Proposal of Chance Index in Co-occurrence Visualized Network

Yukihiro Takayama and Ryosuke Saga

***Abstract*—This study describes a chance discovery method for network that use betweeness centrality and similarity. In prior research of chance discovery, in the chance discovery process, it is required that analysts infer chance from visualized network, because it is difficult that to solve problem like to guess the cause from the data such as non-parametric problem. However, this reasoning process has problem that chance discovery is difficult because chance discovery depends on experience or background knowledge of analysts. Therefore, to solve this problem, we pay attention the mathematical element with the network, and propose chance index that is index of network. Chance index have three calculation methods: the sum of the reciprocal, the product of the reciprocal, and the average reciprocal. Using the proposal method on three kinds of data, results show that proposal method is useful method and chance index that use average reciprocal is most useful calculation method.**

***Index Terms*—Network analysis, betweenness centrality, chance discovery, data mining**

I. Introduction

With the development of information-based societies, increasingly large quantities of data are being stored. As a result, data mining is an attractive way to analyze data, extract knowledge from data, and perform analysis and prediction [1]-[3]. However, in recent years, it has been found that analysts cannot effectively use the knowledge extracted by data mining. Research into chance discovery has been conducted to facilitate effective use of such knowledge. For example, in the case of a store, the purpose of an analyst is to generate profit by analyzing product sales or customer data. However, even if analysts can discover knowledge by data mining, such knowledge is of no use if it cannot be connected to profit opportunities. It can be said that the real purpose of an analyst is to discover a chance rather than knowledge; thus, chance discoveries are important.

Analysts infer chance from visualized data. Visualization is a data mining method. By visualizing data, it is possible to observe data from a different perspective, and as such, various inferred chances can be observed. Visualization is essential to chance discovery. However, in the inference process, chance discovery depends on the analyst because there are individual differences in the experiences and background knowledge among analysts; therefore, various methods are required to catch the data.

KeyGraph is a useful visualization method for chance discovery [4], [5]. KeyGraph simplifies chance inference by devising a graphical representation when visualizing data. However, even though analysts can perform chance discovery more effectively by using a graphical representation, this method has not been able to eliminate the problem of analyst dependence. We focus on a network to solve this problem. We perform mathematical interpretation using graph theory in the network. For example, by using graph theory that exploits an index of nodes and links, we can express a network quantitatively. In addition, graph theory can be applied to a shortest path problem and the traveling salesman problem.

In this study, rather than performing chance discovery by inferring chance from the visualization results for a co-occurrence network, we perform chance discovery using only visualization. For this purpose, we propose a chance index, which is an index of the nodes in the network. To represent this chance index, we use the similarity of links in the network and betweenness centrality, which is an indicator of the centrality of a given node [6].

The remainder of this paper is organized as follows. In Section II, we explain chance discovery. In Section III, we describe related work. Section IV introduces the algorithm used in the proposed method, and Section V presents experimental verification of the proposed method. Conclusions are presented in Section VI.

II. Chance Discovery

A chance is an important event or circumstance that can be used by analysts for decision-making [7]. Using chance discovery, analysts can evaluate the development of a market or predict typhoons by capturing changes [8], [9].

There is a knowledge discovery in databases process in data mining [10]. Analysts can obtain various large data easily; therefore, analysts often use data mining to gain knowledge. However, the purpose of the analyst is not knowledge discovery from data mining. The purpose of an analyst is to understand the causes and background of the gained knowledge and to utilize this knowledge. This is chance discovery.

For example, by analyzing sales data, analysts can identify products that are selling well. This is knowledge discovery. However, this knowledge alone does not result in profit for the store directly. Analysts elucidate the reasons the product sold well and apply those factors to other products. It should be noted that, in this example, data analysis only has meaning or value when analysts can use the obtained knowledge to increase sales. The causality associated with a product that sells well is considered chance. Thus, chance discovery is the final purpose of analysts.

Manuscript received August 12, 2014; revised December 17, 2014.

The authors are with the Osaka Prefecture University, 1-1-1, Gakuen-cho, Naka-ku, Sakai-shi, Osaka-fu, 599-8531, Japan (e-mail: saga@cs.osakafu-u.ac.jp).

DOI: 10.7763/IJKE.2015.V1.14

III. Related Work

Essentially, the final stage of the chance discovery process is analyst inference, because with current computer technology, it is difficult to solve problems like determining a cause from data, such as a non-parametric problem. Thus, analysts perform chance discovery by inferring chance from visualized data.

Visualization is "By making use of a computer, the visual representation of the data in order to expand the recognition Interactive" [11]. Expansion of recognition is the value of visualization. Visualization is used when analysts want to obtain new knowledge. Recently, several visualization methods have been proposed.

For example, the frequency and co-occurrence trend FACT-Graph visualization method has been proposed to visualize keyword trends embedded in documents. FACT-Graph visualizes data by combining class transition information and co-occurrence transition information [12]. In addition, to assist in the interpretation of genome-scale datasets by facilitating the transition from data collection to biological meaning, a visualization system called the database for annotation, visualization, and integrated discovery has been proposed [13]. Furthermore, classification of information visualization and visual data mining methods are based on the data type to be visualized, the visualization method, and the interaction and distortion method. These information visualization methods have been developed over the past decade to support searching large datasets [14]. However, in knowledge discovery, analysts do not always know where they should focus their attention in the visualization results. The knowledge discovery process depends on analyst inference, which is based on the background or experience of a given analyst. The chance discovery process requires chance inference from visualized data. Therefore, chance discovery has the same problem as visualization.

KeyGraph is a popular chance discovery method [4], [5]. KeyGraph is a data visualization algorithm that uses interactive parameters. By changing the color and shape of nodes and links, KeyGraph can identify chance easily. Fig. 1 shows a KeyGraph example. As can been seen, the greater the frequency of a word, the more the shape of a given node changes, and the shape of the link changes by co-occurrence of nodes.

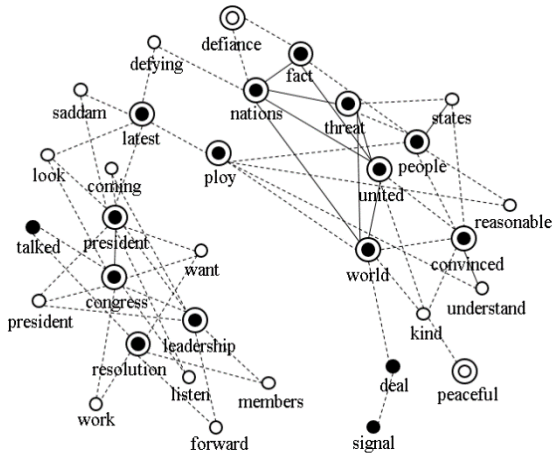

Fig. 1. KeyGraph example (President bush press conference network [15]).

Thus, by devising a way to present visualization results to analysts, KeyGraph makes discovery of chance easy.

However, even if such a method can make chance discovery easy, chance discovery visualization methods still depend somewhat on the tacit or background knowledge of the analyst.

Therefore, in chance discovery process, we consider chance discovery that does not depend on the subjectivity of the analyst, and we focus on a visualized network. By putting data into the network, we can use the mathematical characteristics of the network. By using the mathematical characteristics of the network, we attempt to represent chance in the network.

We propose a chance index that can be used to discover chance quantitatively in a visualized network. Note that the chance index is an index of a node.

We propose the chance index as an index of a node in the network to perform chance discovery without requiring analyst inference. Here, we explain the chance index of the proposed method.

IV. Chance Index

A. Approach

We developed two hypotheses from the features of chance in a network to represent chance without requiring analyst inference.

A: Nodes with small similarity to adjacent links are considered chance.

B: Nodes with large betweenness centrality are considered chance.

Hypothesis A is derived from various chance discovery studies. For example, one study extracted claims from a press conference by President Bush on September 18, 2002 [15]. Fig. 1 shows the results of visualization. That study claimed that "latest" or "ploy" are key points and chance in the network structure. When we considered the position of the chance node in the network, we found that the similarity between the chance node and adjacent links is small.

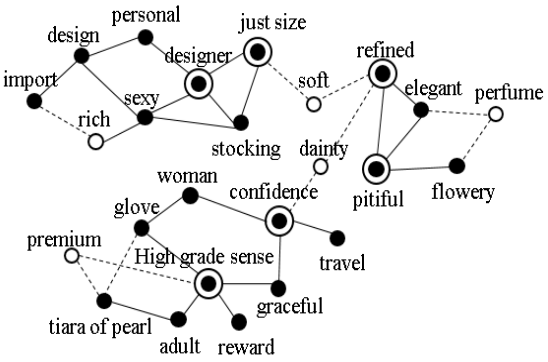

Fig. 2. Visualization results for a questionnaire about women's accessories [16].

For hypothesis B, it can be considered that a node that connects many nodes affects many nodes. For example, a node that connects a set of nodes that represents customer needs bridges two needs. By focusing on this bridge node, analysts may get a chance to attract new customers. Therefore, nodes with high betweenness centrality are considered

chance.

To verify hypothesis B, we surveyed various papers about chance discovery. For example, one study visualized a free descriptive questionnaire about the women's accessories [16]. Fig. 2 shows the results of visualization. That study claimed that "soft" or "dainty" are important key points and chance in the network structure. When we considered the position of the chance node in the network, we found that the chance nodes bridge many nodes.

B. Chance Index

The chance index is expressed by formula (1).

$$chanceindex(i) = C_b(i) + ls(i) \tag{1}$$

Here, $C_b(i)$ is the normalized betweenness centrality of node i ,and $ls(i)$ is the normalized total similarity of links that are adjacent to node i . $C_b(i)$ and $ls(i)$ aresummed after normalization (0–1). Therefore, the chance index can range from 0 to 2.

Formula (2) normalizes the betweenness centrality of node *i*. Here, $C_b'(i)$ is non-normalized $C_b(i)$.

$$C_b(i) = \frac{C_b'(i) - \min_{i \in S} C_b'(i)}{\max_{i \in S} C_b'(i) - \min_{i \in S} C_b'(i)} \tag{2}$$

Let the shortest path from node *j* to node *k* be $g_{jk}(i)$, and the sum of the paths that pass node i in the shortest paths from node *j* to node *k* be $C_b'(i)$. Thus, betweenness centrality of node $i$can be expressed by formula (3).

$$C_b'(i) = \sum_{i \neq j \neq k}^{n} \frac{g_{jk}(i)}{g_{jk}} \tag{3}$$

We considered three methods to calculate the similarity of an adjacent link: the sum of the reciprocal, the product of the reciprocal, and the average reciprocal. In the proposed method, we considered the synthesis method of summing $C_b(i)$ and $ls(i)$, because even if the synthesis method is a product, given the log, the ranking results do not change. However, relative to the visibility of results, a product synthesis method may be better; therefore, an evaluation of optimal synthesis methods will be a focus of future work.

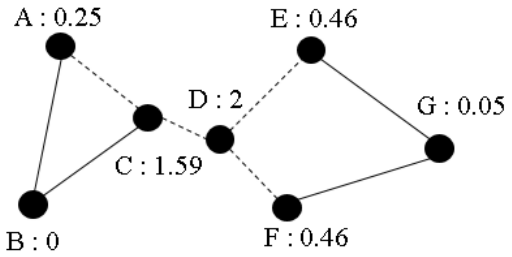

Fig. 3. Chance index example.

Fig. 3 shows an example calculation result of the actual chance index. Note that an alphabet for a node and the value of the chance index are shown. We calculated the similarity of solid line as 1 and the similarity of dotted line as 0.5. Here, the similarity calculation method is the sum of the reciprocal. In this example, nodes "C" or "D" are considered chance. Table I shows the correct calculation results.

TABLE I: EXAMPLE CALCULATION RESULTS

	A	B	C	D	E	F	G
similarity	3	2	5	6	3	3	2
betweenness	0	0	16	19	4	4	1
chance index	0.25	0	1.59	2	0.46	0.46	0.05

We propose three calculation methods for $ls(i)$. However, we have not determined which method is optimal. Therefore, in our experimental verification, we validated the usefulness of the proposed method and the optimum calculation method for $ls(i)$.

V. VERIFICATION EXPERIMENT

A. Verification Experiment Environment

We performed experiments to verify the usability of the chance index and determine the optimum calculation method for $ls(i)$. We applied the proposed method to KeyGraph networks with known nodes to verify if the chance nodes can be extracted by the proposed method. In the verification experiment, we used a network that has been used in other studies. We verified whether the chance index could compensate the analyst inference process from visualized KeyGraph results.

We set the similarity of links for weak links (dotted line) to 0.1–0.3, medium strength links (dark dotted line) to 0.4–0.6, and strong links (solid lines) to 0.7–0.9. We employed three calculation methods: the sum of the reciprocal, the product of the reciprocal, and the average reciprocal.

TABLE II: EXPERIMENTAL DATA

	Bush network	Questionnaire network	Interview network
Nodes	31	25	65
Links	31	33	104

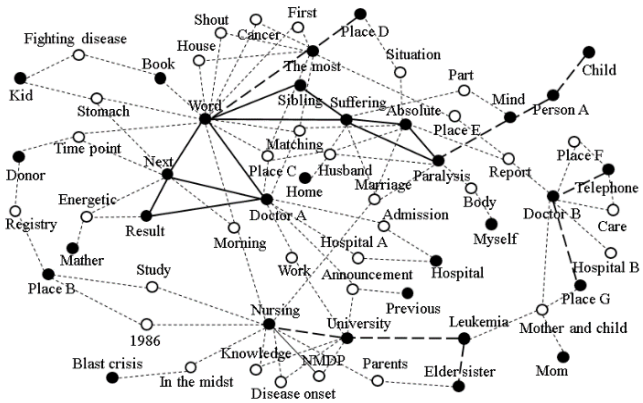

Fig. 4. Visualization results for interview data from medical patients [17].

For network data, we used the President Bush press conference data (Bush network; Fig. 1) [15], questionnaire data about women's accessories (Questionnaire network; Fig.

2) [16], and interview data from medical patients (Interview network; Fig. 4) [17]. Table II shows the details of the data.

B. Results of Verification Experiment

Fig. 5 to Fig. 7 show the results of the verification experiments. The white nodes are chance nodes, and the black nodes are normal nodes.The vertical axisis the chance index value, and the horizontal axis is the number of nodes in the network.

Fig. 5 shows the results of the verification experiments for the Bush network. Globally, we could detect chance nodes in high ranking. These results are considered relatively good. In particular, the results obtained using the average reciprocal are considered the best because chance nodes were detected in the most significant position.

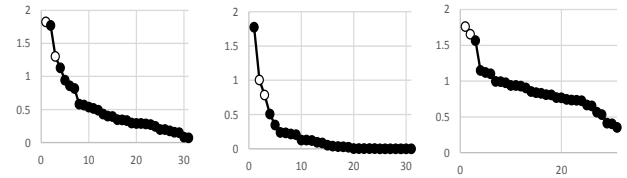

Fig. 5. Results of the verification experiment for the Bush network. From left, results obtained using the sum of the reciprocal, the product of the reciprocal, and average reciprocal.

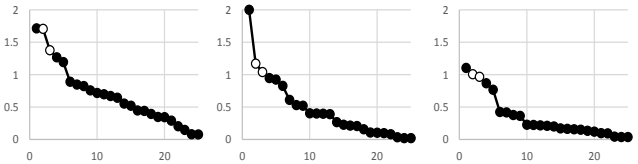

Fig. 6. Results of the verification experiment for the questionnaire network. From left, results obtained using the sum of the reciprocal, the product of the reciprocal, and average reciprocal.

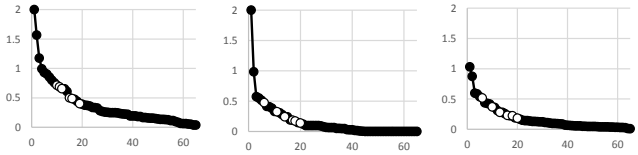

Fig. 7. Results of the verification experiment for the interview network. From left, results obtained using the sum of the reciprocal, the product of the reciprocal, and average reciprocal.

Overall, with this result, we could detect the chance nodes in high ranking. The results obtained using the average reciprocal are considered the best results, however, in the results obtained using the sum of the reciprocal and the product of the reciprocal, other nodes were detected in high ranking. In particular, "President" was detected in high ranking. This word was detected in high ranking because it is used as the subject and is used with other words simultaneously; thus, the betweenness centrality was high. In addition, the chance index of all chance nodes obtained using the average reciprocal was greater than 1.5. Thus, the average reciprocal demonstrated good results even in chance index values.

Fig. 6 shows the results of the verification experiments for the questionnaire network. Globally, we could detect chance nodes in high ranking. This result is considered relatively good. In particular, the results obtained using the average reciprocal are considered the best.

Generally, we could detect chance nodes in high ranking. As mentioned, the results obtained using the average reciprocal are considered the best. However, in the results, "refined" was detected in high ranking because it is used as the subject and is used with other words simultaneously; thus, betweenness centrality was high. In addition, the chance index of all chance nodes obtained using the average reciprocal was approximately 1. Therefore, for the chance index, the results obtained by the other calculation methods are also good.

Fig. 7 shows the results of the verification experiments for the interview network. Globally, we could not detect chance nodes in high ranking. The results obtained using the average reciprocal demonstrate the most detected chance nodes.

With this network, we could not detect chance nodes in high ranking. In particular, words such as "word" and "nursing" were detected in high ranking, because in the medical field, words that everyone uses is many and the words were used with other words like "President." The chance index for all chance nodes obtained by all calculation methods was approximately 0.5. Thus, these results, including the obtained chance index values, are considered poor.

However, when we checked the real network, we discovered that chance nodes connect clusters. Therefore, to improve the proposed method, a new formula must be added.

In summary, it was determined experimentally that using the average reciprocal is most useful. However, the proposed formula is still under development; therefore, various improvements are required.

VI. Conclusion

In the study of chance discovery, analysts infer chance by observing data visualized by a chance discovery process. In this process, chance discovery depends on the background or tacit knowledge of the analysts. To solve this problem, we have focused on the mathematical elements of the network and have proposed a chance index that can be used to extract chance without requiring analyst inference. We have proposed three calculation methods for the chance index; i.e., the sum of the reciprocal, the product of the reciprocal, and the average reciprocal. We have validated the usefulness of the proposed method. The results of verification experiments have shown that the most useful calculation method for the chance index is the average reciprocal method. In future, we plan to add new formulas to the proposed method.

Acknowledgment

This research was supported by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research(A) and (C), 25240049, 25420448.

We acknowledge the support of the ICOM Foundation, Japan, in the form of an International Travel Grant.

References

[1] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification," *International Journal of Computer Science and Information Security*, pp. 136-140, 2012.

[2] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, vol. 1, pp. 12-23, 2000.

[3] R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," *Very Large Databases*, pp. 144-155, 1994.
[4] Y. Ohsawa, N. E. Berson, and M. Yachida, "KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor," *The Institute of Electrical and Electronics Engineers*, pp. 22-24, 1998.
[5] Y. Ohsawa, *KeyGraph: Visualized Structure Among Event Clusters*, Springer-Verlag, pp. 262-275, 2003.
[6] M. E. J. Newman, *Networks An Introduction*, Oxford University Press, 2010.
[7] T. B. Ho and D. D. Nguyen, "Chance discovery and learning minority classes," *New Generation Computing*, vol. 21, pp. 149-161, 2003.
[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. the 20th International Conference on Very Large Data*, pp. 487-499, 1994.
[9] A. Kitamoto, "Spatio-temporal data mining for typhoon image collection," *Journal of Intelligent Information Systems*, vol. 19, pp. 25-41, 2002.
[10] U. Fayyad, G. Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communication of the ACM*, vol. 39, pp. 27-34, 1996.
[11] M. O. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization*, 2010.
[12] R. Saga, M. Terachi, and H. Tuji, "FACT-graph: Trend visualization by frequency and co-occurrence," *Electronics and Communications in Japan*, vol. 95, pp. 50-58, 2012.
[13] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "DAVID: Database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, p. 3, 2003.
[14] D. A. Keim, "Information visualization and visual data mining," *The Institute of Electrical and Electronics Engineers*, vol. 8, pp. 1-8, 2002.
[15] Y. Ohsawa, *Information Technology of Chance Discovery*, Tokyo Denki University Press, 2003.
[16] Kozo Keikaku Engineering Inc. (August 25, 2014). [Online]. Available: http://www2.kke.co.jp/keygraph/example.html
[17] M. Takita, Y. Tanaka, Y. Kodama, N. Murashige, N. Hatanaka, Y. Kishi, T. Matumura, Y. Ohsawa, and M. Kami, "Data mining of mental health issues of non-bone marrow donor siblings," *Journal of Clinical Bioinformatics*, vol. 1, 2011.

Yukihiro Takayama was born in Japan on January 14, 1990. He received his bachelor's degree from Osaka Prefecture University in 2009. Currently, he has joined master's course in the electronic and information engineering in Osaka Prefecture University.

Ryosuke Saga received his 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.