhile, the *easy to use* (usability) factor shows that it is most significant among all factors with p=0.075. It proves that expert users also

feel that *proposed system* is more helpful compared to the *flat system*.

TABLE II. T-TESTING RESULTS

Factors	Model	Expert User				Novice User			
		Mean	t	df	p-value (2-tailed)	Mean	t	df	p-value (2-tailed)
PRQ	Proposed System	**0.6029**	0.461	66	0.646	**0.7685**	5.171	106	0.000
	Flat System	0.5588				0.4259			
PSE	Proposed System	**0.7059**	1.008	66	0.317	**0.8333**	3.891	106	0.000
	Flat System	0.5882				0.5000			
TR	Proposed System	**0.6176**	-0.710	66	0.480	**0.7407**	3.285	106	0.001
	Flat System	0.6912				0.5000			
INF	Proposed System	**0.6765**	0.000	66	1.000	**0.8704**	3.821	106	0.000
	Flat System	0.6765				0.5556			
ETU	Proposed System	**0.6912**	1.808	66	0.075	**0.7870**	5.561	106	0.000
	Flat System	0.5294				0.4074			
EOU	Proposed System	**0.7059**	0.145	66	0.885	**0.8519**	6.086	106	0.000
	Flat System	0.6912				0.4259			

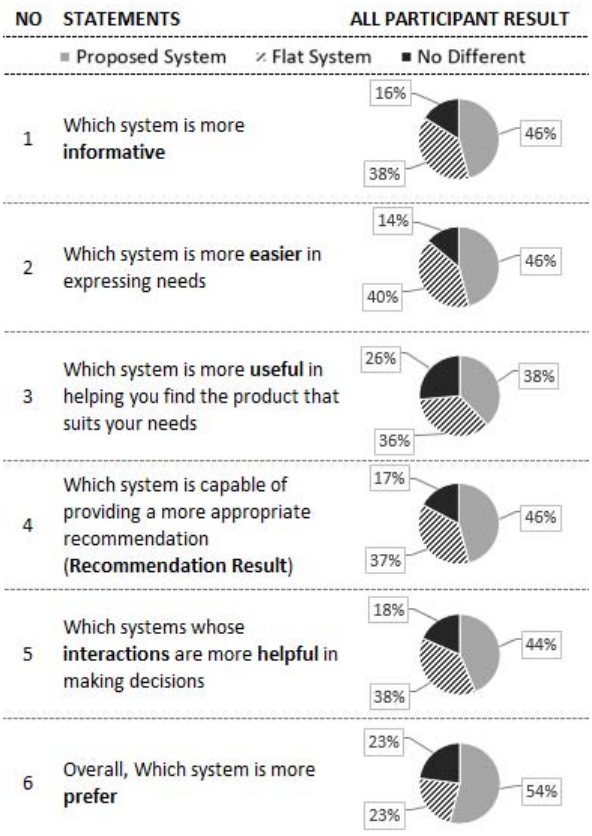
In general, the user satisfaction evaluation on the first stage shows that the *proposed system* has significantly increased the positive perception of novice users ($p < 0.001$ and $p < 0.01$) for all factors. Generally, expert users see the *proposed system* positive with the average score of 0.667 for each factor.

B. *Post Stage Questionaire*

Post stage questionnaire is the second step of user satisfaction evaluation. The test is used as to validate the consistency of the users' answers in the first stage questionnaire [16]. While the first stage questionnaires consist of the statements representing factors or construsts, the post-test user questionnaires consist of implicit questions pertaining to the factors. The post-test questionnaires simply ask each user to vote on which system (*proposed system* or *flat system*) performed better in terms of various criteria such as informative, usability, interaction, etc.

Table III shows the result of post-test questionnaires from all participants (88 users). The result shows that the users consistently feel satisfied with *proposed system* compared to *flat system*, for all factors. However, both of the two systems have a high rate of votes with little difference in the total votes. What makes it interesting is that in the question "Which system is more useful in helping you find the product that suits your needs" (*perceived recommendation quality*), the different rate of votes of the two systems is the smallest compared to the other factors, i.e 33 votes vs 31 votes (different by 2 votes only) and the other 23 users feel that there is no difference. It shows that the two systems give a suitable recommendation. Accordingly, if it is seen from the final question, "Overall, Which system is more prefer", the difference for the two systems is quite big, i.e. 27 votes or 30.68% (System A: 47 votes, System B: 20 votes, No Difference: 20 votes). It shows that recommender system gives not only a suitable recommendation, but also other factors including practicability, presentasion style, and interaction style that influence the users' satisfaction. As we can see, *proposed system* can give a better satisfaction level than the *flat system*.

TABLE III. POST- STAGE QUESTIONNAIRE RESULTS



VI. CONCLUSION

Functional requirement-based compound critiquing is one of the solutions from NBP-based CRS, where the user-system interaction refers to the functional requirements that may not require the users to understand the product technical features. We evaluate this approach from the aspects of recommendation accuracy, query refinement, and user satisfaction (first stage dan post stage).

Based on the first stage user satisfaction evaluation, proposed system (CRS built based on our proposed approach [12]) gets positive response for all factors (perceived recommendation quality, perceived system efficient, trust, informative, easy to use / usability, ease of understanding). Proposed system significantly gets a better satisfaction level (significant level $\alpha = 0.001$ and $\alpha = 0.01$) compared to *flat system* (application that implements common recommender system used in e-commerce) for all factors in novice user category. However, for expert users, the satisfaction level for the two systems is not significantly different. From post stage user satisfaction evaluation, the two systems get a positive result with a little difference in the number of vote. As for the final question, "Overall, Which system is more prefer", the

difference of vote shared between the two systems is quite big, i.e. 30.68%. Overall, based on the user satisfaction evaluation, proposed system can give a higher satisfaction level, compared to the *flat system*.

User satisfaction is also supported by the performance-based evaluation. Based on the evaluation of recommendation accuracy, the average score of recommendation accuracy for proposed-system is as 89.77% (novice user: 90.74% vs expert user: 88.24%). The result shows that the functional requirements-based compound critiquing approach is more helpful either for novice user or expert user. Furthermore, based on query refinement evaluation, functional requirements-based compound critiquing approach works well in refining user preference yet it is not significant in refining the users' needs in the forthcoming interaction. It is because that this approach has been able to limit the recommendation result since the first interaction.

In general, the test on user satisfaction evaluation and performance-based evaluation shows that the functional requirements-based compound critiquing approach has successfully either increased the users' positive perception with high recommendation accuracy or refined the users' needs.

REFERENCES

[1] M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*, 1st Editio., vol. 40. Cambridge University Press, 2011.

[2] D. BRIDGE, M. H. GÖKER, L. McGINTY, and B. SMYTH, "Case-based recommender systems," *Knowl. Eng. Rev.*, vol. 20, no. 3, p. 315, Sep. 2005.

[3] R. D. Burke, K. J. Hammond, and B. C. Young, *Knowledge-based navigation of complex information spaces*. AAAI Press, 1996.

[4] R. D. Burke, K. J. Hammond, and B. C. Yound, "The FindMe approach to assisted browsing," *IEEE Expert*, vol. 12, no. 4, pp. 32–40, Jul. 1997.

[5] L. McGinty and B. Smyth, "Improving the Performance of Recommender Systems That Use Critiquing," in *Intelligent Techniques for Web Personalization*, vol. 3169, 2005, pp. 114–132.

[6] J. Reilly, K. McCarthy, L. McGinty, and B. Smyth, "Incremental Critiquing," in *Research and Development in Intelligent Systems XXI*, London: Springer London, pp. 101–114.

[7] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth, "On the Dynamic Generation of Compound Critiques in Conversational Recommender Systems," *Adapt. Hypermedia Adapt. Web-Based Syst.*, vol. 3137, no. 3, pp. 176–184, 2004.

[8] L. Chen and P. Pu, "Preference-Based Organization Interfaces: Aiding User Critiques in Recommender Systems," *User Model. 2007*, pp. 77–86.

[9] H. Xie, L. Chen, and F. Wang, "Collaborative Compound Critiquing," *22nd Conf. User Model. Adapt. Pers. (UMAP 2014)*, vol. 8538, pp. 254–265, 2014.

[10] D. Jannach, "Advisor suite-a knowledge-based sales advisory system," *ECAI*, vol. 16, p. 720, 2004.

[11] D. H. Widyantoro and Z. K. A. Baizal, "A framework of conversational recommender system based on user functional requirements," in *2014 2nd International Conference on Information and Communication Technology, ICoICT 2014*, 2014, pp. 160–165.

[12] Y. R. Murti and Z. K. A. Baizal, "Compound Critiquing for Conversational Recommender System Based on Functional Requirement," *Adv. Sci. Lett.*, vol. 22, no. 8, pp. 1892–1896, Aug. 2016.

[13] F. Hernández del Olmo and E. Gaudioso, "Evaluation of recommender systems: A new approach," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 790–804, Oct. 2008.

[14] Abdurahman Baizal, D. H. Widyantoro, and N. U. Maulidevi, "Query Refinement in Recommender System Based on Product Functional Requirements," in *The 8th International Conference on Advanced Computer Science and Information Systems (ICACSIS 2016)*, 2016.

[15] J. Reilly, J. Zhang, L. McGinty, P. Pu, and B. Smyth, "A comparison of two compound critiquing systems," in *Proceedings of the 12th international conference on Intelligent user interfaces - IUI '07*, 2007, p. 317.

[16] J. Reilly, J. Zhang, L. McGinty, P. Pu, and B. Smyth, "Evaluating compound critiquing recommenders," in *Proceedings of the 8th ACM conference on Electronic commerce - EC '07*, 2007, p. 114.

[17] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems - RecSys '11*, 2011, p. 157.

[18] Z. K. A. Baizal, D. H. Widyantoro, and N. U. Maulidevi, "Factors Influencing User's Adoption of Conversational Recommender System Based on Product Functional Requirements," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 4, 2016.

[19] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User Model. User-adapt. Interact.*, vol. 22, no. 4–5, pp. 441–504, Oct. 2012.

[20] P. Pu and L. Chen, "Trust building with explanation interfaces," in *Proceedings of the 11th international conference on Intelligent user interfaces - IUI '06*, 2006, p. 93.

[21] W. L. Wilkie, *Consumer behavior*. Wiley, 1990.