Factor Analysis for Game Software Using Structural Equation Modeling with Hierarchical Latent Dirichlet Allocation in User's Review Comments

Rikuto Kunimoto, Hiroshi Kobayashi, and Ryosuke Saga

***Abstract*—While the market investigation is important in game software development, there is a problem that there is no effective way to pursue the factor of user's software evaluation. In this research, we paid attentions to corpus (electric existence of documents) considered that the factor relationships about the user's evaluation were expressed potentially as their opinions. As the way to achieve this idea, we tried to extract useful knowledge by using SEM and topic model for visual and quantitative analysis process. As the related work, there are several researches about Game software market using text mining methods (LSI, or LDA). However, they have the problem concerning to objectivity or explanations because the relationships between topics are not defined based on technical algorithms and expressed only as the frequency of the words that constructs the topics. Experimental results showed that our proposal process can extract effectively the topics that users pay attentions when they evaluate the game software and we can interpret it.**

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

I. Introduction

As of 2012, the game software market, including consumer, mobile, and amusement facilities, has become a large-scale market worth $61.400 million. This remarkable increase is due to the rapid expansion of the platform diffusion rate induced by recent global technological advances of smartphones and tablet terminals. An investigation group from CAPCOM Co. Ltd. Ref. [1] reported that the size of the game software market is expected to reach $86.6 million by 2017.

However, the difficulty in market investigation is one of the most important problems among game software developers, in which rapid growth of the market size is accepted [2]. The difficulty of identifying consumers' purchasing factor is a notable issue, given that many developers are unable to determine whether their products will be popular until they are placed in the market [3]. For luxury goods, identification of the purchasing factor is important, which is generally manifested byuser reviews. Saga *et al*. [4] attempted to analyze factor relationships of game software market using a topic model. Topic models are 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 comprising the corpus belongs to the specific topic. They proposed a path model generation process for structural equation modeling (SEM) using latent semantic analysis (LSA) [5], and then combined user reviews with the model. Other studies such as competitor analysis of consumer situation [6] and specification of negative factors [7] indicated that factor analysis based on latent Dirichletal location [8] (LDA, an example of a topic model) is effective for many applications. However, both LSA and LDA cannot define the relationships among topics in the learned model, and thus, the model is constructed based on the subjectivity of the analyst, by which information without objectivity is extracted. Furthermore, LSA and LDA can express relationships among topics and keywords that constitute the topic only in form of generation probability. Therefore, explanation ability is problematic when we consider total causal relationships among topics or the entire topic model, including keywords.

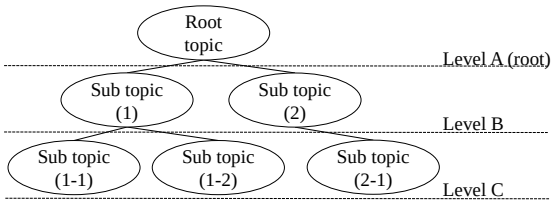

Fig. 1. Hierarchical structure of hLDA.

In this study, we propose an analysis process that uses mainly two information techniques, namely, SEM and hierarchical LDA (hLDA) [9], [10]. hLDA is an advanced technique of LDA and can automatically constitute the relationships among topics hierarchically. The model learned by hLDA is used for SEM. SEM is a causal analysis technique that expresses relationships among items, called latent and observed variables, using a path model. In the path model that treats texts, we regard topics and words as latent and observed variables. This model can visually and quantitatively express their relationships using arrows and path coefficients that have positive and negative values. By using these techniques, we can establish a model that expresses relationships among topics that are given by users when they objectively evaluate a game software; higher explanation ability can be attained when an analyst considers the entire model. In addition, we aim to show the possibility of an effective method for game software market analysis to help developers.

Manuscript received October 17, 2014; revised December 17, 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.

The authors are with Computer Science and Intelligent Systems, Graduate School of Engineering, 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/IJKE.2015.V1.9

II. Hierarchical Latent Dirichlet Allocatoin

hLDA is employed as 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 are unnecessary, and second, the number of topics is estimated automatically by the algorithm of the hLDA process. Hierarchical structure is generated by using the nested Chinese restaurant process in which a visitor and a table (or a restaurant) express the document and 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. The root node of the topicis then set to the node that rides on the path for each document. After that, the node is selected according to the defined probability for each hierarchy level. The parameter of multinomial distribution on words is then chosen. Finally, the level and word (generated by multinomial distribution of topic) are chosen for each place where the word is inserted in the document.

III. Structural Equation Modeling

SEM [11], [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.

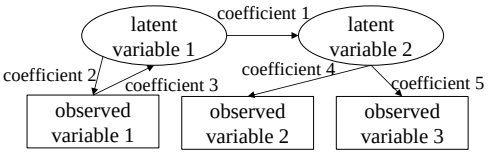

Fig. 2. Path model of SEM.

A path model is used to comprehend the relationships of variables. A path model visualizes factors and their relationships, 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 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 and two latent variables.

IV. Analysis Process Using hLDA with SEM

This chapter describes the concrete process of our proposed factor analysis that combines a SEM and a topic model. As shown in Fig. 3, the process consists of the following four steps:

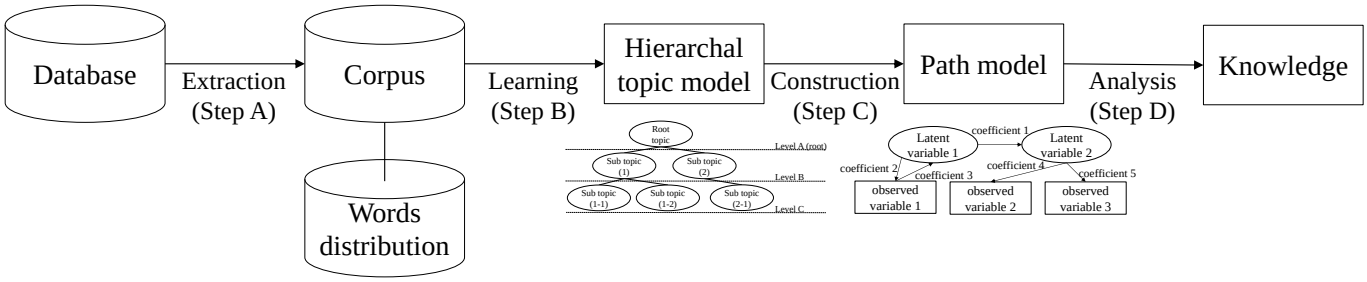

Fig. 3. Analysis process.

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

The corpus must be collected based on the tool to be used to learn the topic model, such as Stanford Topic Modeling Toolbox and Mallet [14]. For example, if we use Mallet, then we must create a dataset file in .csv, .tsv, or .txt format. Data unit should be a row or a file.

In this study, the objective is to extract game software purchase factors. However, our proposed analysis process is not limited to this case. The use of this approach is not an issue when review texts on the Web are used as corpus.

B. Learning Topic Model Using hLDA (Step B)

After acquiring the text source in Step A, we perform the learning of the topic model by using hLDA.

The learned model builds a hierarchical structure, and to construct apath model in Step C, determining whether the hierarchy of which level is incorporated in the path model is necessary. Here, a total number of latent variables of approximately 3 to 10 is desirable because of three reasons. First, a SEM is more likely to fail in the identification of the model when its path model has too much number of latent variables. Second, the reliability is more likely spoiled because data compatibility extremely worsens given the high number of latent variables. Third, this number of variables is desirable to prevent problems that involve too little information of the model for model interpretation or too much information of the model for determining the useful parts. In addition, for the consideration of the model in Step D, estimating what each topic of each hierarchy decided to incorporate (or expressed) in the path model is necessary. We performed the topic estimation artificially by using keyword group.

C. Construction of Path Model (Step C)

We may understand the kind of topic that is expressed by looking at the keyword group that constitutes the learned topic model. The keyword group topic is output in a state that is sorted sequentially to attain a high probability of generationby the learning algorithm. Three high-ranking keywords should be selected, except for the incomprehensible

words. However, we can use all the keywords that constitute the topic when we determine the label of the model. In addition, for the identification problems of SEM, we recommend to avoid a situation in which a repeating word appears and is selected.

TABLE I: 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	7
Model 4	*Fire Emblem: Radiant Dawn*	Yes	9
Model 5	*Super Smash Bros. Melee X*	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 II: 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

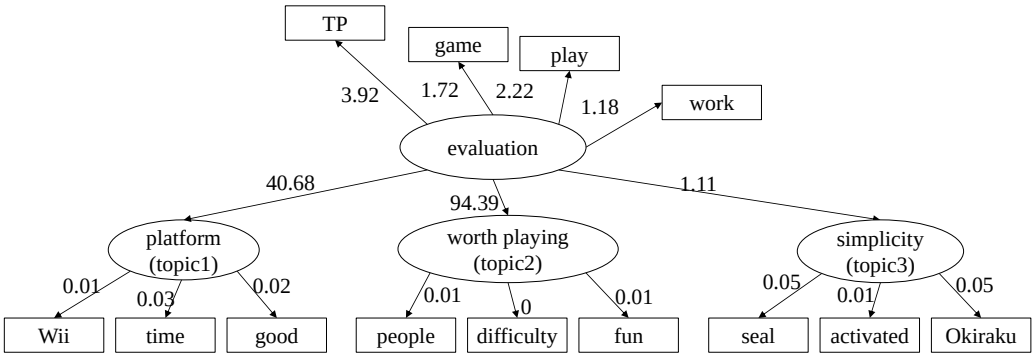

Fig. 4. Visualized causal relationships between topics and keywords of model 12.

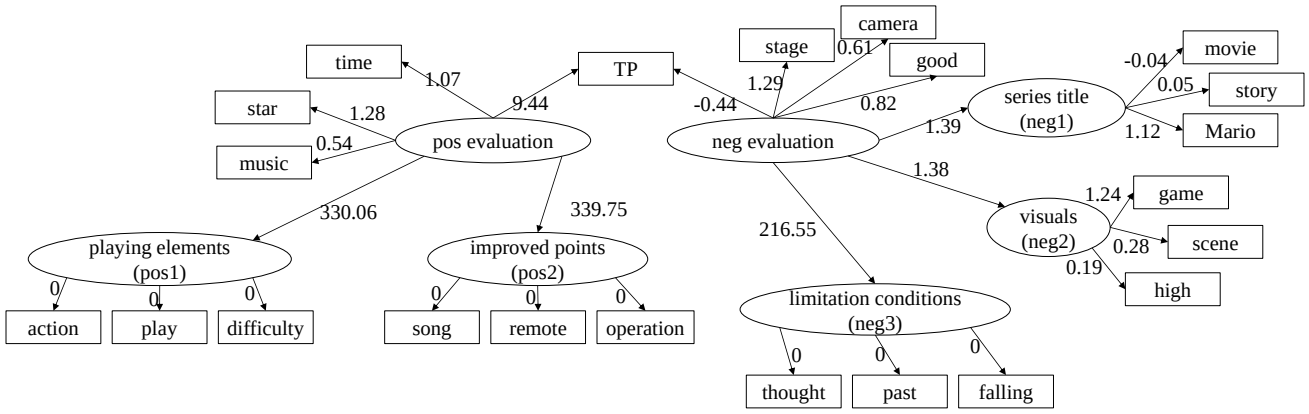

Fig. 5. Visualized causal relationships between topics and keywords of model 3.

In hLDA, each lower-level topic is generated by a higher-level topic. Therefore, setting the path based on the following rules is recommended: 1) drawing the path toward lower-level topics from each higher-level topic, and 2) drawing the path toward each word that constructs a topic from another topic. A clearly identified problem is a precondition for SEM. If we implement the process based on the two rules, then identification problems will not appear in the final stage, and the model is stable. However, if there were any observed variables that an analyst wants to know, then he could add a certain path into the model.

D. Analysis by SEM and Evaluation of Results (Step D)

In this step, we perform SEM analysis using word distribution data, related data acquired together, and the path model that is constructed in Step C.

In SEM, the analysis result indicates that the path model has a calculated contribution degree between each item, as well as some indexes that quantitatively evaluate conformity degree between the model and data or its balance. We used four representative indicators that are usually applied in SEM analysis for reference. The 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; and BIC should be lower than that of 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 analysts with useful knowledge.

We chose a Japan-based game software review site called mk2 [15] as data for this experiment. 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, *Super 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) Extraction (Step a)

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 for every title. The number of datasets is the sum of the five titles (577).

2) Learning and construction (Steps b to c)

We structured the model for multiple interpretability 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 appropriate information quantity model. OR is influenced only by the root topic of the learned topic model.

3) Analysis (Step d)

Steps D should be implemented as mentioned in Chapter IV, which described our proposed process. In this experiment, we used the SEM package supplied in R software [16]-[18] version 12.2.2, a well-known statistical analysis tool. The SEM package of this software provides the source code for using the visualization tool Graph Viz [19], which can present the analyzed model as a figure.

B. Results and Discussion

Experimental results are shown in Table I and Table II, 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," respectively. This topic model shows us how the evaluation point is structured. The abovementioned findings confirm the real relationships of users' evaluation points.

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 2)," "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, pos1and pos2 can be derived as "playing elements," and "improved points," respectively. Similarly, neg1, neg2, and neg3 could be assumed to refer to "series title," "visuals," and "limitation conditions." 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 models with a similar pattern (with categorized comments). 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 relationships 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 an unstructured factor model.

VI. Conclusion

This paper attempted to investigate game software market by using text-based analysis with hLDA 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, which is too complex to analyze effectively.

We proposed a concrete analysis process composed of four steps. Step A involves collecting corpus from the field that the analyst wants to investigate. In Step B, the learning topic model is implemented by using hLDA. In Step C, a path model is constructed for analysis by using SEM. Analysis and evaluation are conducted in Step D.

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 [20] in Step C of the proposed method to improve the precision of our analysis process. Moreover, we will find improved methods of performing Steps B and C 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 25240049, 25420448

References

[1] CAPCOM co. (September 30, 2013). *LTD.: Market Data*. [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] R. Saga *et al*., "Improvement of factor model with text information based on factor model construction process," *IIMSS*, 2013, pp. 222-230.
[5] S. Kawanaka *et al*., "Competitor analysis of consumer situations utilizing topic model," presented at the 25th Annual Conference of the Japanese Society for Articial Intelligence, 2011.
[6] L. Wajima *et al*., "Specific negative factors using latent dirichlet allocation," DEIM Forum, 2014.
[7] S. Deerwester *et al*., "Indexing by latent semantic analysis," *J. Amer. Soc. Info Sci*., 1990, vol. 41, pp. 391-407.
[8] D. M. Blei *et al*., "Latent dirichlet allocation," *The Journal of Machine Learning Research*, 2003, vol. 3, pp. 993-1022.
[9] D. M. Blei *et al*., "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, 2010, vol. 57, no. 2, p. 7.
[10] D. M. Blei *et al*., "Hierarchical topic models and the nested Chinese restaurant process," *Advances in Neural Information Processing Systems*, 2003, vol. 16, pp. 106-114.
[11] J. C. Loehlin, *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*, 4th ed., Routledge, 2004.
[12] J. Pearl, *Causality*, second edition, Cambridge University Press, 2001.
[13] S. Toyokawa *et al*., "Structural equation modeling of the relationship of bone mineral density and its risk factors in Japanese women," *Environmental Health and Preventive Medicine*, 2011, vol. 6, no. 1, pp. 41-46.
[14] MALLET. (2002). A machine learning for language Toolkit. [Online]. Available: http://mallet.cs.umass.edu
[15] mk2 group. [Online]. Available: http://www.psmk2.net/
[16] K. Kita *et al*., *Information Retrieval Algorithms*, 8th ed., Kyoritsu Pub.
[17] The R Project for Statistical Computing. [Online]. Available: http://www.r-project.org/
[18] J. Fox, *Structural Equation Modeling with the SEM package in R. Structural Equation Modeling*, 2006, vol. 13, pp. 465-486.
[19] Graphviz-Graph Visualization Software. [Online]. Available: http://www.graphviz.org/Theory.php
[20] Q. Mei *et al*., "Automatic labeling of multinomial topic models," in *Proc. KDD'07*, 2007, pp. 490-499.

Rikuto Kunimoto was born in Osaka, Japan, 1989. He is pursuing his master's degree in computer science and intelligent systems at the Graduate School of Engineering, Osaka Prefecture University. He is currently engaged in research on knowledge management and text mining with focus on causal analysis and topic model. He is a student member of IEEE.

Hiroshi Kobayashi was born in Mie, Japan, 1991. He is pursuing his master's degree in computer science and intelligent systems at the Graduate School of Engineering, Osaka Prefecture University. He is currently engaged in research on text mining and corpus analysis with focus on keyword extraction. He is a student member of IEEJ.

Ryosuke Saga was born in Tokushima, Japan, 1980. He received his bachelor's degree from Osaka Prefecture University in 2003. He obtained his master's degree and his doctoral degree in electrical and information engineering in 2005 and 2008, respectively. He works as an associate professor at Osaka Prefecture University. He is currently engaged in research on knowledge management, data engineering, and decision support. He is a member of IEEE, etc.