Product improvement opportunity analysis based on text mining of online VOC: application of topic modeling and chance discovery theory

Nam-uk Ko
Department of Industrial Engineering
Konkuk University of, Seoul, Korea
Tel: (+82) 10-5840-4634, Email: kmu1009@konkuk.ac.kr

Janghyeok Yoon †
Department of Industrial Engineering
Konkuk University, Seoul, Korea
Tel: (+82) 2-450-0453, Email: janghyoon@konkuk.ac.kr

Abstract. Voice of customer (VOC) about products is a useful material to set the direction of developing new products and improving current products. Online data posted in social networking services can be considered to be online VOCs because they show plenty of expectations, preferences, opinions and needs that customers think about a certain product. However, this online data are mostly in an unstructured and unorganized textual form, so it is difficult to incorporate them into the product planning processes. Therefore, this paper proposes an approach to identify product opportunities from online VOC by using topic modeling and chance discovery theory. This approach extracts topical issues from large-scale online data related to a specific product by topic modeling, followed by constructing a keygraph based on the co-occurrence relationship among the extracted topics. Then, this approach identifies opportunities by using the breaking points obtained from the keygraph about the product. Our approach is expected to help monitor various types of customer needs expressed in online and thereby generate product improvement ideas.

Keywords: Product improvement opportunity; Social web data; Topic modeling, Chance discovery, KeyGraph

1. INTRODUCTION

Voice of customer (VOC) is a term to describe customer's opinions, needs, preferences and expectations. Due to the increasing importance of customer's feedback, analyzing VOCs become a crucial activity for the success of business activities (Kang, 2016). In customary approaches, VOCs are gathered in an offline form such as telephone surveys, door-to-door interviews and question investigation. However, in recent years, a large number of VOCs have been scattered in an online form, including online sites, forum sites and social network service (SNS). These online VOCs are considered to represent customer's experience, opinion, satisfaction, dissatisfaction in real time.

Product-based companies that create profits through product development should steadily improve their own products or provide new innovative products. Therefore, in setting the direction of developing products and improving current products, it is essential for firms to grasp customer's opinions and views. Thinking from a customer's perspective is to notice the difference between products that companies want and that customers want. For this reason, use of VOCs are increasing for the product innovation process; VOCs are considered the best source to understand customer's view (Franke et al., 2006). Offline VOCs delivered directly to the firm are easy to use in the product planning processes, while online VOCs, as the aggregation of customer's needs, opinions and perspectives, are generally difficult to use because most online data are unstructured and unorganized.

Therefore, this paper proposes an approach to identify product opportunities from online VOCs by using topic modeling and chance discovery theory. The approach is composed of 1) collecting online VOCs related to a specific product, 2) extracting topics from the online data by using

topic modeling, 3) constructing keyword networks based on the co-occurrence relationship among the extracted topics, and 4) identifying product opportunities by using the breaking points obtained from the KeyGraphs about the product.

The suggested method contributes to monitoring various types of customer needs expressed in online and thereby generating product improvement ideas.

2. Theoretical backgrounds

2.1. LDA(Latent Dirichlet Allocation)

The Latent Dirichlet Allocation (LDA) suggested by Blei is a generative statistical topic model that finds latent topics on text-based documents (Blei et al., 2003). In LDA, each document is supposed as a set of various topics and the topic distribution is assumed to have a Dirichlet prior. Then, the probability that keywords in documents are assigned to specific topics is calculated using Dirichlet distribution. Then, we can identify topics and their contributing keywords and the documents related to each topics.

In LDA, M is the number of documents, N is the number of words in document, w is the words, z is the topics, Θ is the topic distribution for document, α is the parameter of Dirichlet prior on the per-document topic distributions, and β is the parameter of the Dirichlet prior on the per-topic word distribution (Figure 1).

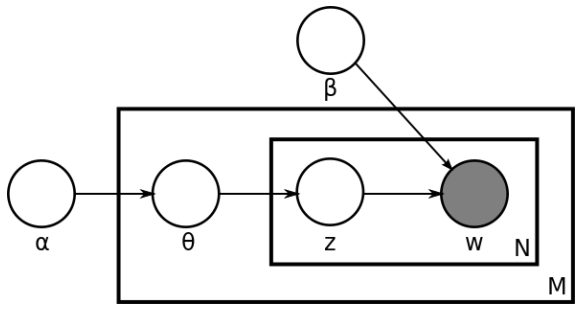

Figure 1. LDA Model

In the model, each node means a random variable, the gray node is an observed variable, and other nodes are latent variables.

2.2 Chance discovery

Chance discovery theory suggested by Ohsawa is a method to identify potential chances (Ohsawa, 2006). In this theory, a chance is defined as an event or a situation that does not frequently occur but is considered relatively significant compared with other events, and thus such event is called a breaking point (Park and Yoon, 2015). KeyGraph, which is used for chance discovery (Ohsawa and Nara, 2002), is a keyword network to visualize text-based documents, and it helps analysts understand the key events (Ohsawa et al., 1998). The six steps to construct KeyGraphs from text-based documents is as follows.

2.2.1. Document preprocessing

A document consists of sentences, which are composed of words. Document D is preprocessed by removing stop-words and relatively low frequency words. Then, D is reduced to document D' composed of valid words (w_i).

2.2.2. Extracting high frequency terms

For each document D', words are arranged in order of their frequency. Then, top N words with a high frequency become nodes (n_{nodes}) in a KeyGraph G.

2.2.3. Connecting nodes

If words co-occur in a same sentence, they can be considered to be associated each other. In the G, a word is a node and an association between a pair of words is a link between the words. Therefore, a measure for co-occurrences of two words (w_i, w_j) is defined as follows.

$$\text{assoc}(w_i, w_j) = \sum_{s \in D'} \min(|w_i|_s, |w_j|_s)$$

2.2.4. Extracting key terms

Key terms are defined as the terms that connect the clusters that contain high frequency terms, and thus a measure to measure the potential as a key term of word w is defined as

$$\text{Key}(w) = 1 - \prod_{g \subset G} [1 - \frac{\sum_{s \in D'} |w|_s |g - w|_s}{\sum_{s \in D'} \sum_{w \in s} |w|_s |g - w|_s}]$$

2.2.5. Extracting key links

A key term has association values (link values). If a link has relatively high association value and connects between two or more clusters, the link becomes a key link.

2.2.6. Extracting keywords

Nodes in G then are sorted by the sum of association values. Then, the top nodes with a high value of the sum of association values are defined as keywords for document D'.

3. Proposed approach

The procedure in this paper is composed of 1) collecting web data related to a specific product, 2) extracting topics from the web data, 3) constructing KeyGraphs and discovering chance topics and 4) identifying opportunities by scenario graphs (Figure 2).

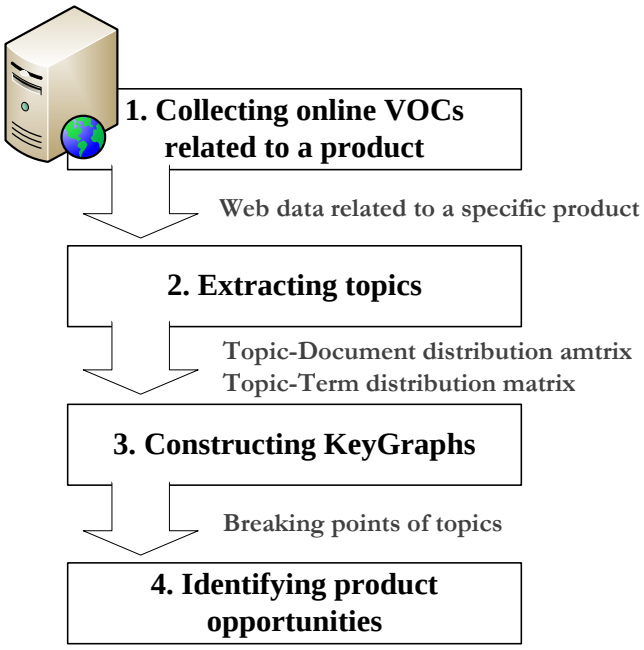

Figure 2. Overall procedure

3.1. Collecting online VOCs related to a product

This paper uses online VOCs data related to a specific product. Online VOCs contain subjective opinions from a customer's view and they are represented in a textual form.

Online VOCs are the collection of texture documents. Each document has various words including garbage words. Therefore, this step extracts words from the set of documents, and removes stop words, meaningless words and low frequency words from set of extracted words. Finally, a set of keywords is prepared.

3.2. Extracting topics

This step identifies the topics from the set of documents. For this process, the LDA model is a useful algorithm for identifying topics contained in a corpus. The LDA defines two input parameters: a document-keyword frequency matrix and the number of topics. The matrix is constructed by word frequencies in each document, and the number of topics is an integer value for extracting topic-document and topic-word distribution matrix. Because the number of topics could not be given automatically in LDA, analysts must set a proper number of topics. In this research, the optimal number of topics is determined by the lowest similarity between topic-word distributions(Kim et al., 2016; Wang et al., 2014). This is because the optimal number of topics well separate from each other.

3.3. Constructing KeyGraphs

This step generates a KeyGraph, which visualizes the relationship between events, is easy to identify breaking points. Analyzing all of numerous VOCs is almost impossible, because it is an extremely inefficient, expensive and time-consuming process. This paper therefore utilizes the chance discovery theory and KeyGraphs. The chance discovery generally needs co-occurrence information of keywords in the same sentence as its input data. Similarly, this study uses a set of topics related to a document as the input data for chance discovery application.

Finally, our approach can identify the various major topics and key topics related to a specific product. Key topics are relatively infrequent but critical topics. In other words, key topics receive relatively less attention from customers but they are potentially critical for customers.

3.4. Identifying product improvement opportunities

The constructed KeyGraphs support identification of product improvement opportunities. A key topic and its connected topics can be used to create scenarios that describe improvement situations from a customer's perspective. Each topic is related to various documents and each document has customer's opinions of a specific product. For the identifying opportunities, this research extracts opportunity graphs, which are subgraphs in each KeyGraph. An opportunity graph is composed of one or more key topic and their related topics. Related topics are the nodes directly connected or closely located to a key topic. Therefore, an opportunity graph plays a role of guidance to identify the opportunities implicitly described in large-scale online data.

4. Case Study : Galaxy Note 5

This paper proposes an approach to identify improvement opportunities related to a specific product from online VOCs. To apply this approach, a target product has various topics. Because the product is simple or has fewer features, customer's voices are about the same. Moreover, the product is recent as possible. Online VOCs are real-time data. The more the product is recent, the more customers mention it on the online. For this reason, this research set a target product is Galaxy Note 5. Because, Galaxy Note 5 is one of the newest smartphones, and it has various components, attachments and functions.

4.1. Data

For the collecting online VOCs, Reddit is a useful social news networking service website. In this site, users can not only submit contents such as news texts or news links, but also vote that articles and comments. Reddit's contents are divided into numerous categories called "Subreddit". A subreddit is not only comprehensive categories such as education, technology, Food, but also, concrete categories that are specific product or service.

In this paper, therefore, online VOCs are gathered from subreddit Galaxy Note 5 until January 2016, and then the collected data eliminate noise data. Finally, 11,123 documents, 3,539 keywords are collected from subreddit Galaxy Note 5.

4.2. Topic extraction and KeyGraph construction

For using LDA, the number of topics should be selected. In order to find the optimal number of topics, we extract similarity between topic-word distributions by changing the number of topics. The optimal number of topic, which has the lowest similarity, is 65 that have a similarity score of 0.0455. Parts of similarities are as follow (Table 1).

Table 1. Parts of similarities by number of topics.

# of topics	similarity	# of topics	similarity
30	0.0812	64	0.0496
40	0.0712	**65**	**0.0455**
50	0.0643	66	0.0563
60	0.0568	67	0.0511
61	0.0502	68	0.0566
62	0.0548	70	0.0538
63	0.0620	80	0.0624

Because LDA is not represent label of topics, analysts would select a meaningful label using keywords that are related to each topics. Topic labeling, which support to discovery latent topics, is a significant process for identifying opportunities. For example, if keywords are "pen", "pens", "sensor", and "pencil", analysts could select topic label is "detect pen". Then, analysts easy to look at documents, are related to the topic "detect pen", emphasis on the word "detect pen". Parts of topic labels and related keywords are as follow (Table 2).

Table 2. Parts of topic labels and related keywords.

Topic Label	Keywords
Detect pen	pen, pens, sensor, pencil
Wifi	Wifi, Wi-Fi, network, Speed, speeds
Expandability	SD, SD-Card, SD card, MicroSD
Physical button	button, buttons, home button, power button, volume button
Charge cable	Charge, cable, charger, fast charge, USB-C, cables

Topic-documents distribution matrix is constructed by related documents in each topics. Because a document is related to all topics, analysts should be cut-off the relationship between topics and documents. However, no algorithm yet set the optimal threshold value in various cases. Moreover the proposed method, which is support method to generate various concepts, need to apply various threshold. Therefore, this paper uses two thresholds that are 0.02 and 0.03.

Polaris, which is a free software for a KeyGraph construction, was used to generate KeyGraphs. Because KeyGraph vary depending on parameters, constructing process need to do sensitivity analysis. By trial-and-error, we found a KeyGraph, which has 15 number of topics, each threshold.

4.3. Identifying opportunities

In the constructed KeyGraph, red nodes are chance topics, and black nodes are major topics. The opportunities could be identified as a collection of nodes, which are gathered around red nodes. This collection, which was a subgraph of KeyGraph, called opportunity graph. Each KeyGraph has more than zero opportunity graphs. In other words, Each KeyGraph has more than zero opportunities of product improvement. Two KeyGraphs and their opportunity graphs are describe Figure 3 and Figure 4.

First KeyGraph, which has threshold value that is 0.02, has three opportunity graphs. First opportunity graph has one chance node (Case) and one black node (Warranty&Repair). A review of documents, which are related to this opportunity graph, reveals waterproofing problem. We ideated about a customer scenario of "A customer works at a restaurant kitchen.

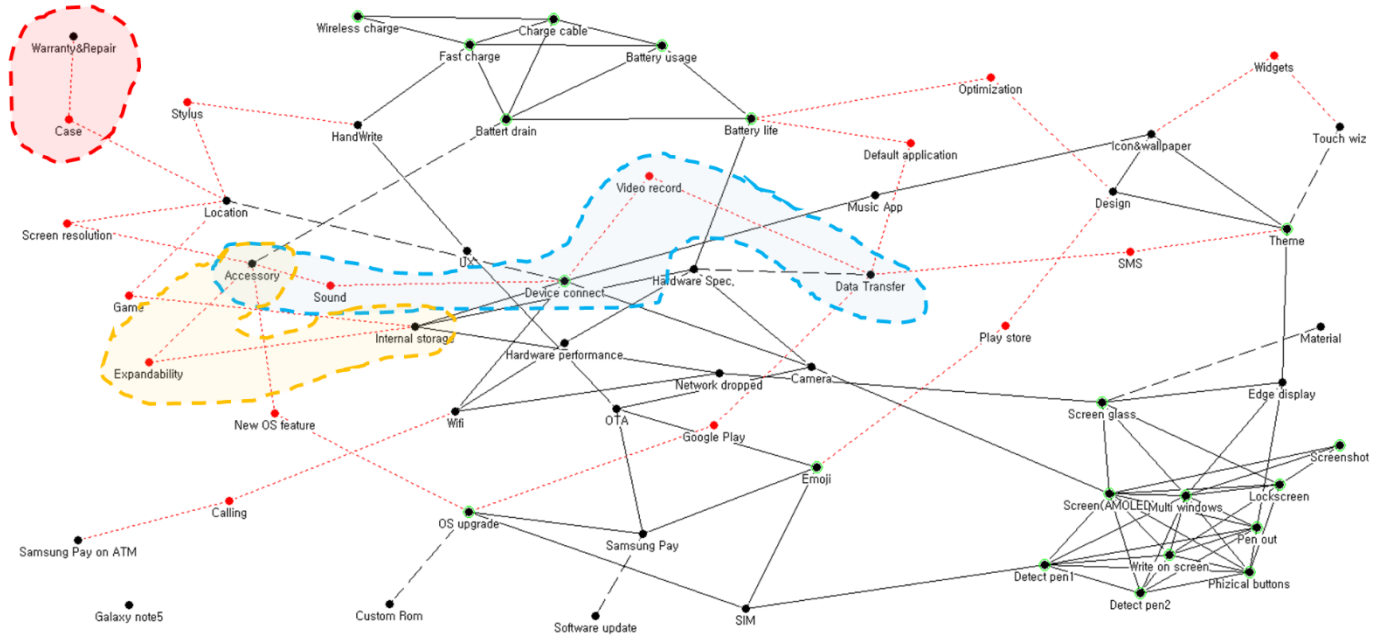

Figure 3. KeyGraph (threshold=0.02)

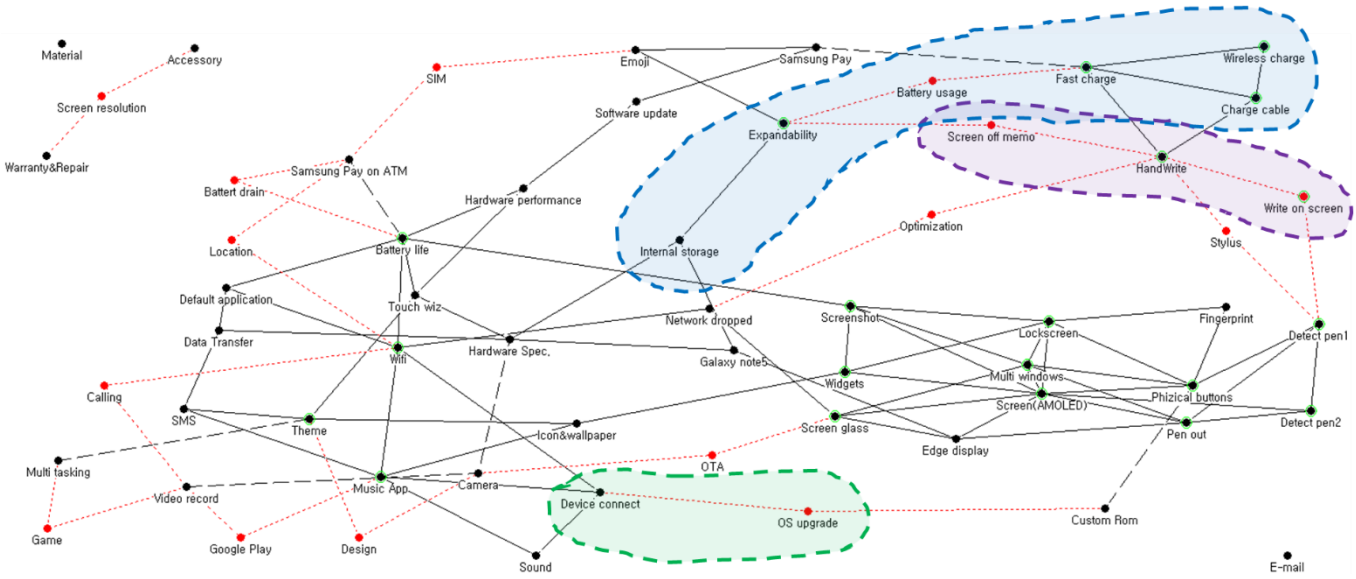

Figure 4. KeyGrpah (threshold=0.03)

Naturally, he spends a lot of time on the kitchen. He always worry about he might drop his cell in water." As an opportunities for this scenario, "develop waterproof cases" and "Mount waterproof function" would be drown. Actually, Galaxy Note 5 is not waterproof. Therefore, many customers need to against water-repellent. Second opportunity graph has two chance nodes (Sound, Video record) and three black nodes (Accessory, Device connect, Data transfer). This area's VOCs explain that Galaxy note 5 has Bluetooth problem. By using the problem, we would ideate about a customer scenario of "A customer took a long drive on Monday. While driving, he did not listen to the music through his car speakers. Because Bluetooth connection of his cell was bad." Actually, many cases that are related to Bluetooth problem of Galaxy Note 5 are appearing. If this problem is most likely in hardware, Samsung needs to examine this issue. Therefore, this problem

Table 3. Opportunities of Galaxy Note 5 from KeyGraph (threshold=0.02)

Topics in opportunity graph	VOCs	opportunities
Case, Warranty&Repair	- About waterproofing : A customer works at a restaurant kitchen. Naturally, he spends a lot of time on the kitchen. He always worry about he might drop his cell in water.	-Develop waterproof cases -Mount waterproof function.
Accessory, Sound, Device connect, Video record, Data transfer	- About Bluetooth sound problem : A customer took a long drive on Monday. While driving, he did not listen to the music through his car speakers. Because Bluetooth connection of his cell was bad.	-Examine Bluetooth and speaker sound
Accessory, Expandability, Internal storage	- About SD card : A customer loves to take a picture using her cell. As she's cell is almost full, she needs more storage memory	-Extend memory -Develop external memory device

Table 4. Opportunities of Galaxy Note 5 from KeyGraph (threshold=0.03)

Topics in opportunity graph	VOCs	opportunities
Internal storage, Expandability, **Battery usage**, Fast charge, Charge cable, Wireless charge	- About battery : Customers have access to their cell at all times of the day. They feel battery level is too low	- Develop battery case - Extend battery capacity - Examine removable battery.
Screen off memo, Handwrite, **Write on screen**	- Desire to compatibility other note apps : A customer use screen off memo often. However, she has to transfer the memo to other note app that she used.	- Examine compatibility apps - Examine controlling memo size
Device connect, **OS upgrade**	- Interlock Mac or iPhone : A customer is going to purchase a Galaxy Note 5. She worried about how to transfer data in iPhone, which she used, to Note 5.	- Develop interlock software

can be an opportunity of Galaxy Note 5 improvement.

Second KeyGraph, which has threshold value that is 0.03, has three opportunity graphs. First opportunity graph has one chance node (Battery usage) and four black nodes (Internal storage, Expandability, Fast charge, Charge cable, Wireless charge). VOCs in this graph are describe to explain the importance of battery. By using them, we would ideate about a customer scenario of "Customers have access to their cell at all times of the day. They feel battery level is too low." As opportunities for this scenario, "Develop battery case", "Extend battery capacity", and "Examine removable battery" would be drown. Second opportunity graph has two chance node (Screen off memo, Write on screen) and one black nodes (Handwrite). One of the most notable features on Galaxy Note 5 is Screen Off memo function, which is allow users to write something when the screen is off. By using them, we ideated about a customer scenario of "A customer use screen off memo often. However, she has to transfer the memo to other note app that she used." As a result, a need to examine compatibility other note app was found.

5. Conclusion

This paper proposes an approach to identify product opportunities from online VOC by using topic modeling and chance discovery theory. Because the increasing importance of customer's feedback, VOCs become a crucial data for the success of business activities. For identifying opportunities of a specific product from online VOC, Reddit data were used in this research. The Reddit data related to a specific product, included customer's needs, opinion, satisfaction, and dissatisfaction of the product, are collection of texture documents. Then, we could identify topics from the set of documents using LDA. These identified topics are the node of a KeyGraph that is visualizes the relationship between topics.

The constructed KeyGraph support identification of product improvement opportunities. The suggested method contributes to monitoring various types of customer needs expressed in online and thereby generating product improvement ideas.

Despite the contribution, this paper, there still exist challenges. First, this paper used only Reddit data. Therefore, Reddit as well as blogs and other social sites as the materials for the proposed approach will achieve outcomes that are more interesting. Second, the proposed method involves experts' intervention. Analysts should set the keyword sets, threshold values, and KeyGraphs. Therefore, we could find a method of reducing experts' intervention.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2015R1A1A1A05027889)

REFERENCES

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Franke, N., Von Hippel, E., & Schreier, M. (2006). Finding commercially attractive user innovations: A test of lead-user theory. Journal of product innovation management, 23(4), 301-315.

Kang, E.-J. S. a. M.-S. (2016). Extraction of Risk Factors Through VOC Data Analysis for Travel Agencies.

Kim, M., Park, Y., & Yoon, J. (2016). Generating patent development maps for technology monitoring using semantic patent-topic analysis. Computers & Industrial Engineering.

Ohsawa, Y. (2006). Chance discovery: The current states of art. In Chance discoveries in real world decision making (pp. 3-20): Springer.

Ohsawa, Y., Benson, N. E., & Yachida, M. (1998). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. Paper presented at the Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on.

Ohsawa, Y., & Nara, Y. (2002). Modeling the process of chance discovery by chance discovery on double helix. Paper presented at the Proc. of AAAI Fall Symposium on Chance Discovery.

Park, H., & Yoon, J. (2015). A chance discovery-based approach for new product–service system (PSS) concepts. Service Business, 9(1), 115-135.

Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. Scientometrics, 101(1), 685-704.