An Extraction Technique for Presentation Schema embedded in Presentation Documents

Shinobu HASEGAWA[a*] **& Akihiro KASHIHARA**[b]
[a] *Center for Graduate Education Initiative, JAIST, Japan*
[b] *Graduate School of Informatics and Engineering,*
The University of Electro-Communications, Japan
*hasegawa@jaist.ac.jp

Abstract: The main topic addressed in this paper is to help a novice graduate/undergraduate student compose his/her presentation document by means of presentation schema that represents heuristics for presenting research contents to be shared by laboratory members. The key idea is to propose a model of presentation structure, which represents roles of and sequences among presentation slides included in the documents with metadata. Following this model, the presentation schema is defined as a typical presentation structure for the laboratory members. This paper accordingly introduces a technique based on association rule mining for automatically extracting the presentation schema from the repository of the documents accumulated in the laboratory. In addition, we report case studies for investigating how to configure the thresholds of the mining and how the schema extracted is valid in comparing the ones between different laboratories.

Keywords: Presentation Schema, Presentation Semantics, Association Rule Mining

Introduction

In our daily research activities, composing presentation documents is one of the most important skills so that researchers and students can report the research findings with well-organized representation. However, it is quite difficult for novice graduate/ undergraduate students in the laboratory to compose the well-organized presentation documents since they have few experiences and heuristics of constructing the presentation structure to be shared by the laboratory members [1]. We call such presentation structure a presentation schema. The presentation schema would often vary according to diverse factors about presentation context, such as presentation time limitation, audiences, research domain, and presentation philosophy in the laboratory. Therefore, it is not so easy for the researchers to prepare such schema for the novices as tangible standards in advance. The final goal of our research is accordingly to help the novices develop the skill for composing the presentation documents by means of the presentation schema that could be extracted from the presentation documents accumulated by the laboratory members.

The main issue addressed in this paper is how to extract the presentation schema automatically from a certain number of the presentation documents the laboratory members have composed. We first introduce a model of the presentation structure with metadata corresponding to each presentation slide [2]. Following this model, we second utilize a machine learning technique to analyze and extract the presentation schema, especially role of and sequence among the presentation slides, from the repository of the documents attached with the metadata in advance. The extracted schema could become a scaffold for the novices to learn the presentation composition skill practically because they can become aware of the typical presentation structure of the laboratory to compose their documents. Scaffolding with presentation schema is accordingly viewed as a part of laboratory education [3]. In addition, each laboratory has its own presentation philosophy. It means that the laboratory would also have its own presentation schema. There is accordingly a great need for extracting the schema automatically with machine learning technique from presentation documents accumulated by the laboratory members.

1. Framework

1.1 Presentation Document Composition Skill

A skill in composing presentation documents is an important research one for brushing up the research itself in laboratory meetings and for reporting the research findings in international/domestic conferences. They generally include a number of slides, which present the research contents. In order to compose the presentation document well, it is necessary (1) to divide the research contents into the slides and (2) to sequence the slides in an understandable way as shown in Figure 1.

On the other hand, it is difficult for the novices to divide their research contents into a number of the presentation slides and to sequence the slides since such presentation structure is often embedded in each document and they have little knowledge about the structure specifying what to present and what order to present logically. Therefore, reading good presentation document is not enough to learn how to compose the presentation documents. In addition, expert researchers are not always good teachers for composing the presentation documents. Of course, they could point out and fix inappropriateness of the presentation documents composed by the novices. But, it is not easy to teach the presentation composition skill directly to the novices. In traditional laboratory education, such skill could be heuristically acquired through daily research activities as cognitive apprenticeship [4].

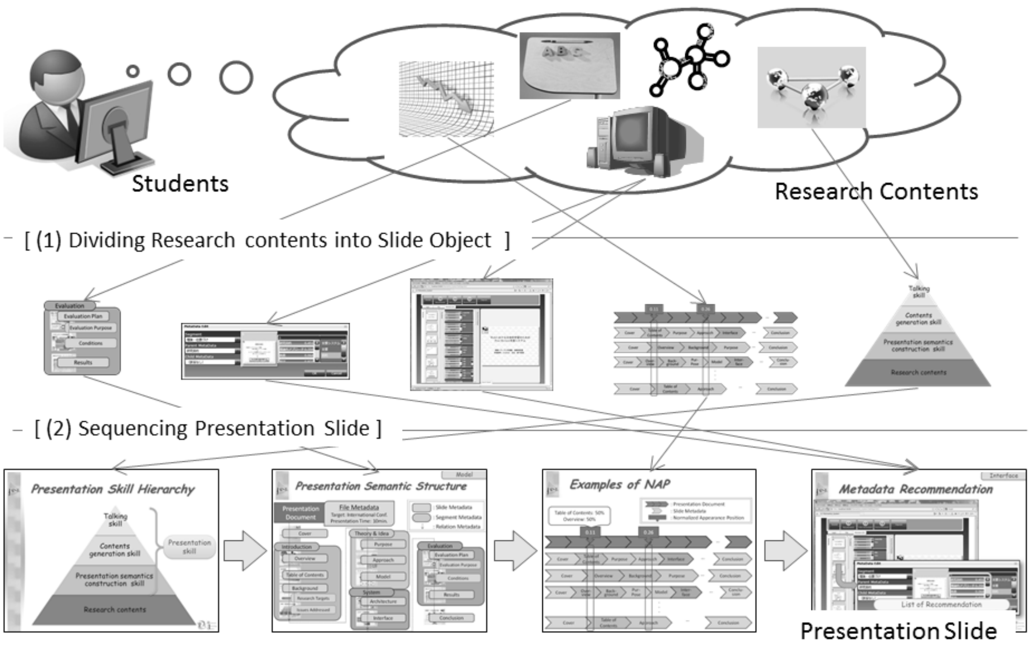

Students
Research Contents
[(1) Dividing Research contents into Slide Object]
[(2) Sequencing Presentation Slide]
Presentation Skill Hierarchy
Talking skill
Contents generation skill
Presentation semantics construction skill
Research contents
Presentation skill
Presentation Semantic Structure
Examples of NAP
Metadata Recommendation
Presentation Slide

Figure 1. Overview of Presentation Document Composition Skill

1.2 *Presentation Schema and Presentation Structure Model*

In this paper, the presentation schema is represented as a typical presentation structure, which implies some heuristics for composing the documents in the laboratory manner. The presentation schema could provide the novices with how to divide the research contents into a number of the presentation slides and how to construct the presentation structure that expresses roles of and sequences among the slides. However, it would be difficult to specify such presentation schema since it is often embedded in the presentation documents accumulated in the laboratory. Our challenge is to extract the presentation schema from the repository of the presentation documents automatically.

In order to deal with the presentation structure and schema explicitly, we provide a presentation structure model which uses three types of metadata for presentation slides as shown in Figure 2 [2]. Slide metadata represent the role that each slide included in the presentation document plays in presenting the research contents. Such metadata does not necessarily correspond to the slide title. Some of them vary according to the presentation context. Others are nested since these slide metadata often appear as compound metadata in one slide. Segment metadata also represents the section of the presentation document that several slide metadata compose for presenting the research contents. We have defined four kinds of main segment metadata. Each segment metadata is associated with several slide metadata in advance. File metadata represent some attributes of the presentation context, which includes the presenter information and presentation contextual information.

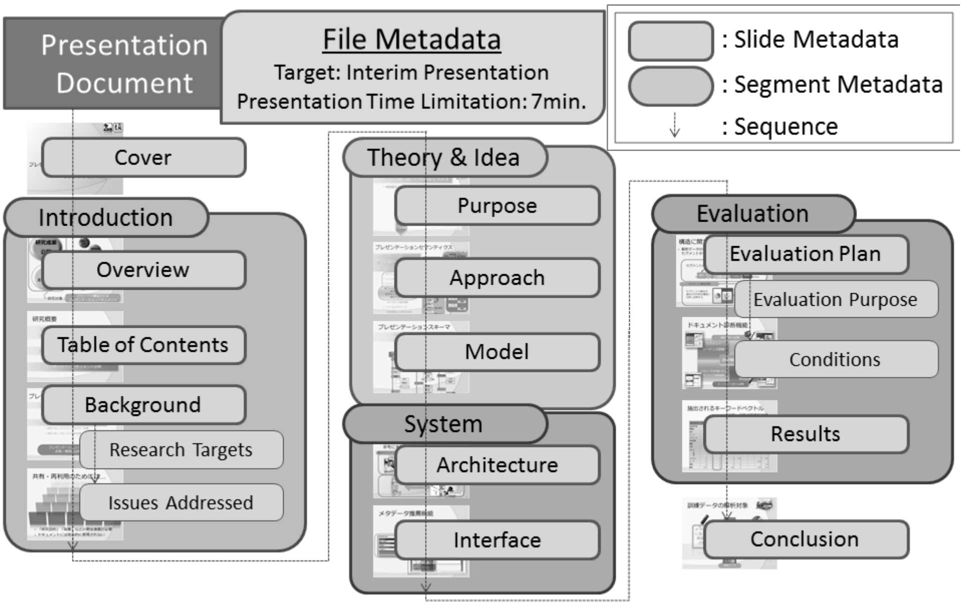


Figure 2. Presentation Structure Model

1.3 *Related Work*

Kohlhase [5] developed CPoint as a semantic PowerPoint extension that allows the authors to enrich PowerPoint documents by means of domain knowledge annotation and concept mapping. Hayama et al. [6] proposed an automatic approach for generating presentation slides from a technical paper. Li and Chang [7] developed the management model and tools that enable users to better exploit and transfer presentational knowledge assets for representing the domain knowledge.

In spite of the significance of the presentation schema and structure, each of these researches did not deal well with such information embedded in the presentation documents. In this paper, we accordingly address the issue of how to extract the presentation schema from the presentation documents accumulated from the laboratory members as training data, which are attached in advance with the metadata.

2. Extraction Technique for Presentation Schema

2.1 Basic Concepts of Association Rule Mining

Association rule mining is one of well-used techniques of data mining in various areas, which was first proposed by Agrawal et. al. [8]. It aims to extract frequent patters and casual structures among sets of items in the transaction databases [9]. In this section, we describe general definitions of the association rule mining to prepare our schema extraction algorithm.

Let $I=\{i_1,i_2,\cdots,i_n\}$ be a set of n distinct items, $T=\{t_1,t_2,\cdots,t_m\}$ be a set of m different transaction records, where each $t_m \subseteq I$. An association rule is indicated by the form of $X \Rightarrow Y$, where $X,Y \subset I$ are sets of items called item sets, and $X \cap Y = \phi$. X is called antecedent while Y is called consequent, the rule means X implies Y.

There are two important basic measures for association rules, a support described in $sup(X \Rightarrow Y)$ and a confidence described in $conf(X \Rightarrow Y)$. $sup(X \Rightarrow Y)$ is defined as the proportion of the number of transactions that contain $X \cup Y$ to the total number of transactions in *T*. $conf(X \Rightarrow Y)$ is also defined as the proportion of the number of transactions that contain $X \cup Y$ to the total number of transactions that contain *X*.

In regard to such $X \Rightarrow Y$, thresholds of support and confidence are usually predefined by users to extract those rules that are important. These thresholds are called a minimal support described in *min_sup* and a minimal confidence described in *min_conf* respectively. The association rules are finally extracted as the item sets that satisfy both *min_sup* and *min_conf*. However, there are several well-known problems in setting the thresholds [6]. The lower the thresholds are, the larger the numbers of rules are extracted, which are difficult to recognize. The higher the thresholds are, the smaller the numbers of just known rules are extracted.

2.2 Schema Extraction Algorithm

Our schema extraction algorithm aims to extract the typical semantic stricture as the presentation schema based on appearance order of the slide metadata of the presentation documents accumulated from the laboratory members, which are attached in advance with the metadata. Therefore, we adopt the association rule mining as shown in Figure 3, where *I* is a set of all kinds of slide metadata, *T* is a set of all presentation documents to be analyzed, *X* is an arbitrary slide metadata appeared in a certain presentation document as an antecedent, and *Y* is a slide metadata appeared next to *X* in the presentation document as a consequent. Suppose *conf(“Overview” ⇒ ”Background”) = 50%* and *sup(“Overview ⇒ ”Background”) = 33%*, it means that *50%* of *“Background”* slides are next to *“Overview”* slides and *33%* documents include such order relation.

In usual association rule mining, antecedent *X* and consequent *Y* are able to include multiple items but are not able to specify the order relation among these items. Therefore, our algorithm restricts the number of items to one per each *X* and *Y*, which means that *X* and *Y* only contain one slide metadata. This makes it possible to extract partial order relations, and to represent whole sequence of presentation schema by accumulating such partial relations as shown in lower right of Figure 3.

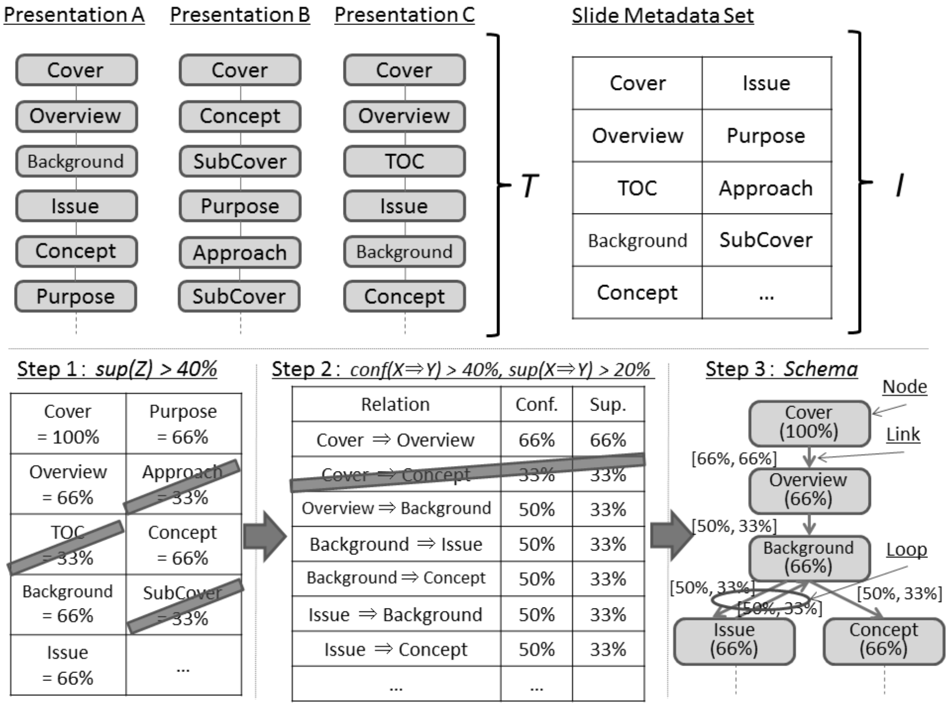
In addition, the infrequently appeared metadata are discarded in order to reduce amount of calculation. *freq*(*Z*), where $Z \subseteq I$, is defined as the proportion of the number of transactions that contain *Z* to the total number of transactions in *T*, and the threshold of frequency described in *min_freq* is also predefined. Such preprocessing would make it simple for the novices.

Based on the above assumptions, the algorithm contains the following steps:

Step 1. The algorithm extracts a set of frequently-appeared metadata that have *freq*(*Z*) larger than or equal to *min_freq*. Suppose *min_freq* is *40%* in Figure 3, then metadata *TOC*, *Approach*, and *SubCover* are discarded.

Step 2. It extracts partial order relations $X \Rightarrow Y \ (X, Y \subseteq Z)$ that satisfy both *min_conf* and *min_sup*. Suppose *min_conf = 40%* and *min_sup=20%*, the relation between *Cover* and *Concept* is discard.

Step 3. It composes a presentation schema diagram by combining the extracted metadata and relations. In the diagram as shown in Figure 3, the nodes are the slide metadata left in Step 1 and the links show the order relations left in Step 2. The loops mean dual-ordered relations such as *Background* and *Issue*, which have the links from node *Background* to *Issue* and from node *Issue* to *Background* at the same time.



Cover	Issue
Overview	Purpose
TOC	Approach
Background	SubCover
Concept	...

Cover = 100%	Purpose = 66%
Overview = 66%	~~Approach = 33%~~
~~TOC = 33%~~	Concept = 66%
Background = 66%	~~SubCover = 33%~~
Issue = 66%	...

Relation	Conf.	Sup.
Cover ⇒ Overview	66%	66%
~~Cover ⇒ Concept~~	~~33%~~	~~33%~~
Overview ⇒ Background	50%	33%
Background ⇒ Issue	50%	33%
Background ⇒ Concept	50%	33%
Issue ⇒ Background	50%	33%
Issue ⇒ Concept	50%	33%
...	...	

Figure 3. Overview of Schema Extraction Algorithm

3. Case Studies

This section describes case studies which investigated how to configure the thresholds of frequency, support and confidence in the association rule mining for extracting presentation schema, and compared the presentation schemas between different laboratories, audiences, and presentation time limitations since the presentation schema would vary according to such factors. The followings are detail information for sets of the presentation documents in these case studies.

The presentation documents accumulated in Laboratory A were final versions of the ones for graduation research of *30* undergraduate students belonged to the laboratory where they focused on development of support systems for self-directed learning, research activity, and experiential learning. The audiences of the presentations were faculties and students of their affiliation of the university, and the presentation time was *7* minutes. These documents were annotated in *30* kinds of the slide metadata to each slide by an experienced researcher in advance. The average and standard deviation for the number of the slides of the documents were *20.1* pages and *3.68*.

The presentation documents accumulated in Laboratory B were also final versions of the ones for domestic conferences of *15* graduate students or researchers belonged to the laboratory where they focused on development of web-based learning support systems and practice of distance learning systems. The audiences of the presentations were related filed researchers, and presentation time was *15 - 20* minutes. These documents were also annotated in *34* kinds of the slide metadata to each slide by the experienced researcher in advance. The average and standard deviation for the number of the slides of the documents were *22.9* pages and *3.93*.

3.1 Case Study 1: Analysis for thresholds

In order to consider the thresholds, we first investigated how the numbers of nodes, links, and loops included in a presentation schema diagram changed by values of *min_freq*, *min_sup* and *min_conf*. Figure 4 compares the numbers of them extracted by the thresholds on the abscissa in proposed schema extraction algorithm. From the results of our previous work [10], we have ascertained that *min_freq* could be set as the same value and *min_sup* could be set as the half value of *min_conf*. In this case study, we also followed this to set these thresholds.

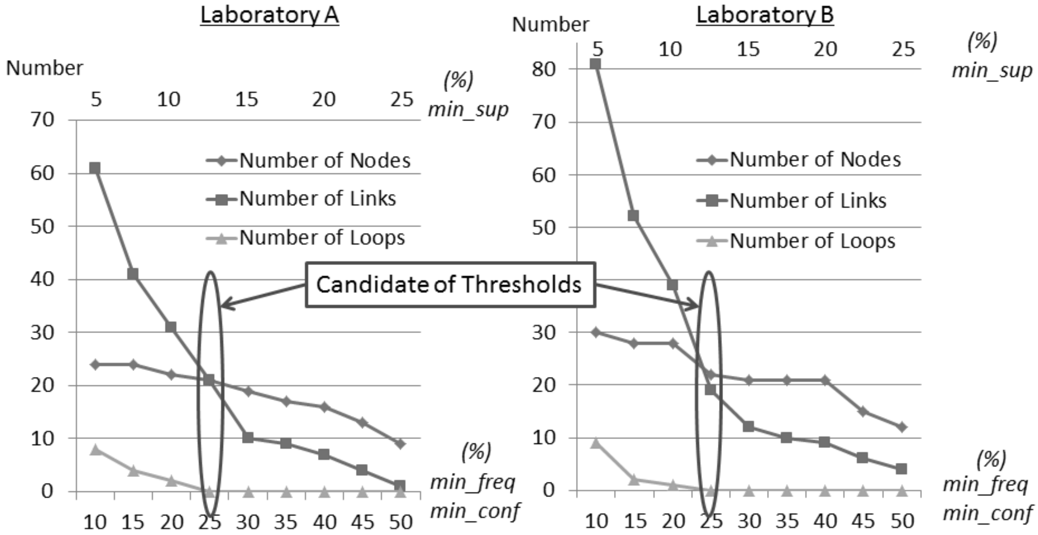


Figure 4. Numbers of Extraction by Changes in Thresholds

In case that the thresholds were sufficiently small, Figure 4 shows the numbers of links are larger than the number of nodes in both laboratories. The larger the thresholds were, the smaller the numbers of nodes, links and loops were. We can see the points (around *min_freq = min_conf = 25%* and *min_sup=12.5%*) at where the numbers of nodes and links were reversed and the numbers of loops were zero. In considering application of the presentation schema, too many links and loops may confuse the novices. Therefore, these points can be important candidates for setting the thresholds. In other words, suitable presentation schema could be obtained by finding out such points to set the thresholds.

3.2 Case Study 2: Assessment of validity for mining technique

The purpose of this case study was to assess the validity of the proposed mining technique by comparing the presentation schemas between Laboratory A and B. Figure 5 illustrates both presentation schema by setting *min_freq* = *min_conf* = *25%* and *min_sup* = *12.5%*. Values in round brackets are probabilities of appeared metadata *freq*(*Z*), and values in square brackets are probabilities of confidence $conf(X \Rightarrow Y)$. Comparing both schemas, for example, the schema regarding to *"Evaluation"* from Laboratory B was different from the one from Laboratory A. This showed a capacity of the presentation schema to represent importance of evaluation in presentation for the domestic conferences. In addition, the schema from Laboratory A tended to have a main path for making smooth presentations of graduate research. On the other hand, there were two paths found in the early segments of the schema from Laboratory B. One path was ordered by *Background*, *Issue*, *Purpose*, *Approach*, and *Technology*. Another was ordered by *Situation*, *Theory*, and *Model*. The reason would be that these presentations included not only researches for system development but also for classroom practice. Some slide metadata did not have any arrows to indicate transition. This showed there are no significant (over thresholds) transitions from the metadata because such metadata had different position in the presentation documents. Following the above consideration, we can say that the presentation schemas extracted satisfy specific conditions of the presentation contexts such as the research domain and/or philosophy each laboratory has. The proposed technique accordingly seems to be valid.

4. Conclusion

This paper has described the presentation structure model and proposed a fundamental technique to represent the presentation schema automatically by accumulating partial order relations extracted with the association rule mining. The diagram of roles of and sequence among the slides would enable the novices to be aware of the presentation schema in the laboratory's manner explicitly. Accordingly this is one of scaffolding ways for the novices to learn the presentation composition skill practically.

We have also discussed preliminary case studies. The results indicate a reasonable setting approach for the thresholds of the mining, and description capability of the schema which depends on the presentation context.

In the near future, it will be necessary to try out the proposed technique to different domain of laboratories. In addition, our research group proposed metadata recommendation, diagnosis, and learning services [2, 11] as the previous work. However, these previous services did not explicitly deal with the presentation schema, especially sequence of the presentation slides. Therefore, we will have to evaluate effectiveness of these services by adding the concept of presentation schema in a more detail.

Acknowledgements

This work is supported in part by Grant-in-Aid for Scientific Research (B) (No. 23300297), Grant-in-Aid for Scientific Research (C) (No. 23501141) from the Ministry of Education, Science, and Culture of Japan.

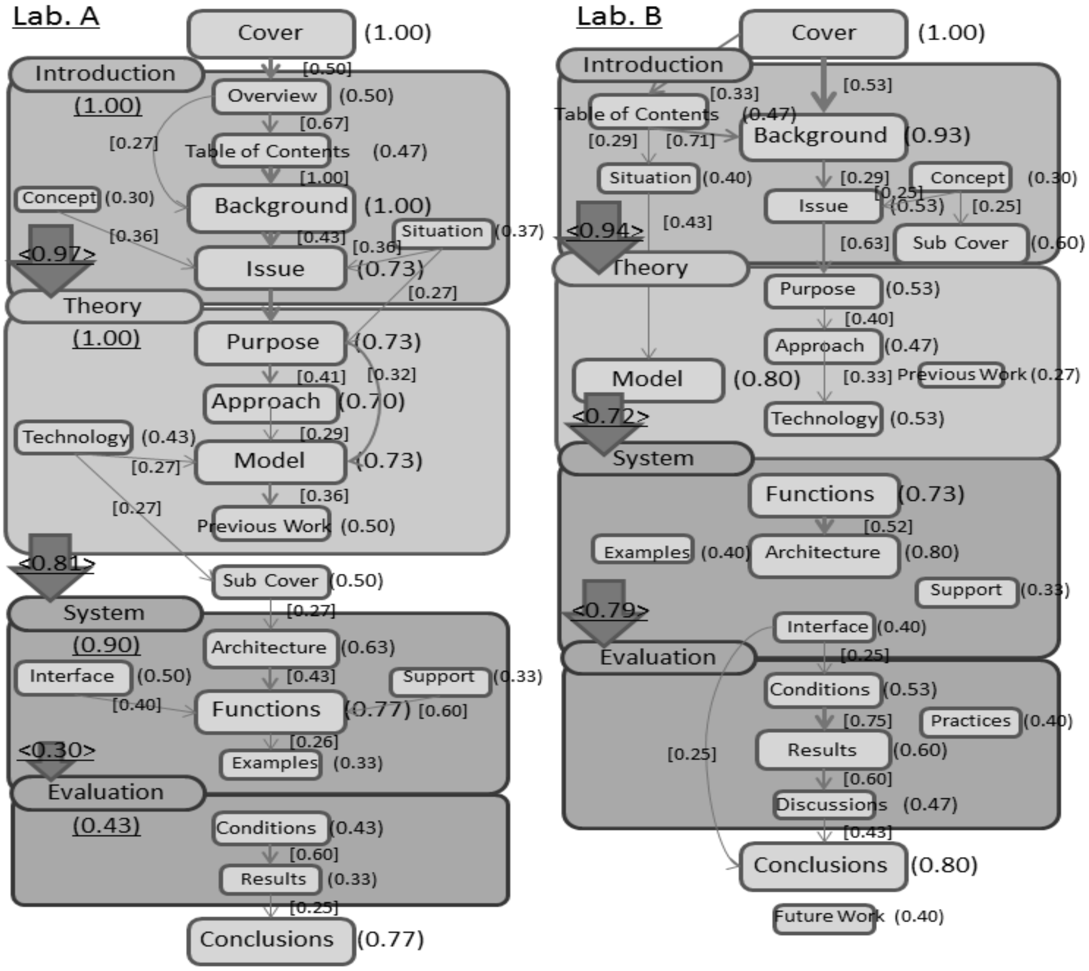


Figure 5. Results of Extracted Presentation Schemas

References

[1] Kirschner, P. A. (1988). The laboratory in Higher Science Education, Problems, Premises, and Objectives, Higher Education, 17, No 1, pp. 81-90.

[2] S. Hasegawa, and A. Kashihara. (2011). Recommendation and Diagnosis Services with Structure Analysis of Presentation Documents, Proceedings of KES2011, Part I, LNAI 6881, pp. 484-494.

[3] Convertino et al. (2004). A laboratory method for studying activity awareness, Proc. of the Third Nordic Conference on Human-Computer Interaction, pp. 313-322.

[4] Collins, A. (2006). Cognitive apprenticeship: The Cambridge Handbook of the Learning Sciences, R.Keith Sawyer (Ed.), Cambridge University Press, pp.47-60.

[5] Kohlhase. (2007). Semantic PowerPoint: Content and Semantic Technology for Educational Added-Value Services in MS PowerPoint, Proc. of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA2007), 3576-3583.

[6] T. Hayama, H. Nanba, and S. Kunifuji (2005). Alignment between a technical paper and presenta-tion sheets using a hidden Markov model, Proc. Active Media Technology 2005, pp.102-106.

[7] S. T. Li, and W. C. Chang. (2009). Exploiting and transferring presentational knowledge assets in R&D organizations, Expert Systems with Applications 36, pp. 766-777.

[8] Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[9] Kotsiantis, S., Kanellopoulos, D. (2006). Association Rules Mining: A Recent Overview, GESTS International Transactions on Computer Science and Engineering, Vol.32(1), pp.71-82.

[10] K. OTA and A. KASHIHARA. (2010). Mining Collective Knowledge for Reconstructing Learning Resource, Proc. of ICCE2010, pp.104-106.

[11] K. Saito, A. Tanida, A. Kashihara, and S. Hasegawa. (2010) An Interactive Learning Environment for Developing Presentation Skill with Presentation Schema, Proc. of E-Learn2010, pp. 2696-2703.