IADIS International Journal on Computer Science and Information Systems
Vol. 8, No. 1, pp. 130-141
ISSN: 1646-3692

AN AWARENESS SYSTEM OF RISK CONTROL ACTION BY CONSTRAINED CLUSTERING ON PROJECT REPORTS

Masaki Samejima, *Graduate School of Information Science and Technology, Osaka University, 2-1, Yamadaoka, Suita, Osaka, 5650871, Japan*

Yuuki Imanara, *Graduate School of Information Science and Technology, Osaka University, 2-1, Yamadaoka, Suita, Osaka, 5650871, Japan*

ABSTRACT

In order to avoid project managers' failing to perform appropriate risk control actions in a planning phase of the project, we propose an awareness system to display the appropriate risk control actions to the project managers. Because the project manager needs practical risk control actions, the awareness system extracts sentences of risk control actions from past project reports that are accumulated in the information system development company. The proposed system identifies the sentences that are related to the target risk in the project reports and extracts sentences of the risk control action from the project reports with the related sentences. In order to identify the related sentences to the target risk, the proposed system makes clusters of sentences with constraints to include sentences of risk control actions, and relates the target risk to the cluster that has the similar sentence to the sentence of the target risk. The experimental result shows that the proposed system can extract the sentences of risk control actions at F-measure of 65.7%.

KEYWORDS

Awareness system, Risk Control Action, Project Report, Constrained Clustering.

1. INTRODUCTION

As the scale and the complexity of the information system increase, it becomes more difficult to succeed in the project of information system development. Several report say that 30% of the projects fail (Cerpa, N. and Verner, J. M., 2009). As indicated in Project Management Body Of Knowledge (Project Management Institute, 2008), project managers in information development companies put their experience of the project into documents; a project manager writes a project plan document in planning a new project and a project report in completing the project. In order to mitigate the risk in planning the project, the project manager plans risk control actions for the target risk by checking the list of the standard risk control actions that are defined in the company. The list of the standard risk control actions only the abstract information of the risk control actions. Therefore, the project manager needs to collect the concrete information of the risk control actions to know how to perform the risk control actions.

The project reports that the project managers wrote at the end of the project include the information of what risk control actions are done in the past projects. The project manager often finds the concrete information of the practical risk control actions in the project reports. For the purpose of the support to find project reports that are related to the target risk, many project report management systems have been proposed. The project report management systems register project documents and retrieve the documents by keyword matching (Chapman, C. and Ward, S., 2007, Patterson, F. D. and Neailey, K., 2002). However, the project managers have to read the project reports to obtain the information of the risk control actions, which is a time-consuming task for the project managers. This paper addresses making the project managers aware of information of the risk control actions without reading the project reports.

The awareness system displays the information of the risk control actions for the risks that the project manager faces by extracting the information from project reports. In order to identify the risk that the project manager faces, our research group has already developed the method to extract the related sentences to the target risk from the accumulated project plan documents with a project manager's writing one that includes the sentence of the target risk (Imanara, Y. and et al., 2012). Because the project plan document that includes the extracted sentences corresponds to the project report for the same project, it is possible to identify not only the project plan documents but also the project reports for the target risk by using our method. However, the problem still remains that the sentences of the risk control actions for the target risk are not extracted from the identified project report.

To realize the awareness system, we develop the extraction method of sentences of risk control actions from the project reports that are identified by the project plan document. Because the project reports have information of multiple risks, the proposed method identifies which sentences correspond to the target risk. We call the sentences that correspond to the target risk "risk-related sentences". And the proposed method extracts the sentence of the risk control action from the risk-related sentences.

The rest of the paper is organized as follows. Section 2 outlines the document-based management of project risk. Section 3 describes the awareness system with the extraction method of the risk control actions from project reports. Section 4 describes the evaluation experiment of the proposed method. Section 5 deals with the conclusion derived from the experimental results.

2. DOCUMENT-BASED MANAGEMENT OF PROJECT RISK

2.1 Conventional Researches on the Support of the Project Risk Management

The project reports in the past projects have been used for the risk management in the conventional researches. One of the major purposes to use the project report is to identify success factors of the project management (Jiang, J. and Klein, G. 2000, Kwak, Y.H. and Stoddard, J., 2004, Schalken, J., and et al., 2006). By analyzing the numerical values such as budget, productivity and so on in the project report statistically, the success factors are identified as the frequent factors in the past success projects. The success factors are reflected to the standard process of the project management or used for training the project managers.

Other researches focus on that the project reports are useful for making the project manager aware of the risks and the solutions in the ongoing project. The literature (Alhawari, S., et al., 2012) describes the risk awareness system using the project document. The awareness system displays the knowledge to manage the risks to the project managers. In order to realize the system, the sentences and numerical values in the past project documents have to be changed to the knowledge based on the theory on the knowledge management such as SECI model (Nonaka, I. and Takeuchi, H. 1995). For making the analysis process easier, it is proposed that the document is described with the common language or model (Tah, J. and Carr, V. 2001).

However, the analysis process of the past project document is still a hard work and needs high experience on the project management because the characteristics of the projects is too various to describe with the common model. Most of the software development companies do not have enough human resource to analyze the past project document. Therefore, the project manager searches and reads the project reports to lead the ongoing project to the success one without the knowledge derived by the analysis, which makes the project result in failure. Conventional researches have addressed supporting project documents retrieval for the project manager (Menzies, T. and Marcus, A., 2008) with using text processing technique (Salton, G., 1989). In this paper, we address the project risk management to make the project managers aware of the risk control actions without the project managers' workload by using the retrieval method of the project documents with text processing technique.

2.2 Problem on Document-Based Management of Project Risks

The target of this research is the project risk management by using the accumulated document sets of a project plan document and a project report. A project plan document has sentences of cost estimation, the schedule, risks and so on. A project report has sentences of the occurred risks, risk control actions and so on. Project managers use these document sets for planning risk control actions. The process and the problem on the project risk management using the documents are shown in Fig. 1.

As shown in Fig. 1, a project manager writes a project plan document with sentences of risks that the project manager detected by the risk detection tool such as a checklist. We call the detected risk "target risk" in the project plan document. For planning risk control actions for the target risk, the project manager refers a list of standard risk control actions that are predefined in the company as abstract actions (e.g. define goal). In order to know the detail of the actions, the project manager retrieves the project reports that are related to the target risk from the documents sets. Because we have already developed the retrieval system of the document sets by using the sentences of the target risks (Imanara, Y. and et al., 2012), the document sets of the same risk as the target risk can be shown to the project managers. Reading project reports that are retrieved by the conventional method, the project manager can find what the risk control actions are done in the past project. However, the project manager has to read long project reports, which takes much time. So, our goal of this research is to develop an awareness system by extracting and displaying sentences of the risk control actions in the project reports.

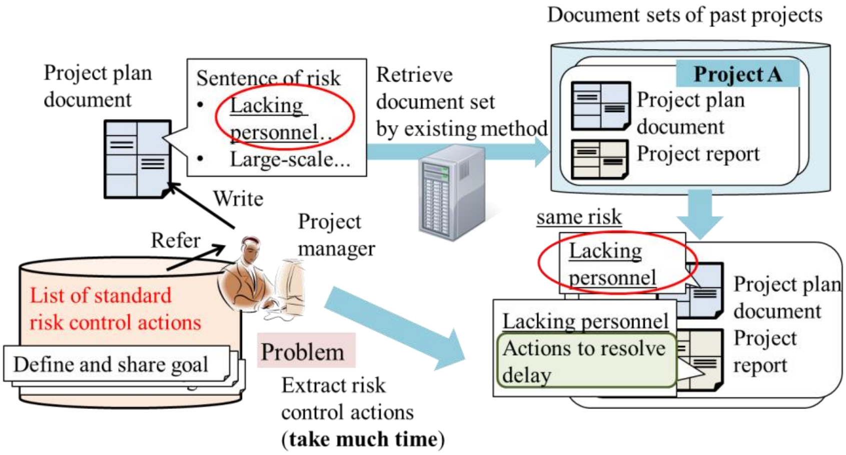

Figure 1. Problem on project risk management using documents of the project

2.3 Research Issues on the Awareness for Risk Control Actions

As described in the section 2.2, it is possible to identify the document sets related to the target risk by our conventional method. So, we address extracting the sentences of the risk action controls from the project reports in the document sets. Because the sentences of the target risk in the project plan document and the sentences of the risk control actions are related to each other, the sentence of the risk action control is extracted from project report by the sentence of the target risk in the project plan document. Fig. 2 shows the flow of the extracting the sentence of the risk control action.

A typical approach to extract the sentences of the risk action is to extract similar sentences to the sentence of the target risk in the project plan document because of the relation between both sentences. Jaccard coefficient (Hamers, L., and et al., 1989) is a well-known way of measuring the similarity between sentences. Let s_i and s_j denote the sets of words in the sentences, and $\mathrm{s}_i \cap \mathrm{s}_j$ denote the set of common words in both sentences.

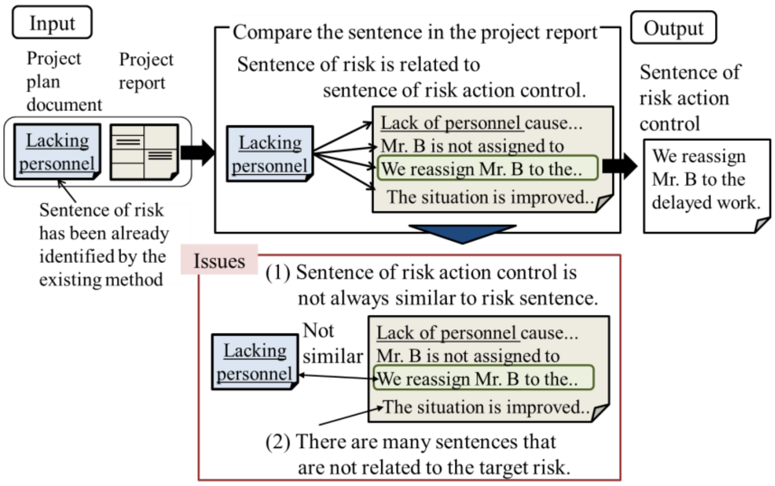
Input
Project plan document
Project report
Lacking personnel
Sentence of risk has been already identified by the existing method
Compare the sentence in the project report
Sentence of risk is related to sentence of risk action control.
Lacking personnel
Lack of personnel cause...
Mr. B is not assigned to
We reassign Mr. B to the..
The situation is improved..
Output
Sentence of risk action control
We reassign Mr. B to the delayed work.
Issues
(1) Sentence of risk action control is not always similar to risk sentence.
Lacking personnel
Not similar
Lack of personnel cause...
Mr. B is not assigned to
We reassign Mr. B to the..
The situation is improved..
(2) There are many sentences that are not related to the target risk.

Figure 2. The flow of the extracting the sentence of the risk control action

$$\text{Jaccard coefficient}(s_i, s_j) = \frac{n(s_i \cap s_j)}{n(s_i) + n(s_j) - n(s_i \cap s_j)}$$

However, by the approach of extracting the sentence that indicates the largest similarity, wrong sentences are often extracted. The reasons of the wrong extraction are shown in the following:

(1) The sentence of the risk control action is not always similar to the risk sentence.
Because a project report describes the result of the project that is different from the plan of the project, the expressions in a project report are also different from the expressions in a project plan document.

(2) There are many sentences that are not related to the target risk.
A project report describes various sentences related to the evaluation of the project, the outline of the development, and so on. Especially, the project report has sentences of the risks except for the target risk. These sentences may be wrongly extracted as the sentence of the risk control action for the target risk.

3. AN AWARENESS SYSTEM OF RISK CONTROL ACTION

3.1 Outline of the Awareness System by Extraction of Risk Control Actions

The outline of the proposed awareness system is shown in Fig. 3, and the process of the system is described in the following.

When the project manager writes the project plan document, the awareness system monitors what is described in the document automatically. Based on the content of the project plan document, the awareness system retrieves the sets of the project plan document and the

project report that are related to the target risk from the project document database by the conventional method in section 2.2. Next, the awareness system extracts the sentence of risk control actions from the document sets (Samejima, M. and Imanara, Y., 2013.).
Finally, the awareness system outputs the risk control actions to the project manager. The project manager can be aware of the risk control actions by checking the risk control actions that are output from the awareness system.

In the extraction step, it is necessary to resolve the issues that are indicated in section 2.3. The sentence of the risk action control is not always similar to the sentence of the target risk in the project plan document, but similar to the sentence of the target risk in the project report. As indicated in section 2.3, both sentences of the target risk in the project plan document and the project report are similar to each other. So, the proposed method identifies the sentence of the target risk and extracts the sentence of the risk control action based on the identified sentence.

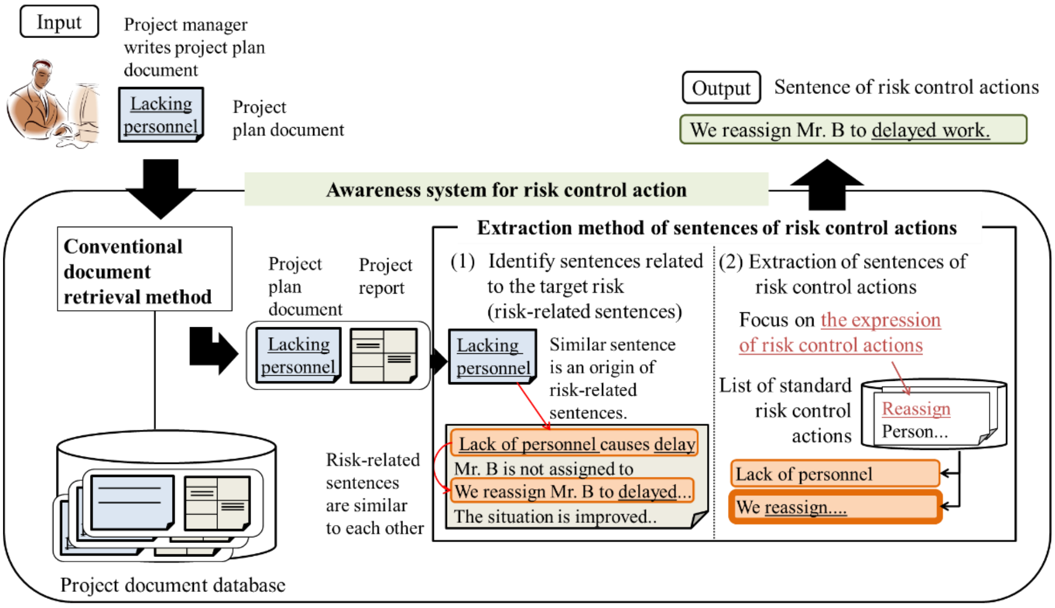
Input
Project manager writes project plan document
Lacking personnel
Project plan document
Output
Sentence of risk control actions
We reassign Mr. B to delayed work.
Awareness system for risk control action
Conventional document retrieval method
Project plan document
Project report
Lacking personnel
Project document database
Risk-related sentences are similar to each other
Extraction method of sentences of risk control actions
(1) Identify sentences related to the target risk (risk-related sentences)
Lacking personnel
Similar sentence is an origin of risk-related sentences.
Lack of personnel causes delay
Mr. B is not assigned to
We reassign Mr. B to delayed...
The situation is improved..
(2) Extraction of sentences of risk control actions
Focus on the expression of risk control actions
List of standard risk control actions
Reassign
Person...
Lack of personnel
We reassign....

Figure 3. Outline of the awareness system

Therefore, we develop the extraction method based on the characteristics of the similarity. First, focusing on that the sentences related to the target risk are similar to each other in a project report, the proposed method identifies such sentences as “risk-related sentences”. In order to identify which sentence is related to the target risk, the proposed method generates clusters of sentences that are related to each risk with constraints. Second, the proposed method extracts the sentences of the risk control actions from the risk-related sentences by using the list of the standard risk control actions. The following sections describe the above 2 functions: identification of risk-related sentences in section 3.2 and extraction of sentences of risk control actions in section 3.3.

3.2 Identification of Risk-Related Sentences by Constrained Clustering

Fig. 4 shows the outline of the identification of risk-related sentences. The identification of risk-related sentences consists of 2 steps of the identification of the origin of the risk-related sentences and the identification of the entire of the risk-related sentences from the origin sentences.

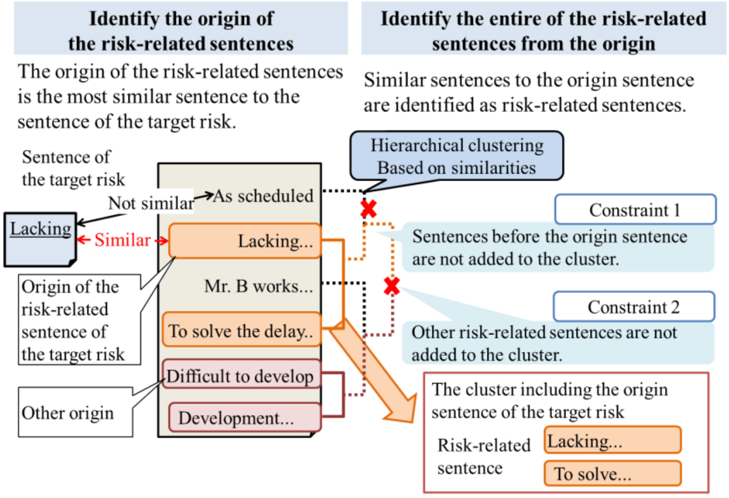
Identify the origin of the risk-related sentences
The origin of the risk-related sentences is the most similar sentence to the sentence of the target risk.
Identify the entire of the risk-related sentences from the origin
Similar sentences to the origin sentence are identified as risk-related sentences.
Sentence of the target risk
Not similar
As scheduled
Lacking
Similar
Lacking...
Mr. B works...
To solve the delay..
Difficult to develop
Development...
Origin of the risk-related sentence of the target risk
Other origin
Hierarchical clustering Based on similarities
Constraint 1
Sentences before the origin sentence are not added to the cluster.
Constraint 2
Other risk-related sentences are not added to the cluster.
The cluster including the origin sentence of the target risk
Risk-related sentence
Lacking...
To solve...

Figure 4. Outline of the identification of risk-related sentences

The sentences of the risk in both of a project plan document and a project report are similar to each other if the target risk is the same. Before identifying the risk-related sentences, the proposed method identifies the sentence that is similar to the sentence of the risk in a project plan document. The similar sentence can be regarded as the sentence of the target risk in the project report. We call the similar sentence “origin sentence” because the proposed method starts to identify the risk-related sentences based on the similar sentence. In order to identify the origin sentence, the proposed method derives the similarities between the sentence of the risk in the project plan document and the sentences in the project report by Jaccard coefficient. If Jaccard coefficient is larger than a certain threshold, the sentences are regarded as similar sentences. Here, it is necessary to decide the threshold of the similarity by comparison to Jaccard coefficients on the sentences of all the project reports. So, the proposed method derives the distribution of Jaccard coefficients on the sentences of all project reports, and set the threshold to the 95th percentile of the distribution.

Next, the proposed method identifies the risk-related sentences based on the origin sentence that has the similarity over the threshold. Focusing on that the risk-related sentences are similar to each other, the proposed method makes clusters based on the similarities between sentences in the project report. And, the proposed method identifies the risk-related sentences as the cluster that includes the origin sentence. As our research goal is to extract the sentence of the risk control action, the cluster has to be generated with including the sentence of the risk control action. Therefore, the proposed method generates the cluster with constraints to include the sentence of the risk control action for the target risk. The constraints for including the sentence of the risk control actions are shown in the following:

Constraint1 The sentences described before the origin sentence are not added to the cluster.

Because the origin sentence is similar to the sentence of the risk, the origin sentence often indicates the same kind of the risk in the project report. The sentence of the risk control action tends to be described after the sentence of the risk.

Constraint2 Other risk-related sentences are not added to the cluster.

Because the sentences of the project report describe various risks, the risk-related sentences have to be identified for each risk.

In order to apply the above constraints to the cluster, the proposed method uses the hierarchical clustering method. The hierarchical clustering uses the similarity distance between sentences:

$$\text{Similarity distance}(s_i, s_j) = 1 - \text{Jaccard coefficient}(s_i, s_j)$$

Because the height of the hierarchy that means the similarity distance is small, the sentence is added to the cluster in the ascending order of the height. The algorithm of the constrained clustering in the proposed method is described in the following:

(1) The proposed method selects the sentence that has the smallest similarity distance to the origin sentence.

(2) If the origin sentence and the selected sentence satisfy the constraints, this process continues to (3). Otherwise, the origin sentence is identified as one risk-related sentence.

(3) The selected sentence and the origin sentence are combined to a cluster.

(4) The proposed method selects the sentence that has the smallest sum of distances to the sentences belonging to the cluster.

(5) If the sentences in the cluster and the selected sentence satisfy the constraints, this process continues to (3) repeatedly. Otherwise, the sentences in the cluster are identified as the risk-related sentences.

3.3 Extraction of Sentences of Risk Control Action

As shown in Fig. 3, the proposed method extracts the sentence of the risk control actions from the risk-related sentences by the expression of the risk control action in the list of the standard risk control actions. We call the expression of the risk control action “action expression”. Especially, verbs in the sentences represent the risk control actions, e.g. “add”, “deal”, “agree” and so on. On the other hand, typical verbs in information system development are in the list of the standard risk control actions, e.g. “design”, “delay”, and so on. By just keyword matching with all the verbs in the list, sentences that do not include the risk control actions are also extracted wrongly.

Such typical words are not included in the risk-related sentences but are included in the other sentences. So, the proposed method weights the verbs that are frequently appeared in the risk-related sentences based on tf-idf (term frequency - inverse document frequency) that is the general method of weighting words for information retrieval (Manning, C.D., et al., 2008).

Fig. 5 shows the extraction of sentences of risk control actions by weighting words. The extraction process consists of weighting action expressions and extracting risk control actions.

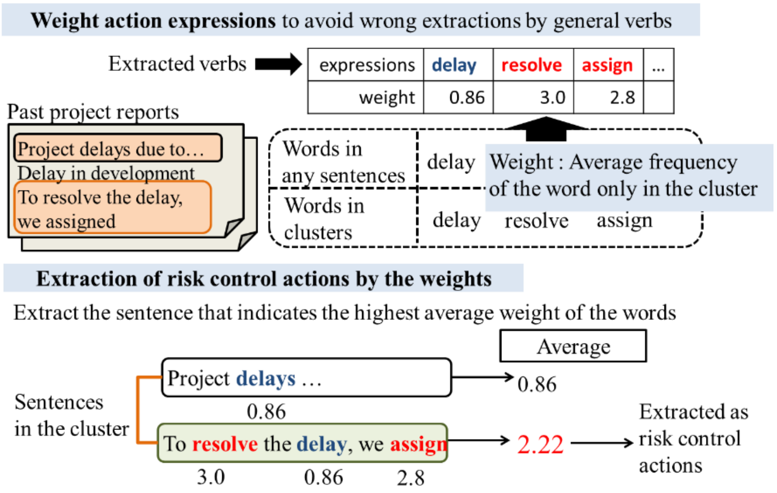
Weight action expressions to avoid wrong extractions by general verbs
Extracted verbs
expressions
delay
resolve
assign
...
weight
0.86
3.0
2.8
Past project reports
Project delays due to…
Delay in development
To resolve the delay, we assigned
Words in any sentences
delay
Weight : Average frequency of the word only in the cluster
Words in clusters
delay
resolve
assign
Extraction of risk control actions by the weights
Extract the sentence that indicates the highest average weight of the words
Average
Sentences in the cluster
Project delays …
0.86
0.86
To resolve the delay, we assign
3.0
0.86
2.8
2.22
Extracted as risk control actions

Figure 5. Extraction of sentences of risk control actions

(1) Weight action expression
The verbs that are described only in the cluster tend to express the risk control action because the clusters are generated so as to include the risk control actions. So, the kth $verb_k$ in the lth $cluster_l$ of the risk-related sentences is weighted by the following:

$$weight_{k,l} = \text{Frequency of } verb_k \text{ in } cluster_l \times \frac{\text{The number of clusters}}{\text{The number of clusters including } verb_k}$$

In Fig. 5, the verb of “delay” appears in any sentences but the verbs of “resolve” and “assign” appear only in the cluster. So, the weights on “resolve” and “assign” are higher than one on “delay”.

(2) Extract risk control actions
The proposed method decides the score of the inclusion of the risk control action for each sentence s_i in $cluster_l$. Let $w_{i,k}$ denote whether $verb_k$ appears in sentence s_i or not. The $score_{i,l}$ is based on how many weighted words are included in the sentence s_i that belongs to the $cluster_l$.

$$score_{i,l} = \frac{\sum_k w_{i,k} weight_{k,l}}{\sum_k w_{i,k}}$$

The proposed methods extract the sentence that has the largest score from the cluster of the risk-related sentences.

4. EVALUATION EXPERIMENT

4.1 Target of Experiment

In order to evaluate the proposed method, we gathered the problem description in the qualifying examination for project managers in Japan. Together with a senior project manager, we extracted 27 sets of a project plan document and a project report from the problem description. And, we found 33 sentences of the risk control actions in the project report. By using the list of standard risk control actions published in Japan, we applied the following methods for comparison:

- The method by Jaccard coefficient
  The most similar sentence to the sentence of the target risk in the project plan document is extracted from the project report.
- The method by just using constrained clustering
  All the sentences in the cluster that the proposed method identifies are extracted.
- The proposed method

The evaluation criteria are the recall rate, the precision rate and F-measure that are generally used in the researches on information retrieval (Manning, C.D., et al., 2008):

$$Recall\ rate = \frac{\text{The number of the correctly extracted sentences}}{\text{The number of the correct sentences}}$$

$$Precision\ rate = \frac{\text{The number of the correctly extracted sentences}}{\text{The number of all the extracted sentences}}$$

$$F - measure = \frac{2 \times Recall\ rate\ \times Precision\ rate}{precision\ rate\ +\ recall\ rate}$$

4.2 Experimental Result

Fig. 6 shows the experimental results in applying the method by Jaccard coefficient (called “Jaccard”), the method by just using constrained clustering(called “cluster”), and the proposed method (called “proposed”).

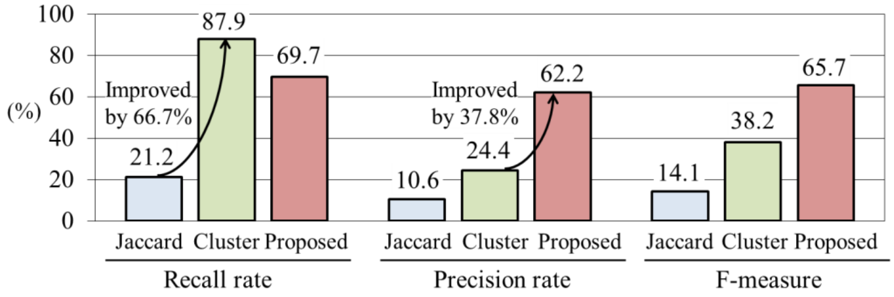

Figure 6 Experimental result

The method by Jaccard coefficient extracts only 21.2% of the correct sentences of the risk control action because the sentence of the risk control action is not always similar to the sentence of the risk as previously discussed. By using the constrained clustering that is a part of the proposed method, the recall rate is improved by 66.7%. This means that the clustering method can include the sentences of the risk control action. The precision rate is also improved by 13.8%, but it is not enough for the practical use. The proposed method can improve the precision rate by 37.8% because the sentences that do not include the risk control action are removed. Although the proposed method decreases the recall rate by 18.2%, F-measure is improved by 27.5% compared to the method by constrained clustering. Therefore the proposed method is more effective for the extraction of the risk control actions than the other methods.

The sentences that are not extracted by the proposed method tend to include more words than the other sentences. The proposed method calculates the score to extract the risk control actions by using all the verbs in the sentence. But some of the sentence includes the information of not only the risk control actions but also the target risk, the situation of the risk, and so on. While the long sentence includes the risk control actions, the score of the long sentence tend to be small. In order to improve the accuracy of the extraction, it is necessary to consider the length of the sentence, e.g. splitting the long sentence into small ones.

5. CONCLUSION

In this paper, we proposed the awareness system of the risk control actions by extracting from the project reports based on a project manager’s writing project plan document. Because the sentence of the risk control action is not always similar to the sentence of the target risk, the proposed method identifies the origin sentence as the most similar sentence and make the cluster of the risk-related sentences that include the origin sentence. In order to extract the sentence of the risk control action from the cluster of the risk-related sentences, the proposed method weights the words that often appear in the risk-related sentences. As a result of the experiment, it has been confirmed that the proposed method improves the F-measure by 51.6% compared to the extraction method of the most similar sentence by Jaccard coefficient. However, 65.7% of the F-measure is not enough in the practical use. In order to improve the F-measure, we have to refine the extraction method of the sentences of the risk control action described in section 3.3. By just using the list of the standard risk control actions, it is impossible to identify the sentences of the risk control actions from the risk-related sentences. So, it is necessary to extract the other verbs to express the risk control actions from not only the list of the standard risk control actions but also the accumulated project reports.

REFERENCES

Alhawari, S., et al., 2012. Knowledge-Based Risk Management framework for Information Technology project, *In International Journal of Information Management*, Vol. 32, pp. 50-65.

Cerpa, N. and Verner, J. M., 2009. Why did your project fail?. *In Communications of the ACM*, Vol. 52, No. 12, pp.130-134.

Chapman, C. and Ward, S., 2007. *Project Risk Management: Processes, Techniques and Insights*. John Wiley & Sons, New York, USA.

El Emam, K. and Koru, A.G., 2008. A replicated survey of IT software project failures, *In IEEE Software*, vol. 25, no. 5, pp. 84-90.

Hamers, L., and et al., 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *In Information Processing and Management* , Vol. 25, No. 3, pp.315-318.

Imanara, Y. and et al., 2012. An Identification Method of Risks in Project Plan Document by Automatic Acquisition of Risk Expression. *In Proceedings of 2012 IEEE International Conference on Systems, Man & Cybernetics.* Seoul, Korea, pp.1240-1244.

Jiang, J. and Klein, G., 2000. Software development risks to project effectiveness, *In Journal of Systems and Software*, Vol. 52, No. 1, pp. 3–310.

Kwak, Y.H. and Stoddard, J., 2004. Project risk management: lessons learned from software development environment, *In Technovation*, Vol. 24, pp. 915-920.

Manning, C.D., et al., 2008. *Introduction to Information Retrieval*, Cambridge University Press, UK.

Menzies, T. and Marcus, A., 2008. Automated severity assessment of software defect reports, *In Proceedings of IEEE International Conference on Software Maintenance*, ICSM 2008, pp. 346-355.

Nonaka, I. and Takeuchi, H., 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, USA.

Patterson, F. D. and Neailey, K., 2002. A Risk Register Database System to aid the management of project risk, *In International Journal of Project Management*, Vol. 20, No. 5, pp. 365–374.

Project Management Institute, 2008. *A Guide to the Project Management Body of Knowledge* (*PMBOK Guide*). Project Management Institute, Pennsylvania, USA.

Salton, G., 1989. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley Longman Publishing, USA.

Samejima, M. and Imanara, Y., 2013. An extraction method of risk control action from project reports by constrained clustering, *In Proceedings of IADIS Information Systems Post-implementation and Change Management Conference 2013*, pp.11-18.

Schalken, J., and et al., 2006. A Method to Draw Lessons from Project Postmortem Databases, *In Software Process Improvement and Practice*, Vol. 11, pp. 35-46.

Tah, J. and Carr, V., 2001. Knowledge-based approach to construction project risk management, *In Journal of Computing in Civil Engineering*, Vol. 15, pp. 170–177.