Situational Analysis Based on Graph Structuralization

Taketo Matsunaga[1,2], Koji Kitamura[2], Yoshifumi Nishida[2], Hiroshi Takemura[1]

[1] Department of Mechanical Engineering, Tokyo University of Science, Chiba, Japan
[2] Digital Human Research Center,
National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Abstract—Despite the pervasiveness of situation data such as incident reports and situation reports, we lack methods for describing and analyzing situation data. In this research, the authors propose a new system for a situational structure analysis by formulating the problem of situation mining as a problem of graph structural analysis. The proposed system consists of two basic functions: a function for graph structuralization and a function for situation data mining. The graph-structuralizing function allows users to create a semantic graph from free-description sentences about a situation. The data mining function allows users to conduct clustering, search similar situations, and visualize typical situation processes. To evaluate the effectiveness of the developed system, we analyzed 818 child-bicycle accident data.

Key Words: Situational Analysis, Graph Structural Analysis, Data Mining, Information Search

1 Introduction

In recent years, an enormous volume of situation reports related to injury has become publicly available. Knowledge creation from such a large number of reports is strongly required for a scientific approach to injury prevention, risk management, consumer product improvement, and risk communication [1][2]. Knowledge creation from serious and relatively rare accidents such as airplane crashes, plant accidents, and power plant accidents has been studied in conventional research. For example, as a pioneering work, a failure knowledge database was created and is available in Japan [3][4]. However, we still lack a good methodology for dealing with the knowledge creation of accidents that are individually relatively small in scale but very large in number, such as childhood injuries. For these types of accidents, the total scale becomes very large. According to the world report on childhood injuries published by the World Health Organization (WHO) in 2008 [5], unintentional injury is a major killer of children under the age of 18 and is responsible for approximately 950,000 deaths per year.

Finding typical situations by using situation reports, which are given as text data, and counting the frequency of each typical situation is one of the most important steps for analyzing a situation. However, it is difficult to accomplish this task by a conventional keyword search method or a text mining system. For example, when trying to find the number of situations of "a child rode a bicycle," the keyword search gives us the results of irrelevant situations such as "a mother rode a bicycle while her child was sitting in the back seat" because the keyword search system searches for text that includes "child," "rode," and "bicycle."

The second problem lies in finding the typical process of a situation. To prevent injury, we have to clarify situational structures and find the factors calling for intervention. In this paper, a situational structure indicates two kinds of structural data: the relationship among factors such as environment, consumer products, and persons, and the process of time-series change of the relationship among the factors. The time-series change means that, in general, the relationship among factors changes before the incident, during the incident, and after the incident. So we have to clarify not only the relationship among factors but also the time-series process of the relationship to intervene and control the situation so as to prevent incidents. However, no good technologies exist to support this task.

To solve these problems, this study applies a method for a graph structural analysis to a situational analysis. Technologies for a graph structural analysis [6][7] are available and have been applied to fields such as social networks [8][9], bioinformatics, and molecular structure analysis [10]. Analyzing tools and visualizing tools have also been developed [11]. If we can structuralize incident situation data as graph data, we can formulate the problem of situation mining as a problem of structural analysis.

This paper proposes a situation mining system that allows a user to find situational structures based on a method of graph structuralization, to cluster situations based on the structuralized situational data, and to visualize a typical situation process using the large amount of text data on the situation. To evaluate the effectiveness of the system, we analyzed the real data of 818 child-bicycle incidents.

2 Development of Situational Analysis System Based on Graph Structuralization

Figure 1 shows the configuration of the developed system. The developed system has two basic functions: a function for graph structuralization and a function for situation data mining. The first function, graph structuralization, enables the creation of semantic graphs from free-description sentences about an incident situation. Specifically, the system has functions for graph structuralization of situation data, management of a situation graph database, and management of a domain-specific terminology dictionary. The situational analysis allows us to conduct situation semantic search, situation clustering,

situation linkage analysis and visualization of a typical situation process.

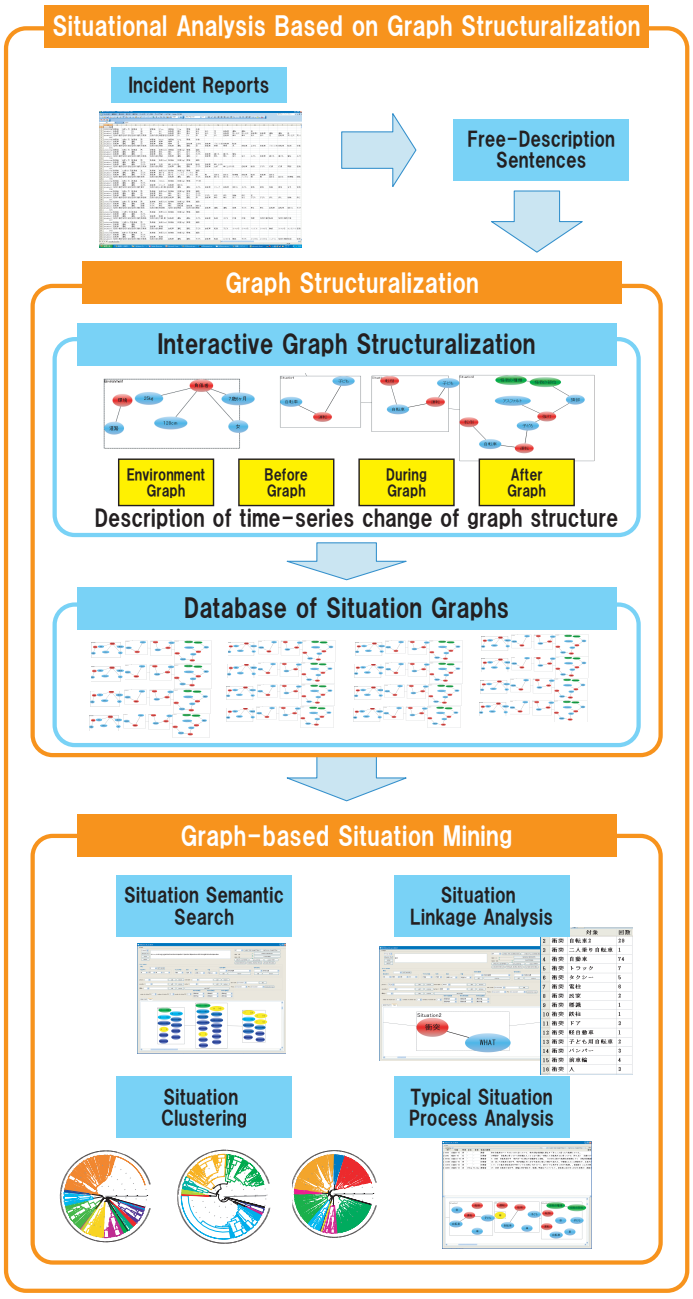

Fig. 1. Configuration of the developed situation mining system based on graph structuralization

2.1 Graph Structuralization

2.1.1 Interactive Graph Structuralization

A user can conduct graph structuralization interactively by using the developed system. We utilize MeCab [12], which is a text mining software, to support the user's graph structuralization task. For example, in the case of road accidents, the system automatically divides free-description sentences about the incident situation into individual words and categorizes them into five groups: Environment, which is used for describing attributes of an occupant and the environment, Persons, Vehicles, Things except vehicles, and Action. Secondly, it outputs nodes with labels of the dissolved words into the software GUI for creating situation graphs. This graph-editing GUI is used for describing both the relationship among components of the environment where an incident occurs and the time-series process of change of the relationship. Then, using the GUI, a user can create a graph structure by connecting the nodes. The user creates four kinds of graph structures of the situations: the environment of the situation (Environment graph), before an incident (Before graph), during an incident (During graph), and after an incident (After graph). The following are the directions for graph structuralization for accident data.

Directions for Graph Structuralization

- Rule 1. In the "Environment box," which is the GUI of the system for editing an Environment graph, the user creates the graph data that describes the attributes of the environment and the occupants. In the "Before box," the user creates the graph data that describes the situation before the incident. In the "During box," the user creates the graph data that describes the situation from the time when the incident happens to the time when the occupant is injured. In the "After box," the user creates the graph data that describes the situation after the incident.
- Rule 2. In each box, the user creates graph data such as an Environment graph, a Before Graph, a During graph, and an After graph by selecting nodes from the candidates displayed in a pop-up menu.
- Rule 3. If the situation sentence pattern is Subject + Verb, the user connects a Subject node to an Action node (a Verb node).
- Rule 4. If the situation sentence pattern is Subject + Verb + Direct Object, the user connects a Subject node to a Direct Object node through an Action node (a Verb node).
- Rule 5. If the situation sentence pattern is Subject + Verb + Indirect Object + Direct Object, the user connects a Subject node to a Direct Object node, and an Indirect Object node and a Direct Object node to an Action node (a Verb node).
- Rule 6. The user creates each graph by adding a new graph to the situation graph of the previous phase. For example, the user can create a During graph by adding an additional graph to the previously created Before graph.

2.1.2 Management of Situation Graph Database

The database function manages the graph-structuralized situation data. The user can retrieve the registered data and modify the data. This function is mainly used for the situation mining functions, described later.

2.1.3 Management of Terminology Dictionary

The text mining engine requires a dictionary of words. In the developed system, we have to create a dictionary of domain-specific terminology and classify the words into five categories, such as Environment, Persons, Vehicles, Things, and Action. The data mining function is used for supporting this task. If the system finds unknown words while the user

proceeds with graph structuralization, the system registers them into the dictionary by interactively asking the user for a suitable category for the unknown words. As the graph-structuralization task proceeds, the dictionary becomes rich and the user is able to skip the process of registering the unknown words by hand.

We also implemented a dictionary of synonyms. In this function, one representative word is selected for each group of synonyms so that we can convert every word into its representative word, if a representative word exists. This is one of the common functions of general text mining software. We expand this function for use in the situation graph database. When the system finds a new word, the system lists the candidates of synonyms for the new word by checking the words connected to the new word and the semantic connection in the graph database.

For example, if the graph database has a situation graph corresponding to the situation "a bicycle collided with a car," then in the dictionary the word "collide" is registered as a word connected to nodes with the labels "bicycle" and "car." When creating a situation graph corresponding to the new situation "a bicycle crashed with a car," which includes a new word "crash," the system recognizes that "crash" has the same semantic structure as "collide" by using the synonym function. The user can register new words by representatives from the candidates.

2.2 Situation Mining

2.2.1 Situation Semantic Search

The situation semantic search function enables us to search data considering the time-series of the situational structures. It is difficult to do this kind of search by using a simple keyword search algorithm. Users can search situation data by creating a situation graph as a search query. If users set any nodes as "necessary condition," the user will get results that not only satisfy Eq. (2), but also the nodes existing in the graph structures of the results.

In the algorithm of the situation semantic search, we can compute the similarity between two different situations by the rule that "two graphs are similar if they share many edges in common." We represent each phase of a graph-structuralized situation as G (Before graph: G_B(v,e), During graph: G_D(v,e), After graph: G_A(v,e)), and the number of edges that two different situations G_i G_j share in common as Sum($G_i \cap G_j$). The similarity among two situation graphs can be defined as follows:

$$Sim(G_i,G_j)=\frac{Sum(G_{iB}\cap G_{jB})+Sum(G_{iD}\cap G_{jD})+Sum(G_{iA}\cap G_{jA})}{Sum(G_i)}. \quad (1)$$

We can search the desired situation data by calculating the similarity expressed by Eq. (1). Specifically, by defining the following condition expressed by Eq. (2), we can find a similar situation.

$$Sim(G_i, G_j) > S_{min}, \quad (2)$$

where Sim_{min} is the minimum value of similarity that ranges from 0 to 1.

2.2.2 Linkage Analysis on Situation Graph

The linkage analysis function allows us to compute which nodes are connected to a specific node and the frequency of of connections. Using the developed software stated later, the user can set a search condition by simply connecting the specific node to a "WHAT" node. Then, by pushing a "Search WHAT" button, the system finds the nodes connected to the specific node and counts how many times the node is linked to other nodes.

2.2.3 Situation Clustering

The clustering function conducts cluster analysis by various kinds of clustering methods such as K-means, hierarchical clustering, and graph kernel. The cluster analysis is conducted separately for each phase of situations, such as Before graphs, During graphs, and After graphs. The detail of the algorithm is as follows. First, the system computes unique graph structures in each graph G' (G'$_B$(v,e), G'$_D$(v,e), G'$_A$(v,e)) by the following formulas:

$$G'_B = G_B, \quad (3)$$
$$G'_D = G_D - G_B, \quad (4)$$
$$G'_A = G_A - G_D. \quad (5)$$

For example, Fig. 2 and Fig. 3 show a situation graph for "a child rode a bicycle and fell down with the bicycle." In this case, the graph in Fig. 2 is G'$_B$(=G_B), and the graph in Fig. 3, in which the nodes and edges colored gray are expressed to be the common graph with G_B, is G'$_D$. The similarity matrix is expressed by Eq. (6), in which the suffix "D" indicates the matrix for the During graph.

$$\mathbf{Msim_D} = \begin{pmatrix} 1 & Sim_{D12} & ... & Sim_{D1n} \\ Sim_{D21} & 1 & ... & ... \\ ... & ... & 1 & ... \\ Sim_{Dn1} & ... & ... & 1 \end{pmatrix}, \quad (6)$$

where Sim_{Dij} indicates the following value:

$$Sim_{Dij} = \frac{Sum(G'_{Di} \cap G'_{Dj})}{Sum(G'_{Di})}. \quad (7)$$

The distance matrix $Mdist_D$ for this clustering is the same as the average of $Msim_D$ and the transposed $Msim_D$.

$$\mathbf{Mdist_D} = \frac{\mathbf{Msim_D} + \mathbf{Msim_D^T}}{2}. \quad (8)$$

2.2.4 Visualization of a Typical Situation Process

The visualization function visualizes the process of typical situations by using the results of clustering. First, the system conducts a cluster analysis separately for each phase of the graphs (Before graph, During graph, and After graph). In this procedure, typical situations for each phase are clarified and displayed as nodes in the developed software. Second, the system analyzes the typical connections among different

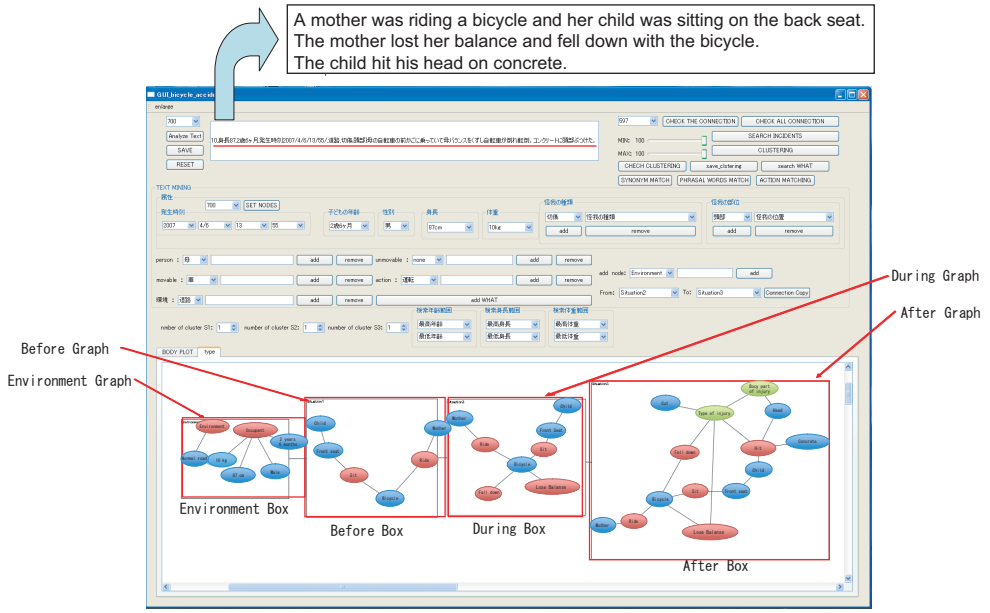

Fig. 4. Example of graph structuralization

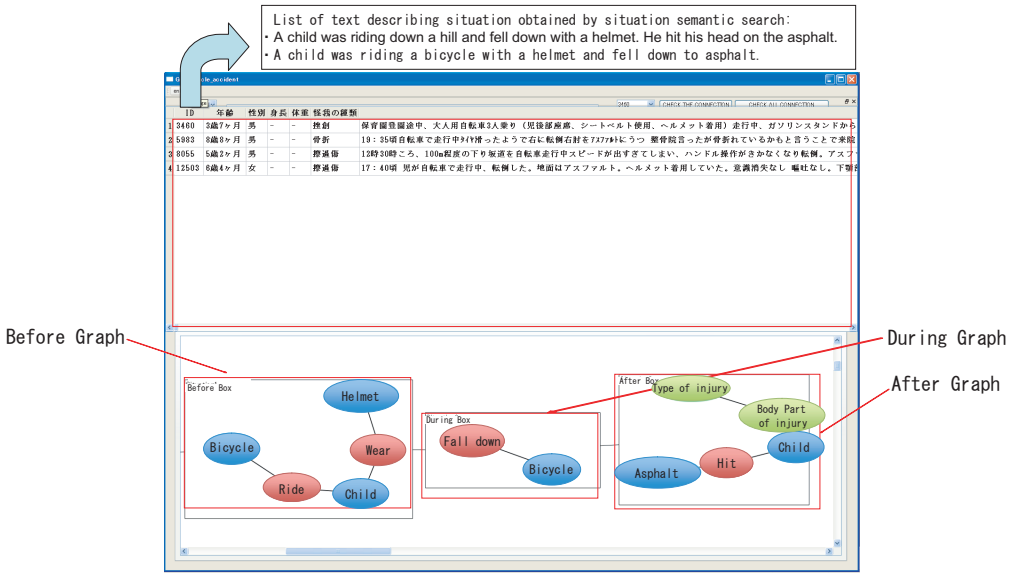

Fig. 5. Example of the result of the situation semantic search

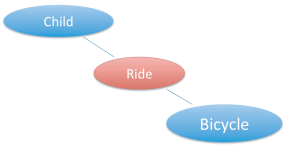

Fig. 2. Example of a Before graph (G_B)

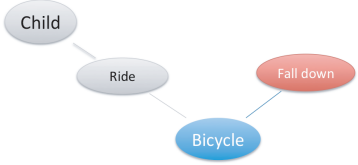

Fig. 3. Example of a unique During graph (G'_D)

phases, namely, the connection between a Before situation and a During situation, the connection among a Before situation,

a During situation, and an After situation, and so forth. In the developed software, the connections are displayed as edges. In addition to this visualization, we use a function that shows the ratio of the number of incidents belonging to the typical situation to the number of all incidents. The visualization allows us to understand which are the components of each typical situation and which components are more important than others.

3 Evaluation of the Situational Analysis System

We performed experiments to evaluate the effectiveness of the developed system by using the real data of 818 child-bicycle incidents, collected at the National Center of Child Health and Development [13].

3.1 Dataset for Evaluation

The 818 child-bicycle incident data include the attributes of occupants, injuries, date of incidents, and site where incidents happened, as well as free-description sentences about the incident situations. The contents in the four situation graphs (Environment, Before, During, and After) are as follows. An Environment graph is created by using the "attributes of an occupant and the environment." A Before graph is created by finding information on the "situation before the incidents" from the free-description sentence. A During graph is created by finding information on the "situation during the incidents" from the free-description sentence. An After graph is created by combining information on the "situation after the incidents" extracted from the free-description sentence, "type of injuries and body parts of injuries" and "which actions caused the incidents."

3.2 Creation of Situation Database

As a specific example of graph structuralization, we show a graph-structuralized situation with one item of data. Figure 4 shows the GUI of the developed software system for the following accident case. Injured child's age: 2 years and 6 months old, Sex: male, Height: 87 cm, Weight: 10 kg, Type of Injury: cut, Body Part of Injury: head, Date and Time of Incident: 13:55 on 6th July 2007, Site of Incident: normal road, Free-description sentence: "A mother was riding a bicycle, and her child was sitting on the back seat. The mother lost her balance and fell down with the bicycle. The child hit his head on concrete." The area at the bottom of Fig. 4 shows the four created graphs (Environment, Before, During, and After) in this case.

3.3 Evaluation of Situation Semantic Search

We conducted an efficacy analysis on the situation semantic search. For the setup, we set Sim_{min}=1. For the search, we graph-structuralized the situation "A child rode a bicycle with a helmet and fell down. After that, he hit his body on concrete" as a condition setting. Figure 5 shows the results at the top and the graph structure for the condition setting at the bottom. We calculated Precision (P), Recall (R), and F-measure (F) in Eqs. (9), (10), and (11).

$$R = \frac{w}{(w+x)}, \tag{9}$$

$$P = \frac{w}{(w+y)}, \tag{10}$$

$$F = \frac{2PR}{(P+R)}, \tag{11}$$

where w is the number of correct documents, x is the number of unexpected documents, and y is the number of missing documents.

The evaluation result is shown in Fig. 6. This figure indicates that by setting Sim_{min}=1 we can search the situation data that matches completely the given situation graph and we can also obtain similar situation data by changing the Sim_{min} value from 0 to1.

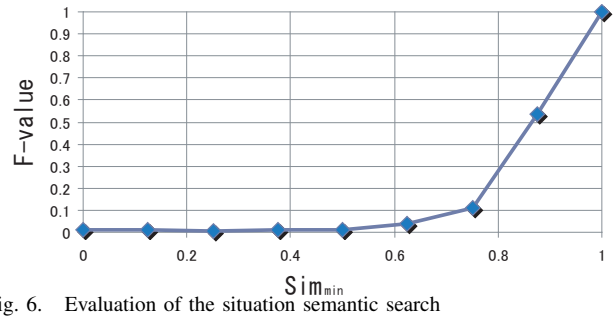

Fig. 6. Evaluation of the situation semantic search

3.4 Effectiveness of Situation Clustering

The developed system supports various kinds of clustering. We show a few examples of clustering by the supported functions. Figure 7 shows an example of the polar dendrogram obtained by hierarchical clustering for the Before graphs. Figures 8 and 9 show the results of clustering using the group average method for the Before graphs and the During graphs, respectively. In the two figures, the number of clusters is 11 for the situational analysis for the Before graphs and 15 for the situational analysis for the During graphs. The number of clusters was determined by the authors by considering the free-description sentence in each cluster. Thus, the system can conduct graph-structuralization based clustering.

We evaluated the effectiveness of situation clustering using the F values. First, we define a typical situation graph for each cluster ID by using the clustering results. Second, we conduct a situation semantic search by giving this typical situation graph to the system as a search query. Then, we can obtain the complete set of situation graphs corresponding to the search query. By comparing this complete set and the results of clustering, we calculate the P, R, and F values. Table I shows the evaluation of clustering for the Before graphs. The F values of the table suggest that the average performance is high in the case of the clustering results shown in Fig. 8. Since this performance depends on the number of clusters that the user gives, the user should set it adequately in actual use; namely, the user should change the grain size of each cluster

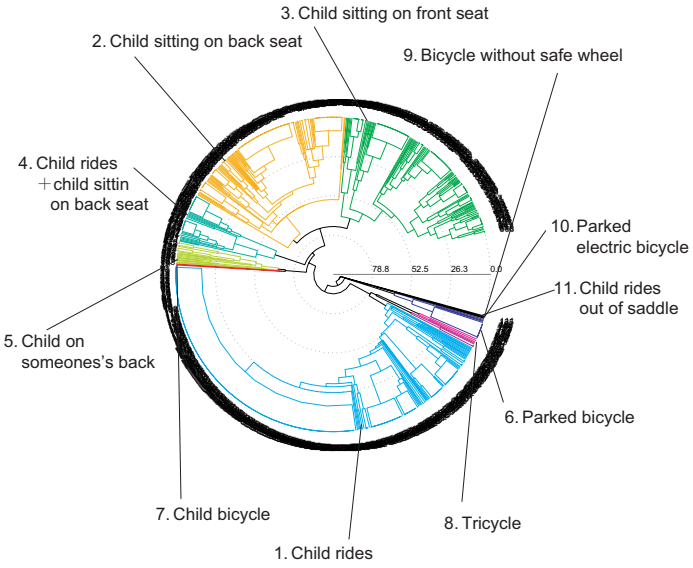

Fig. 7. Polar dendrogram of situation clustering for the Before graphs (number of clusters = 11)

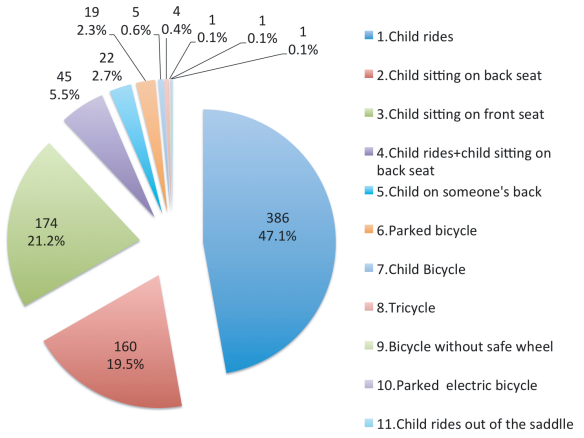

Fig. 8. Result of situation clustering for the Before graphs (number of clusters = 11)

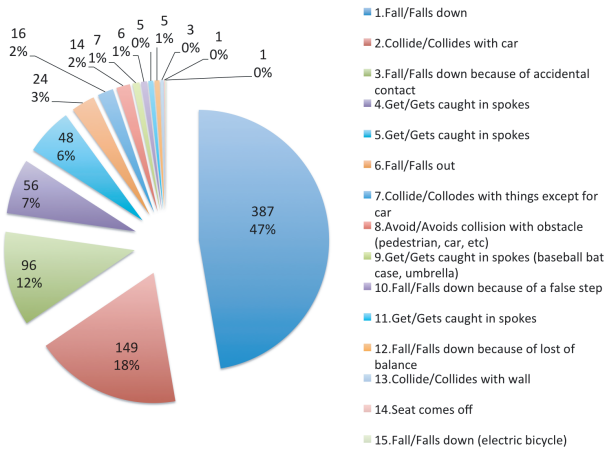

Fig. 9. Result of situation clustering in the case of the During graphs (number of clusters = 15)

so that the user can have an insight into each clustered situation from the viewpoint of injury prevention and situation control.

TABLE I
EVALUATION OF CLUSTERING FOR THE BEFORE GRAPHS

Cluster ID	1	2	3	4	5	6
P (Precision)	0.995	0.994	0.856	0.889	0.273	1.000
R (Recall)	0.948	0.975	0.797	0.068	0.857	0.415
F (F-measure)	0.971	0.985	0.825	0.126	0.414	0.253
Cluster ID	7	8	9	10	11	Average
P (Precision)	1.000	1.000	1.000	1.000	1.000	0.910
R (Recall)	0.833	1.000	1.000	1.000	1.000	0.874
F (F-measure)	0.909	1.000	1.000	1.000	1.000	0.771

3.5 Effectiveness of Visualization of a Typical Situation Process

As a specific example of a typical situation process analysis, we show in Figure 10 the visualization of a typical situation process with the results of clustering we obtained above. The three boxes indicate typical Before situations, During situations, and After situations. The red nodes in each box indicate the most frequent situation. In this figure, for example, we can find the following typical processes. ”Get/Gets caught in spokes (baseball bat case, umbrella)” in a During situation, which is colored yellow in Fig. 10, is connected to two Before situations, ”Child rides” and ”Child sitting on back seat.” The During situation is also connected to the three After situations of ”Bruises,” ”Hit/Hits body” and ”Non-categorized.” Thus, this visualization allows us to understand the components of each typical situation and which components are more important than others.

Using the function of the typical situational analysis, for example, we identified typical situations such as ”a child rides a bicycle and collides with a car,” ”a child drives a bicycle and falls down,” ”someone rides a bicycle with a child on the back seat. The leg of the child gets caught in the spokes,” and ”someone rides a bicycle with a child on the front seat, the bicycle falls down because of accidental contact.”

4 Conclusion

In this research, as a new situational analysis system to extract situation structures from a large number of situation data, we proposed a new situational system that consists of two basic functions: a function for graph structuralization and a function for situation mining based on the situation graph data. The feature of the system lies in formulating a situational analysis as a graph structuralization analysis. We implemented functions for a situation semantic search, linkage analysis of a situation, situation clustering and visualization of a typical situation process. To evaluate the effectiveness of the developed system, we analyzed the real data of 818 child-bicycle incidents. Using this system, we created a database of situation graphs by transforming the 818 bicycle incident data into situation graphs. With the situation semantic search, users can search similar situation data, which is difficult by conventional keyword search methods or text mining methods. Situation clustering is implemented by applying a hierarchical clustering method to the situation graph data. This function

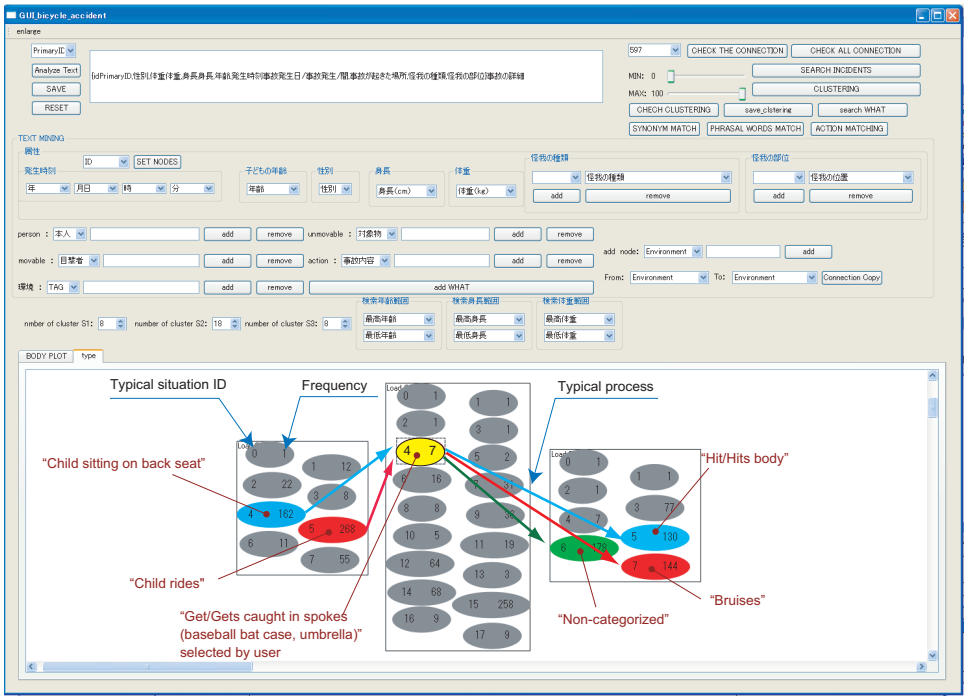

Fig. 10. Visualization of a typical situation process

allows the user to grasp the typical situation of each phase of situations, such as situations before an accident, during an accident, and after an incident. Based on the function of situation clustering, the developed system can visualize typical situation processes.

For future work, we plan to develop a graph-structuralization supporting function. This function will help users to input necessary data for prevention. For example, if a user tries to graph-structuralize an incident and forgets to input information of helmet use status, this new function will let the user know about the necessity of helmet use information. This function will enhance the quality of graph-structuralized situations. In addition, we also plan to make an open database of various incidents. By making a database of graph-structuralized situations of various incidents in accordance with various products or various places, and by sharing the information of situational analyses, we hope to make society safer and more cooperative for injury prevention.

REFERENCES

[1] ISO 3100: 2009 Risk Management-Principles and Guidelines, 2009.
[2] ISO/ICE, Guide 51 Safety Aspects-Guidelines for Their Inclusion in Standards, 1999.
[3] JST Failure Knowledge Database, http://shippai.jst.go.jp/fkd/Search.
[4] Y. Hatamura, Learning from Design Failure, Springer, 2009.
[5] World Health Organization, World Report on Child Injury Prevention, 2008.
[6] J. Shawe-Taylor and N. Christianini, "Kernel Method for Pattern Analysis," Cambridge University Press, 2004.
[7] T. Garther, Kernels for Structured Data, World Scientific Publishing Co. Pte. Ltd., 2008.
[8] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina, "Web Graph Similarity for Anomaly Detection," Journal of Internet Services and Applications, Volume 1 (1). pp. 19-30, 2010.
[9] A. Mislove, M. Marcon, K. P. Gummandi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," the 5th ACM/USENIX Internet Measurement Conference (IMC'07), San Diego, CA, October 2007.
[10] D. W. Mount, "Sequence and Genome Analysis," Cold Spring Harbor Laboratory Press, 2004.
[11] B. S. Everitt and T. Hothorn, A Handbook of Statistical Analyses Using R, Second Edition, Chapman and Hall/CRC, 2009.
[12] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.
[13] the National Center of Child Health and Development, http://www.ncchd.go.jp/English/Englishtop.htm.