Purchase Factor Expression for Game Software Using Structural Equation Modeling with Topic Model in User's Review Texts

Rikuto Kunimoto and Ryosuke Saga

***Abstract*—Considering user opinion in game software development is important from a marketing viewpoint, because there are no effective ways to analyze the market of game software. In this research, we attempted to develop an analysis process for consumers' review comments by using topic model and structural equation modeling. By using this approach, we aimed to extract the relationships of elements to which users seem to direct their attention visually and quantitatively, and we expected to extract meaningful knowledge for game software development. Experimental results suggest that our proposed process can analyze the market as effectively as the text-based model generation method for confirmatory factor analysis.**

***Index Terms*—Causal analysis, factor expression, game software, structural equation modeling, topic model, hierarchical latent dirichlet allocation.**

I. Introduction

With the rapid expansion of the platform diffusion rate spurred on by the spread of smartphone and tablet terminals due to recent global technological advances, the game software market, including consumer, mobile, and amusement facilities, has become a large-scale market worth $61.400 million as of 2012. A report by CAPCOM co. LTD. investigation group indicates that the size of the game software market is expected to reach $86.6 million by 2017 [1], [2]. However, the difficulty of market investigation is one of the most important problems for any game software developers, whereas rapid growth of the market size is accepted. The difficulty of identifying consumers' purchasing factor is a notable issue, given that many developers unanimously say that they are unable to know whether their products will be popular until they send it off in the market [3], [4]. As in the related work, Kunimoto *et al*. attempted to extract the important factors by using KJ method [5] and model integration methods with structural equation modeling (SEM) [6]. They propose the path model generation process, which uses the idea of collective intelligence. Saga et al. attempted to improve the analysis process with SEM by using text information [7]. They showed the effectiveness of using a combination of models and text information for factor analysis with SEM. However, issues still exist because of the explanation ability of the factor model, which inadequately expresses the text-based confirmation analysis process. In this research, we propose an analysis process that uses mainly two information techniques, namely, SEM and topic model, as one of the approaches for the problem based on the idea of visualization to analyze invisible phenomenon as a latent factor in text data. Higher collective intelligence exists in users' comments and reviews. Thus, factor analysis that uses such information source will have higher explanation ability. The topic model will enable us to structurally understand specific topicality when users evaluate game software by electronically analyzing existing text data (corpus). We suggest combining the factor analysis method, namely, SEM, with the obtained structured topic model to analyze users' interests visually and quantitatively. We aim to show the possibility of an effective investigation method for the game software market and to help developers.

II. Topic Model

Topic model is a machine learning technique that clarifies the structure of a document group by estimating words that constitute a topic based on the premise that each document group that constitutes the corpus belongs to the specific topic. Several topic model methods are available, such as latent semantic indexing (LSI) [8], latent Dirichlet allocation (LDA) [9], and hierarchical LDA (hLDA) [10], which is a progressive LDA technique. We use LDA as the foundation of the topic model because a reviewer's comment can be safely assumed to have several background topics. Furthermore, we adopted hLDA, which is highly compatible with SEM, as a concrete step in our analysis process.

A. hLDA

hLDA is the representative hierarchical topic model. In hLDA, the potentiality topic constitutes the part tree of infinite height and the hierarchy structure branches off endlessly, unlike LDA, which assumes a flat potentiality topic. Adopting hLDA has two advantages. First, relationships between topics do not need to be considered, and second, the number of topics will be estimated automatically by the algorithm of the hLDA process. Hierarchy structure is generated by using the nested Chinese restau-rant process [11] in which the visitor and the table (or a restaurant) expresses the document and the topic, respectively. The generation process of hLDA is as follows: First, the parameter of multinomial distribution (Dirichlet allocation) on words for each topic is chosen, as shown in Fig. 1. Then, the root node

Manuscript received October 17, 2014; revised December 19, 2014. This work was supported by a Grant-in-Aid for Foundation of the Fusion of Science and Technology; FOST, and MEXT/JSPS KAKENHI 2524049, 25420448.

R. Kunimoto and R. Saga are with Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai-City, Osaka, Japan (e-mail: kunimoto@mis.cs.osakafu-u.ac.jp, saga@cs.osakafu-u.ac.jp).

DOI: 10.7763/IJIMT.2014.V5.551

of the topic to the node that rides on the path for each document is set. After that, the node is selected according to the defined probability for each hierarchy level. Next, the parameter of multinomial distribution on words is chosen. Finally, the level and word (generated by multinomial distribution of topic) for each place where the word will be inserted in the document are chosen.

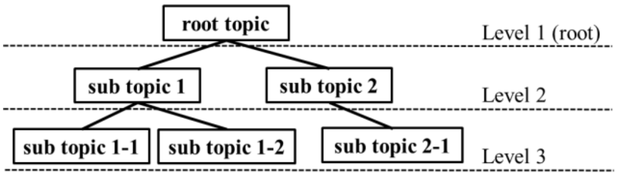

Fig. 1. Hierarchical structure of hLDA.

III. Structural Equation Modeling

SEM [12] analyzes various relationships among several factors, i.e., latent and observed variables. A latent variable is an invisible concept for target analysis. For instance, “bone” and “mineral” are used in biology [13]. An observed variable is an observable item from a target analysis and is used to estimate a latent variable. These variables have relationships, such as causal and co-occurrence relationships. SEM can quantify the influence and strength of these relationships [14].

A path model is used to comprehend the variables’ relationships. A path model visualizes factors and relationships among factors, as shown in Fig. 2. In the path model, an observed variable is expressed as a rectangle and a latent variable as an ellipse. The relationships among variables are expressed by unidirectional arrows and bidirectional arrows, which correspond to causal relationships and co-occurrence relationships, respectively.

The path model shown in Fig. 2 consists of three observed variables (B, C, and D) and one latent variable (A). The relationship between A and B, denoted as α_1, is co-occurrence, and the other relationships, denoted as α_2 to α_5, are causal relationships.

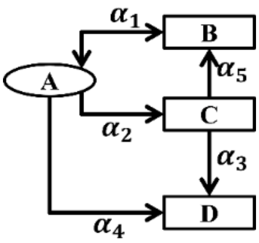

Fig. 2. Path model of SEM.

IV. Analysis Process Using Topic Model with SEM

This section describes the concrete process of our proposed factor analysis that combines SEM and topic model. The proposed process consists of the following five steps:

A. Obtaining the Corpus of the Research Target for the Learning Topic Model

The corpus must be collected based on the tool to be used to learn the topic model, such Stanford Topic Modeling Toolbox [15] and Mallet [16]. For example, if we use Mallet, we must create a dataset file in .csv, .tsv, or .txt format. Data unit should be a row or a file.

In this research, the objective is to extract game software purchase factors. How-ever, our proposed analysis process is not limited to this case. Using the approach is not an issue if review texts on the Web are used as corpus.

B. Learning Topic Model Using hLDA and Structuring Path Model

After acquiring the text source in step 1, we carry out the learning of the topic model by using hLDA. In hLDA, each lower-level topic is generated by the higher-level topic. Therefore, setting the path based on the following rules is recommended [14]: (1) drawing the path toward lower-level topics from each higher-level topic, and (2) draw-ing the path toward each word that constructs the topic from the topic. A clearly identified problem is a precondition for SEM, as mentioned in Section III. If we implement the process based on the above two rules, identification problems will not appear in the final stage, and the model is stable.

C. Estimating Learned Topics and Selecting Representative Keywords

We may understand the kind of topic that is expressed by looking at the keyword group that constitutes the learned topic model. Table 1 shows an example of key-words that constitute a topic. The keyword group topic is output in a state that is sorted sequentially to have a high probability of being generated by the learning algorithm. We should choose three high-ranking keywords, except for the incomprehensible words, and construct the model by following the approach described in step 2, as in Fig. 3. However, we can use all keywords that constitute the topic when we guess the label of the model. The oval figures express topic node as the latent variable, and the black one is the root topic. The rectangular figures express keywords as observed variables. Each alpha 1 to 15 expresses the contribution degree that will be calculated in SEM. The topicality can be understood visually and quantitatively by using the path model in SEM based on the topic model learned from the text data.

In addition, from the viewpoint of the identification problem of SEM, we recommending avoid a situation in which a repeating word appears and is selected.

D. Generation of Pseudo-Statistical Data to Create Variance-Covariance Structure Data

TABLE I: The Arrangement of Channels

Evaluation (root topic)	quality originality satisfaction good fun play game evaluate …
…	…
usability (sub-topic 1)	operate system interface control load memory continue platform …
…	…
visual (sub-topic 2)	character part strong mystery enemy scenario grow support chapter …
…	…

As mentioned in Section III, a structural equation is constructed in SEM under the assumption that some correlations exist as multiple regressions, and confirmation is performed according to such assumption. Therefore, analyzing numerical data is necessary.

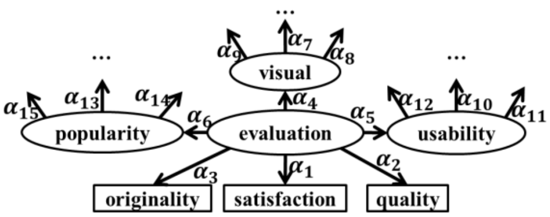

Fig. 3. Example of creating the model from the topic model.

As a basic idea, we consider one document to be the unit data in this set. The set without repeating words that constitute the topic is set as the data field. The number of times the words appeared in each document is set as the value.

In addition, the characteristic of data is expected to be expressed strongly by considering the weight of words used in the information retrieval algorithm for those values. We tested the five weightings (simple appearance frequency, global frequency IDF, inverse number of document frequency, probabilistic frequency IDF, and entropy) [17] in the preliminary experiment. The use of simple frequency obtained the most stable result. Thus, we recommend using the simple appearance frequency degree in the analysis.

E. Performing Analysis by SEM and Evaluation of Analysis Result

In SEM, the analysis result indicates that the path model has a calculated contribution degree between each item and some indexes that evaluate conformity degree between the model and data or its balance quantitatively. We use four representative indicators that are usually used in SEM analysis for reference. These indicators are as follows: GFI and AGFI should be closer to 1 and over 0.9 to indicate compatibility between the model and the data. RMSEA should be closer to 0 and under 0.1, which indicates estrangement with the true model. BIC should be lower than the other models, which indicates balance between compatibility and information quantity.

V. Experiment

A. Goal, Dataset, and Process of Experiment

This experiment tests the proposed process by using actual data to determine whether the process can visually and quantitatively provide sale strategists or software developers with useful knowledge.

We chose a Japan-based game software review site called mk2 as data for this experiment [18]. Game software for different gaming consoles (PS4, PS3, PSP, PSV, N3DS, NDS, Xbox, Xbox 360, etc.) are evaluated by users and then collected and published on the review site. The contributed data are totaled according to each title. Quantitative evaluation of the quality (including graphics, music, originality, and comfort) and texts about the pros, cons, and general comments of every reviewer are registered. These sets, which were contributed by every user, are treated as unit data in the experiment.

We collected corpus about five major software titles of Nintendo Wii, namely, Su-per Mario Galaxy, Mario Kart 9, Monster Hunter Tri, Fire Emblem: Radiant Dawn, and Super Smash Bros. Melee X. The number of data of each title is 101, 82, 108, 101, and 185, respectively.

With the text dataset, we collected the overall rating (OR) as given by the users to indicate the overall quality of the title. The rating is evaluated as the numerical indicator and defined from 0 to 100; a high rating indicates that the game is popular and interesting for users.

During the experimental process, we were aware that the analysis result of a preliminary experiment may change greatly by changing the analysis conditions of steps 1 and 2. Therefore, we present several ways of analyzing results. A concrete experiment process is described below.

1) Step 1

As mentioned in the previous section, we collected comment datasets from mk2. For the learning topic model in the next step, we prepared review comments for each software (five titles) in this step. We expected to acquire a more detailed topic model than that which was learned in the set of review comments of many software titles.

As another way to collect corpus, we regarded review comments from five titles as a set. We expected that the bias for each title will decrease and that the acquired topic model will express more global topics compared with the method that uses each text every title. The number of datasets is the sum of the five titles (577).

2) Step 2

We structured the model for multiple interpretabilities of our experimental results in two ways.

First, we divided comments into “good” and “bad,” and implemented the learning topic model for each corpus. This method aims to inform the analyst how positive and negative factors influence OR.

Second, no division pattern is obtained. By analyzing overall comments and the learning topic model, we expect to obtain a simple and more appropriate information quantity model. OR is influenced only by the root topic of the learned topic model.

3) Step 3 to 5

Steps 3 to 5 should be implemented as mentioned in Section IV, which described our proposed process. In this experiment, we used SEM package supplied in R software version 12.2.2 [19], [20], a well-known statistical analysis tool. The SEM package of this software provides the source code for using the visualization tool GraphViz [21], which can present the analyzed model as a figure.

B. Results and Discussion

Experimental results are shown in Tables II and III, and include Fig. 4 and Fig. 5.

Fig. 4 is the visualized figure of Model 12, which had the highest evaluation score among the 12 models that we constructed. The number of latent variables (four) is the size that is appropriate for examining the entire model. Latent variables labeled “evaluation” and “topic 1 to 3” express the root topic and its lower-level topics. In this model, each indicator score was excellent. When we looked at the entire model to discern meaningful information, we found that this model provided only a few interesting results. Words such as "good," "time," and "play" are common and unhelpful for estimating the topic, and they do not provide new or important knowledge. However, we could understand the topicality of the review comments of users. Latent variable topics 1, 2, and 3 are "platform," "worth playing," and "simplicity." This topic model shows us how the evaluation point is structured. The abovementioned findings confirm the real relationships of users' evaluation points.

TABLE II: DESCRIPTION OF EACH MODEL DATA

Model name	Title name	Comment division	Number of latent variables
Model 1	Super Mario Galaxy	Yes	7
Model 2	Mario Kart 9	Yes	9
Model 3	Monster Hunter Tri	Yes	8
Model 4	Fire Emblem: Radiant Dawn	Yes	9
Model 5	Super Smash Bros. Melee	Yes	12
Model 6	Five titles combined	Yes	8
Model 7	Super Mario Galaxy	No	3
Model 8	Mario Kart 9	No	3
Model 9	Monster Hunter Tri	No	4
Model 10	Fire Emblem: Radiant Dawn	No	4
Model 11	Super Smash Bros. Melee	No	5
Model 12	Five titles combined	No	4

TABLE III: RESULTS OF EACH MODEL INDICATOR

Model name	GFI	AGFI	RMSEA	BIC
Average of Models 1–5	0.722	0.676	0.0891	-1016
Average of Models 7–11	0.850	0.788	0.0866	-217
Model 6	0.768	0.719	0.0986	65.2
Model 12	0.968	0.953	0.0398	-275

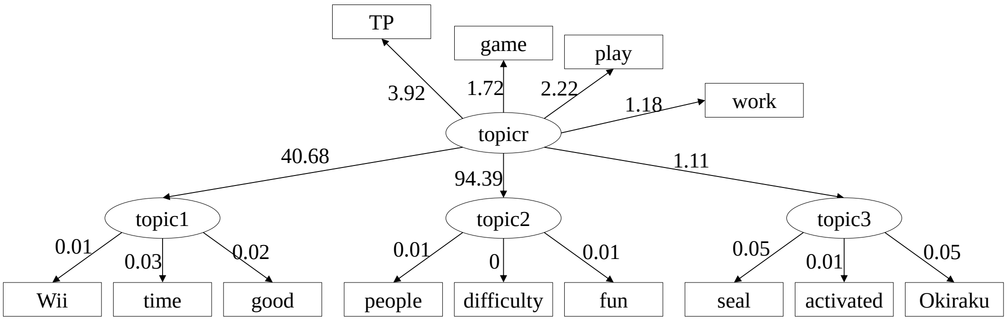

Fig. 4. Visualized causal relationships between topics and keywords of Model 12.

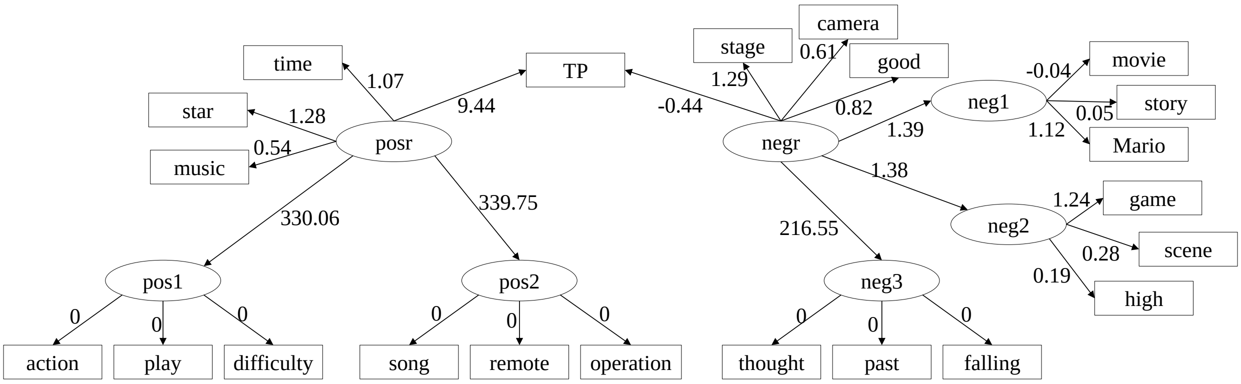

Fig. 5. Visualized causal relationships between topics and keywords of Model 3.

Fig. 5 shows that Model 3 had the highest score among the models of the review comment patterns categorized as "good" or "bad" in step 2. Latent variables named "pos evaluation," "pos 1 (to 3)," "neg evaluation," and "neg 1 (to 3)" denote the root topic of positive comments, its lower-level topic, the root topic of negative comments, and its lower-level topic, respectively. Based on keywords, pos1, pos2, and pos3 can be derived as "playing elements," "improved points," and "battle actions," respectively. Similarly, neg1, neg2, and neg3 could be assumed to refer to "boss battle," "limitation conditions," and "visuals." This model shows the small positive contribution degree (0.2) from the root of the negative factor and the large positive contribution degree (1.98) from the root of the positive factor. The proportion of meaningful words such as "weapon," "armor," and "monster," increased compared with those in Model 3 or in Model 12, in which review comments were not categorized. This trend is seen in other similar pattern (with categorized comments) models. The fact that the degree from the negative factor is not a negative value indicates that the user review is not an insult or abasement, but provides productive criticism and suggests improvements for the software developers to consider. This trend is observed in real contents of review comments in mk2. The above findings confirm that our proposed process not only expresses relation-ships visually and quantitatively between topics written by users as review comments, but also identifies which element users tend to evaluate in game software, unlike other factor analysis methods that use a non-structured factor model [4], [7].

VI. Conclusion

This paper proposed a process of analyzing game software market by using topic model and SEM. The basic idea of the proposed method is that visual and quantitative analysis that uses text data contributed by many people will make it possible to know the factor structure of the game software market that is too complex to analyze effectively.

We proposed a concrete analysis process composed of five steps. Step 1 involves collecting corpus from the field that the analyst wants to investigate. In Step 2, the learning topic model is implemented by using hLDA. In Step 3, a path model is constructed for analysis by using SEM. In Step 4, pseudo-statistical data are created by using appearance frequency of keywords in the corpus as the numerical dataset. Analysis and evaluation are conducted in Step 5.

In the experiment, we collected comment text corpus from the Japanese game software review website mk2. We found that our proposed method may serve as a tool for discovering useful or confirmatory knowledge for analysts.

For future work, we will consider the use of automatic labeling method [22] in step 3 of the proposed method to improve the precision of our analysis process. Moreover, we will find improved methods of performing steps 3 and 4 using other concepts or weighting methods. In addition, the proposed process needs to be evaluated by real experts in the fields of game software development and market analysis.

Acknowledgment

This research was supported by a Grant-in-Aid for Foundation of the Fusion of Science and Technology; FOST, and MEXT/JSPS KAKENHI 2524049, 25420448

References

[1] CAPCOM Co., LTD.: Market Data. (September 30, 2013). [Online]. Available: http://www.capcom.co.jp/ir/english/business/market.html.

[2] METI Japan. About the overseas development measure of contents. [Online]. Available: http://www.meti.go.jp/committee/kenkyukai/seisan/cool_japan/pdf/011_05_00.pdf

[3] Textbook Production Committee of the Digital Game, "Textbook of the digital game," Softbank Creative co., LTD, 2010.

[4] K. Kitami, R. Saga, and K. Matsumoto, "Comparison analysis of video game purchase factors between Japanese and American consumers," *Lecture Notes in Computer Science*, vol. 6883, pp. 285-294, 2011.

[5] R. Scupin, "The KJ method: A technique for analyzing data derived from Japanese ethnology," *Human Organization*, vol. 56, no. 2, pp. 233-237, 1997.

[6] R. Kunimoto and R. Saga, "Path models integration methods for structural equation modeling by OR and probability concepts," *IEEE SMC*, pp. 257-259, 2013.

[7] R. Saga, T. Fujita, K. Kitami, and K. Matsumoto, "Improvement of factor model with text information based on factor model construction process," *IIMSS 2013*, pp. 222-230, 2013.

[8] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Info Sci*., vol. 41, pp. 391-407, 1990.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[10] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol. 57, no. 2, pp. 7, 2010.

[11] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," *Advances in Neural Information Processing Systems*, vol. 16, pp. 106-114, 2003.

[12] J. C. Loehlin, "Latent variable models: An introduction to factor," *Path, and Structural Equation Analysis*, 4th ed., Routledge, 2004.

[13] S. Toyokawa, H. Nishikawa, M. Ueji, K. Motegi, and K. Kano, "Structural equation modeling of the relationship of bone mineral density and its risk factors in Japanese women," *Environmental Health and Preventive Medicine*, vol. 6, issue 1, pp. 41-46, 2011.

[14] J. Pearl, *Causality*, second edition, Cambridge University Press, 2001

[15] Stanford Topic Modeling Toolbox. [Online]. Available: http://www-nlp.stanford.edu/software/tmt/tmt-0.4/

[16] MALLET. (2002). A machine learning for language toolkit. [Online]. Available: http://mallet.cs.umass.edu

[17] K. Kita, K. Tsuda, and M. Shishibori, *Information Retrieval Algorithms*, 8th ed., Kyoritsu Pub.

[18] Mk2 Group. [Online]. Available: http://www.psmk2.net/

[19] The R project for statistical computing. [Online]. Available: http://www.r-project.org/

[20] J. Fox, "Structural equation modeling with the SEM package in R. structural equation modeling," vol. 13, pp. 465-486, 2006

[21] Graphviz-Graph Visualization Software. [Online]. Available: http://www.graphviz.org/Theory.php

[22] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," presented at KDD'07, CM 9781595936097/07/0008, August 12–15, 2007.

R. Kunimoto is a master's degree student at computer science and intelligent systems, Graduate School Of Engineering, Osaka Prefecture University. He is now engaged in research on knowledge management and text mining, especially structural equation modeling and topic model. He is a student member of IEEE.

R. Saga received a bachelor's degree from Osaka Prefecture University in 2003 and completed the master's course in electrical and information engineering and the doctoral course in 2005 and 2008. He works as an associate professor in Osaka Prefecture University. He is now engaged in research on knowledge management, data engineering and decision support. He is a member of IEEE, etc.