2011 Ninth International Conference on Creating, Connecting and Collaborating through Computing

Slide KWIC: Snippet Generation for Browsing Slides based on Conceptual Relationship and Presentational Structure

Yuanyuan Wang Kazutoshi Sumiya
University of Hyogo, Japan
nd09s005@stshse.u-hyogo.ac.jp, sumiya@shse.u-hyogo.ac.jp

***Abstract*—Currently, several presentation files such as SlideShare and MPMeister are shared by many universities over the Web. These files are useful and valuable to students. However, such files have to be retrieved for self-learning purposes, and there is still a lack of support for self-learners browsing through slides containing information that might be irrelevant to their query. We describe a slide retrieval method involving the snippet generation of target slides, and we discuss how to present the retrieval results to users by considering what portions of the slides are relevant to a user query, on the basis of the relationships between slides. This method is based on (1) the keyword conceptual structure of the conceptual relationship that implicitly exists between keywords that are extracted from the slide text of the presentation content by WordNet and (2) the presentational structure of the indent levels in the slides. With our novel snippet-generation method, users can easily determine which slides to learn by browsing the relevant portions of the desired slides in the presentation and by focusing on portions from either detailed or generalized slides. We experimentally confirmed that our method enables the users to browse snippets of the retrieved slides efficiently and effectively.**

***Keywords*-multimedia; e-learning; presentation content retrieval; snippet generation; conceptual relationship**

I. Introduction

These days, a considerable amount of lecture materials, which are often prepared by using teaching materials used in actual classes conducted in universities or other education organizations, is freely shared on websites such as SlideShare [1] and MPMeister [2]. Thus, not only students who missed a lecture or presentation but also those interested in the topic being discussed in the lecture can review it and study its content on their own, according to their convenience.

A user must formulate a query in the proper manner if he or she wishes to retrieve the required lecture slides on the basis of matching keywords. If the keywords in a query appear repeatedly, many irrelevant slides could be retrieved, and this would make it difficult to obtain an appropriate retrieval result on the basis of a query. Moreover, one of the important functions necessary for archiving presentation slides is the ability to retrieve the desired slides using the given keywords. For the benefit of users, it is essential that certain keywords are supported to enable the retrieval of important slides. However, retrieving the important slides only on the basis of certain keywords can destroy the implicit relevant information between slides and decrease the relevance of the retrieved slides to the given context, lowering the user's understanding of the information on the slides. Additionally, these methods do not consider the relevant information existing between the slides related to the query; it is impossible to easily obtain successful retrieval results through the concepts represented by a query.

We present a novel slide retrieval method to meet users' requirements using snippet generation (see Figure 1). This retrieval method can be implemented by (1) understanding the portions of slides that satisfy a user query and (2) generating snippets for the retrieved slides to present the relevant portions of slides on the basis of the relationships between the slides that include what users need to browse. To achieve our goal, we analyzed the implicit semantic relationship; in particular, we focused on the conceptual relationship between keywords, e.g., an is-a relationship. We derived a keyword conceptual structure consisting of the conceptual relationship between keywords extracted from the slide text. However, the usage of keywords in slides varies depending on the author. We derived a presentational structure by focusing on certain features of the slides, such as the level of indents in the slide text. It was then necessary to use the conceptual relationship and presentational structure to determine the portions of slides related to the user query; furthermore, we detected the relationships between slides in terms of the query.

As an example, consider a user query "vegetable" whose snippet in slide x is shown in Figure 2. In fact, some presentation slides may be related to other slides in terms of detailed and generalized information. Therefore, we generated snippets of the relevant portions of the retrieved slides on the basis of the relationships between slides related to the query. For instance, the explanation provided in slide y, "cabbage and spinach are leafy vegetables" is more likely to be more specific and detailed than the general one provided in slide x, "vegetables." Therefore, slide y has a $detailed$ relation with slide x in terms of "vegetable." In this case, a snippet for slide x would look like a portion P_x of slide x with a portion P_y of slide y related to "vegetable."

The next section describes our approach and presents the

[1]http://www.slideshare.net/
[2]http://www.ricoh.co.jp/mpmeister/

978-0-7695-4363-5/11 $26.00 © 2011 IEEE
DOI 10.1109/C5.2011.18

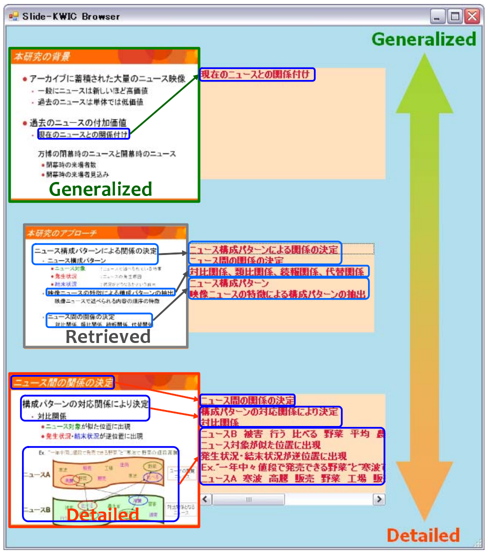

Figure 1. Screen image of a snippet for retrieved slide

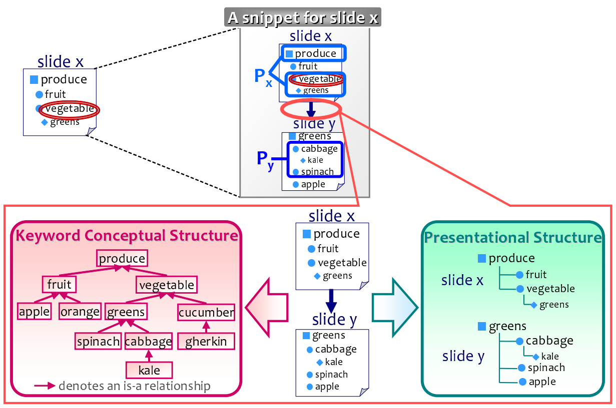

Figure 2. Snippet of slide using the relationships between slides

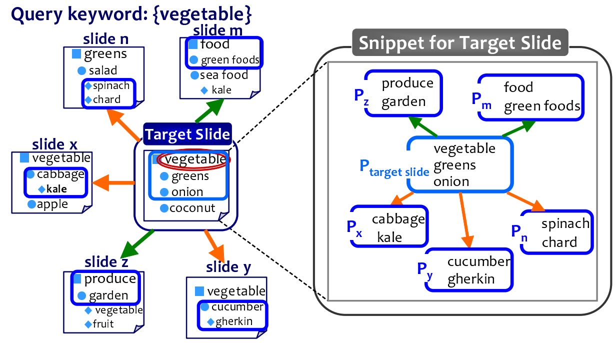

Figure 3. Concept of generating snippet for target slide

concepts of Slide KWIC as well as reviews of related work. Section 3 explains the keyword conceptual structure and presentational structure. Thus, we mathematically determine the relationships types. Section 4 describes snippet generation using the relationships between slides. In Section 5, we present an experimental evaluation of our method. Finally, Section 6 presents the conclusions of this paper.

II. Our Approach and Related Work

A. Our Approach

In this paper, we discuss how to obtain user retrieval results efficiently and effectively. A traditional snippet used to obtain a retrieval result consists of a portion of the retrieval result containing the user query with its surrounding text. We propose Slide KWIC, which helps users understand the target slides related to a user query for generating snippets. A snippet for the target slide would look like the one shown in Figure 3. It captures the portion of the target slide related to the query keyword with the portions of other slides that are related to the target slides. We consider that the retrieved slide in particular contains information that is irrelevant to the query, and we only present a portion of the retrieved slide; users might not understand the meaning of this portion of the presentation content in terms of the query. Therefore, our snippet generation method identifies a portion of the retrieved slide along with the relevant portions from other slides, on the basis of the relationship between the slides in terms of the query.

Thus, we analyze the relationship between keywords and how the keywords in different levels of indents in slides are related to a user query. Actually, there are implicit semantic relationships between keywords, such as an is-a or has-a relationship. In this study, we consider that users grasp the conceptual content easily, and hence, we focus only on an is-a relationship between keywords. We define the keyword conceptual structure of keywords extracted from the slide text using the conceptual dictionary WordNet [3], which is focused on a conceptual relationship, e.g., an is-a relationship. Additionally, we define the presentational structure of indents in the slide text. As mentioned above, we believe that an efficient Slide KWIC should extract the portions of the retrieved slides that provide relevant information about a user query. On the other hand, we generate snippets of the relevant portions of the retrieved slides that help users to browse the retrieval results in terms of the query from the presentation content.

B. Related Work

Most of the research related to academic content has been focused on the retrieval of slides. Yokota et al. [1] proposed a system named Unified Presentation Slide Retrieval by Impression Search Engine (UPRISE) for retrieving a sequence of lecture slides from archives containing a combination of slides and recorded videos. Kobayashi et al. [2] proposed a method based on the use of laser pointer information for

[3]http://wordnet.princeton.edu/

retrieving lecture slides by UPRISE. Le et al. [3] proposed a method for extracting important slides by automatically generating digests from recorded presentation videos. Their method extracts important slides from unified content on the basis of the metadata features of a single medium or two heterogeneous media. However, we considered that retrieving only the important slides decreases the relevance of the results of a user query to the given context, and their method cannot be used to browse important slides containing information related to a query. This does not enhance the user's understanding. Therefore, our objectives are to effectively retrieve slides by implicitly accumulating the relevant information between slides in terms of a user query and to generate snippets of the retrieved slides using the relationships between the slides related to the query.

Kushki et al. [4] proposed a novel XML-based system for the retrieval of presentation slides. This system analyzing contextual information, such as structural and formatting features, is extracted from the open format XML representation of slides. Our complementary method considers both the structural features of the slides and the semantic information, such as the conceptual relationship between keywords in the slide text, and it analyzes these two features to determine the relationships between slides. Kitayama et al. [5] proposed a method for extracting slides on the basis of their semantic relationships and roles. Wang et al. [6] presented a method for automatically generating learning channels using the semantic relationships in slides of a lecture that had an accompanying recorded video. These studies are similar to ours, in that the researchers have proposed a method for the retrieval of slides using the relationships between slides. Our method is an extension of these methods by generating snippets on the basis of the relationships between slides.

Parapar et al. [7] proposed a method for snippet generation for Blog Search based on sentence selection, using comments to guide the selection process. We propose a snippet generation method for slide retrieval based on the relevant portions of slides, and we use the conceptual relationship between keywords and the presentational structure of indents to detect the relationships between slides. Penin et al. [8] extended existing work on ontology summarization to support the presentation of ontology snippets for semantic web search engines, and the proposed solution leveraged a new semantic similarity measure to generate snippets on the basis of the given query. This ontological method is similar to our method; we focus on the keyword conceptual structure as ontology. In addition, Penin et al. [8] computed the similarity between bags of words to compare sentences and thus generate snippets that displayed query related topics and sentences. However, our snippet generation is for slides that are not semantic web documents, and we determine the relationships between slides using the conceptual relationships and presentational structure, to provide snippets of the portions of slides that are only relevant to the given query.

III. Determining Relationships between Slides using Conceptual Relationship and Presentational Structure

A. Keyword Conceptual Structure and Presentational Structure

We consider that a semantic relationship implicitly exists between keywords extracted from slide text. In particular, the conceptual relationship is called an is-a relationship [9], [10] and it is used as a basis for the semantic relationship between keywords. "X subsumes Y, or Y is-subsumed-by X" (Y is-a X) usually means that concept Y is a specialization of concept X and that concept X is a generalization of concept Y. For example, a "fruit" is a generalization of an "apple," an "orange," a "mango," and many other fruits; in other words, an apple is a fruit (apple is-a fruit). Therefore, we define a keyword conceptual structure as consisting of an is-a relationship between keywords extracted using WordNet.

We define a presentational structure as a slide that appears in the outline pane, on the basis of indents in the slide text extracted from the Office Open XML in Microsoft Office 2007. The slide title (1st level indent) is the upper level. The first item of the text is on the 2nd level, and the depth of the subitems increases with the level of indentation (3rd level, 4th level, and so on). Indents outside the text, such as figures or tables, are on the average level of the slide. If a given keyword appears in the title of the slide or in lines with smaller indents, we implicitly assume that the lower-level indented keywords are supplementary and that they explain the upper-level keywords.

Therefore, the conceptual relationship between keywords and the level of indents in a slide should both be considered for slide retrieval.

B. Determination of Relationship Types

We define a focused slide that a slide in question and other slides that have specific relationships as being conceptually related to the focused slide through one of two types of relations: $detailed$ and $generalized$. If a slide has a $detailed$ relationship with the other slides, it is called a $detailed$ slide. If a slide has a $generalized$ relationship with the other slides, it is called a $generalized$ slide. In our previous work [11], we developed a presentation retrieval engine based on the relationships between slides and evaluated its effectiveness through experiments. This section briefly explains the manner in which the types of relationships are determined. Let x be the number of a focused slide and y be the number of the slide that we want to retrieve. Slide x contains keywords k_i and k_j. The types of relationships are determined for all slides for keyword q in a user query.

1) Determination of Detailed Relationships: If a slide has more information about a user query than the focused slide, its relationship with the focused slide is a $detailed$ one. We explain the determination of $detailed$ slides using the query

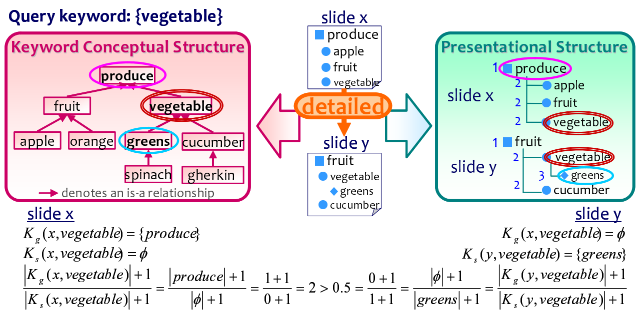

Figure 4. Example of detailed relationship between slides

keyword q that is present in the focused slide x and slide y, which needs to be retrieved. Figure 4 shows an example of the determination of the $detailed$ relationship between slide x and slide y for a query on the word “vegetable.”

When q and other keywords in slide x and slide y satisfy certain conditions, slide y is determined to be the detailed slide of slide x. This is because the amount of content in slide y that is specific to q is greater than that in slide x.

$$K_g(x, q) = \{k_i | k_i \in x, l(q) \geq l(k_i), q \text{ is-a } k_i\} \quad (1)$$

$$K_s(x, q) = \{k_j | k_j \in x, l(q) < l(k_j), k_j \text{ is-a } q\} \quad (2)$$

Here, $K_g(x, q)$ is a set of keywords in slide x. The levels of the keywords in this set are not lower than the level of q in the presentational structure, and q has an is-a relationship with each one of them in the keyword conceptual structure. In Eq. (1), k_i, (e.g., “produce”) belongs to the set of keywords $K_g(x, q)$ in slide x, its level $l(k_i)$ is not lower than the level $l(q)$ of q (e.g., “vegetable”) in the presentational structure, and q is-a k_i in the keyword conceptual structure (see Figure 4). In our method, the keyword conceptual structure is extracted as a tree-shaped structure. In general, an is-a relationship between keywords is equivalent to a parent-child relationship, and our method may classify an is-a relationship as a descendent relationship. $K_s(x, q)$ is a set of keywords in slide x, the levels of the keywords are lower than the level of q in the presentational structure, and each keyword has an is-a relationship with q in the keyword conceptual structure. In Eq. (2), k_j (e.g., “greens”) belongs to the set of keywords $K_s(x, q)$ in slide x, its level $l(k_j)$ is lower than the level $l(q)$ of q (e.g., “vegetable”) in the presentational structure, and k_j has an is-a relationship with q in the keyword conceptual structure (see Figure 4).

$$\frac{|K_g(x, q)| + 1}{|K_s(x, q)| + 1} > \frac{|K_g(y, q)| + 1}{|K_s(y, q)| + 1} \quad (3)$$

where the function $|K_g(x, q)|$ extracts the total number of k_i in $K_g(x, q)$, and $|K_s(x, q)|$ extracts the total number of k_j in $K_s(x, q)$ in slide x. $K_g(y, q)$ is also a set of keywords in slide y, and it satisfies the same conditions as $K_g(x, q)$ in Eq. (1); $K_s(y, q)$ is a set of keywords in slide y, and it satisfies the same conditions as $K_s(x, q)$ in Eq. (2). Thus, Eq. (3) can be used to calculate the ratio of $|K_g(x, q)|$ to $|K_s(x, q)|$ for slide x and the ratio of $|K_g(y, q)|$ to $|K_s(y, q)|$ for slide y. If the ratio calculated for slide x is higher than that calculated for slide y using Eq. (3), slide y is determined to be the detailed slide of slide x with regard to q.

2) Determination of Generalized Relationships: If a slide contains content about the query in the outline given in a generalized slide, it is described in relation to the focused slide. We explain the determination of $generalized$ slides using the query keyword q that is present in the focused slide x and slide y, which needs to be retrieved.

$$\frac{|K_g(x, q)| + 1}{|K_s(x, q)| + 1} < \frac{|K_g(y, q)| + 1}{|K_s(y, q)| + 1} \quad (4)$$

When the query keyword q and other keywords in slide x and slide y satisfy Eqs. (1), (2), and (4), then slide y is determined to be a generalized slide of slide x. This is because the amount of content in slide y that is general with respect to q is greater than that in slide x. Eq. (4) can be used to calculate the ratio of $|K_g(x, q)|$ to $|K_s(x, q)|$ for slide x and the ratio of $|K_g(y, q)|$ to $|K_s(y, q)|$ for slide y. When the ratio calculated for slide x is lower than that calculated for slide x using Eq. (4), slide y is determined to be the generalized slide of slide x with regard to q.

As can be seen, detailed and generalized slides are functionally interchangeable, whereas a focused slide is a generalized slide from the viewpoint of a detailed slide.

C. Expansion of Query

Some slides have relationships with the focused slides but do not contain a user query. It is possible to expand a user query to detect the relationships between slides in terms of the query. In this section, a method for expanding a user query using the keyword conceptual structure and the presentational structure that is explained. In this method, the query keyword q is expanded in two ways with the aim of determining $detailed$ relationships and $generalized$ relationships. Let x be the number of a focused slide containing the specified keyword q_s and the generalized keyword q_g of the query keyword q (this means that q_s is-a q and q is-a q_g). If the distance between the position of $q_s(q_g)$ and q in both the presentational structure and in the keyword conceptual structure is only one, the relevance of $q_s(q_g)$ and q is high, and hence, the meanings of q_s (q_g) and q are very nearly identical.

In the case of $detailed$ relationships, if a slide does not contain a user query but has specific, detailed information in common with the focused slide in terms of the query, its relationship with regard to the focused slide is a $detailed$ one. We explain query expansion to determine the $detailed$ relationship using the specified keyword q_s of the query keyword q present in the focused slide x conforms to certain conditions; q_s is an expanded query keyword of q.

$$Q_s(x,q) = \{q_s | q_s \in x, l(q_s) - l(q) = 1, \\ pos(q_s) - pos(q) = 1, q_s \text{ is-a } q\} \quad (5)$$

q_s belongs to the set of keywords $Q_s(x,q)$ in the focused slide x. The function $l(q_s) - l(q) = 1$ extracts the q_s whose level $l(q_s)$ is lower than the level $l(q)$ of q by only one level in the presentational structure, and $pos(q_s) - pos(q) = 1$ extracts the q_s whose position $pos(q_s)$ is lower than the position $pos(q)$ of q by only one level in the keyword conceptual structure. When q and q_s used in slide x and slide y satisfy Eqs. (1), (2), (3), and (5), slide y is determined to be the detailed slide of slide x.

In the case of *generalized* relationships, if a slide does not contain a user query but contains content regarding the focused slide in terms of the query in the outline, its relationship with the focused slide is a *generalized* one. We explain query expansion to determine the *generalized* relationship by using the generalized keyword q_g of the query keyword q present in the focused slide x that satisfies certain conditions; q_g is an expanded query keyword of q.

$$Q_g(x,q) = \{q_g | q_g \in x, l(q) - l(q_g) = 1, \\ pos(q) - pos(q_g) = 1, q \text{ is-a } q_g\} \quad (6)$$

Here, q_g belongs to the set of keywords $Q_g(x,q)$ in the focused slide x, its level is higher than the level of q by only one level in the presentational structure, and its position is higher than the position of q by only one level in the keyword conceptual structure. When q and q_g in slide x, and q_g in slide y satisfy Eqs. (1), (2), (4), and (6), slide y is determined to be a generalized slide of slide x. Based on these two cases, we can expand the query keyword for retrieval to detect the relationships between slides related to the query keyword.

IV. Snippet Generation using the Relationships between Slides

To generate snippets, Slide KWIC takes the relevant portions of the retrieved slides in terms of a user query using the relationships between slides. It is difficult for users to understand the relevant information between portions of slides in terms of the query. For example, a user wants to study slide 4 to further understand "vegetable" in the lecture content about Nutrition. Our method generates a snippet for slide 4 using a portion P_4 of slide 4 with a portion P_2 of slide 2 and a portion P_5 of slide 5 (see Figure 5). In this case, slide 2 explains that "produce vegetables" has a generalized relationship with slide 4 in terms of "vegetable," and slide 5 explains that "cabbage and spinach are leafy vegetables," as a detailed relationship with slide 4 in terms of "vegetable." When the user browses the snippet for slide 4 that consists of a portion P_4 of slide 4 with the relevant portions P_2 and P_5 from slide 2 and slide 5, he or she

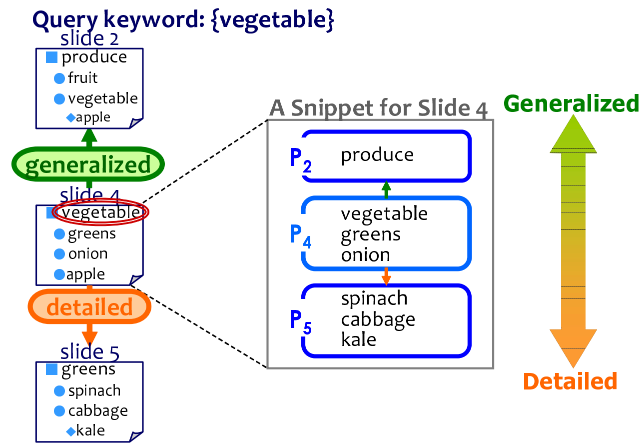

Figure 5. Example of snippet generation

gets more information on "vegetable" from slide 4, and this enables the user to further his or her understanding easily. Therefore, our snippet generation method is based on the concept of slides that present snippets containing related portions in terms of a detailed, ordered user query. This section describes how to generate snippets on the basis of the relationships between slides related to the query by the following procedures:

A. Identifying the Portions of Retrieved Slides

Our method can retrieve slides related to a user query. However, whether the information contained on the retrieved slides is relevant or irrelevant to the query must be determined. Therefore, our method first identifies the portions of the retrieved slides related to a user query on the basis of the keyword conceptual structure and presentational structure. Let x be the number of the retrieved slide containing keywords k_i and k_j. When the query keyword q and other keywords in slide x satisfy Eqs. (1), (2), (7), (8), and (9), the portion P of slide x is determined to be related to q.

$$L_g(x,q) = \{l_n | l(k_i) \leq line(l_n) \leq l(q), k_i \in K_g(x,q)\} \quad (7)$$

$$L_s(x,q) = \{l_m | l(q) \leq line(l_m) \leq l(k_j), k_j \in K_s(x,q)\} \quad (8)$$

$$P = L_g(x,q) \cup L_s(x,q) \quad (9)$$

where the function $L_g(x,q)$ extracts a set of texts l_n in the slide x such that the levels of the texts range from the depth of the level containing q to the depth of the level containing the generalized keyword k_i of q that belongs to the set of keywords $K_g(x,q)$ in slide x; the extraction is performed using Eq. (1). In Eq. (7), $l(q)$ is the depth of the level containing q in slide x, and its depth is not less than the depth $line(l_n)$ of the text from the given level; in addition, $l(k_i)$ is the depth of the level containing k_i in slide x and its depth is not greater than the depth $line(l_n)$ of the text from the given level. $L_s(x,q)$ extracts a set of the texts

l_m in slide x such that the texts of the levels ranges from the depth of the level containing q to the depth of the level containing the specified keyword k_j of q, that belongs to the set of keywords $K_s(x, q)$ in slide x; the extraction is performed using Eq. (2). In Eq. (8), $l(q)$ is the depth of the level containing q in slide x and its depth is not greater than the depth $line(l_m)$ of the text from the given level; in addition, $l(k_j)$ is the depth of the level containing k_j in slide x and its depth $line(l_m)$ is not less than the depth of the text from the given level. Thus, Eq. (9) can be used to extract the portion P of slide x and thus combines the sets of texts from levels $L_g(x, q)$ and $L_s(x, q)$.

B. Determining the Relevant Portions of Related Slides

Our method generates snippets on the basis of the relationships between slides related to a user query. The relevant portions are then extracted from slides that have relationships with the retrieved slides in terms of the query.

1) Determining the Portions of Generalized Slides: When slide x_g is a generalized slide that has a generalized relationship with the retrieved slide x related to the query keyword q, the portion P_g of slide x_g provides the general content of the portion P of the retrieved slide x related to q. Therefore, the portion P_g of the generalized slide x_g is determined using the query keyword q and the generalized keyword k_i is determined from the retrieved slide x.

$$P_g = L_g(x_g, k_i) \cup L_g(x_g, q) \quad (10)$$

When q and other keywords in slide x_g satisfy Eqs. (1), (2), (7), and (10), then the portion P_g of the generalized slide x_g is determined. This is because slide x_g contains more general content related to q than does slide x. For slide x_g containing the generalized keyword, k_i is determined from the retrieved slide x and q. When the respective functions $L_g(x_g, k_i)$ and $L_g(x_g, q)$ are used to extract a set of the texts l_n from levels in that slide and satisfy the same conditions as the function $L_g(x, q)$ (these conditions apply to slide x and are given by Eq. (7)), Eq. (10) can be used to determine the portion P_g of slide x_g that combines the sets of text having levels $L_g(x_g, k_i)$ and $L_g(x_g, q)$.

2) Determining the Portions of Detailed Slides: When slide x_d is a detailed slide that has a detailed relationship with the retrieved slide x related to the query keyword q, the portion P_d of slide x_d provides specific, detailed information about the portion P of the retrieved slide x related to q. Therefore, we determine the portion P_d of the detailed slide x_d using the query keyword q, and the specified keyword k_j is determined from the retrieved slide x.

$$P_d = L_s(x_d, q) \cup L_s(x_d, k_j) \quad (11)$$

When q and other keywords in slide x_d satisfy Eqs. (1), (2), (8), and (11), the portion P_d is determined from the detailed slide x_d. This is because the amount of content in slide x_d that is specific to q is greater than that in slide x. For slide x_d containing the specified keyword, k_j is determined from the retrieved slide x and q. When the respective functions $L_s(x_d, k_j)$ and $L_s(x_d, q)$ are used to extract a set of the texts l_m from levels in that slide and satisfy the same conditions as the function $L_s(x, q)$ (these conditions apply to slide x and are given by Eq. (8)), Eq. (11) can be used to determine the portion P_d of slide x_d that combines the sets of text having levels $L_g(x_d, q)$ and $L_g(x_d, k_j)$.

As mentioned above, our method for generating snippets of the retrieved slides satisfies user demand by relating portions of the generalized, retrieved, and detailed slides to provide content that varies from generalized to detailed content in terms of a user query for specific content.

Table I
RESULTS OF IDENTIFYING THE PORTIONS OF SLIDES

	Our method				Matching keywords
	Content X	Content Y	Content Z	All contents	All contents
Precision	73.4% (149/203)	62.3% (48/77)	69.2% (63/91)	70.1% (260/371)	69.8% (215/308)
Recall	79.7% (149/187)	92.3% (48/52)	46.7% (63/135)	69.5% (260/374)	57.5% (215/374)
F-measure	0.76	0.74	0.56	0.70	0.63

V. EVALUATION

A. Experiment 1: Validity of Identifying the Portions of Slides

Seven participants freely captured portions of text from the indents of slides and assessed 16 sets of slides containing four query keywords from three sets of actual academic content [4]: content X has 22 slides that explain "Tree patterns," content Y has 24 slides that explain "Broadcasting Geographic Information," and content Z has 7 slides that explain "News Structure Patterns." A correct answer was defined as a portion where four or more test participants found the text of indents in the slides that they had captured. In this study, we evaluated the validity of the rules for identifying the portions of slides in terms of the query keywords by precision and recall using the results obtained by our method and those obtained from participants who gave correct answers. The results of the portions of slides related to the given query keywords are listed in Table I.

We calculate the precision and recall of portions in each set of content X, Y, and Z. Additionally, in this experiment, we compared the results containing portions of slides in all sets of content extracted by our proposed method and the corresponding results obtained by the matching keywords method. These experimental results show the recall of all sets of experimental academic content was low. In particular, the recall of content Z was low, and according to our method, many of correct answers did not contain keywords related to

[4]DBSJ Archives:
http://www.dbsj.org/Japanese/Archives/archivesIndex.html

Table II
RESULTS OF GENERATING SNIPPETS

	Content X		Content Y	Content Z	All snippets
	Snippet A	Snippet B	Snippet C	Snippet D	
Precision	70.0% (49/70)	76.7% (46/60)	62.0% (31/50)	78.3% (47/60)	72.1% (173/240)
Recall	66.2% (49/74)	66.7% (46/69)	91.2% (31/34)	51.6% (47/91)	64.6% (173/268)
F-measure	0.68	0.71	0.74	0.62	0.68

the query keywords. We believe that one of the reasons for obtaining these results is that our method uses only an is-a relationship between keywords, and not a has-a relationship. For instance, while the keywords, "subject" and "status" are likely to be part of the query keyword "structure," since they would be used in more detailed explanations of "structure," our method based on WordNet cannot extract the has-a relationship between them. However, the recall of content Y was very high, and the number of results containing portions extracted by our method was greater than correct answers. We believe that the participants did not consider the title of the slides in terms of the query keywords when our method was used. This reduced the precision of our method. Although the results of the comparison of the two methods look similar, the methods did not extract portions of slides that explained the query keywords, even though they did not contain them; this is because our method does not extract portions of slides only on the basis of the query keywords. From the results of this experiment, we confirmed that our method can extract the appropriate portions of slides using the conceptual relationship between keywords and the presentational structure of indents. However, we would like to use an enhanced method for extracting not only the is-a relationship between keywords but also the has-a relationship between keywords.

B. Experiment 2: Validity of Generating Snippets

We showed the participants the following four data sets that are composed of four query keywords in the slides retrieved from the three sets of actual academic content used in Experiment 1. We then presented four snippets of the relevant portions and the portions of four slides pertaining to the four query keywords; the snippets presented a detailed explanation of the keywords in the order of the relevant portions in the slides. Five participants took part in this experiment. A correct answer was defined as snippets of the portions of slides where three or more test participants found the snippets that they presented in their free descriptions.

As shown by the results in Table II, note that the recall of almost all experimental snippets was low here, e.g., the recall of snippet D for slide 3 with the query keyword "structure" (see Figure 6). When our method was used in experiment 1, it was determined that many of the correct answers did not contain the text at indents C_7, C_3, and

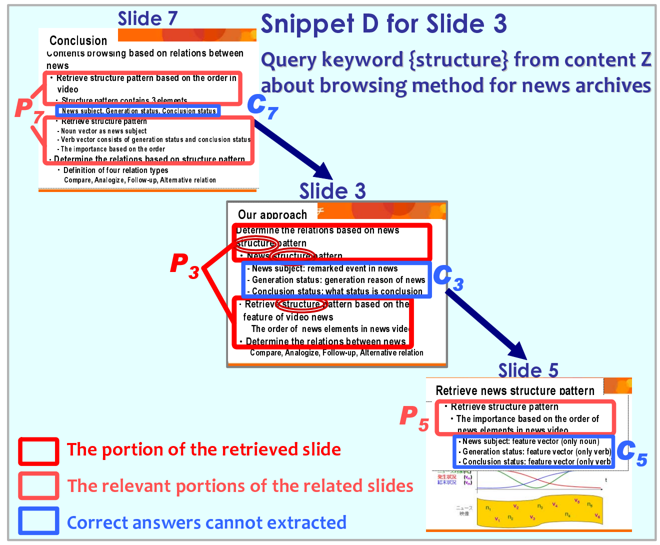

Figure 6. Snippet D cannot extract some correct answers by our method

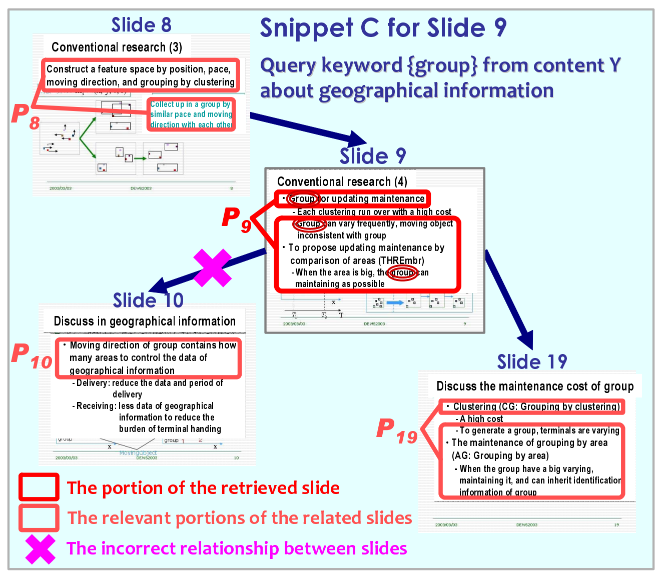

Figure 7. Snippet C extract incorrect relationship between slides by our method

C_4, namely, the text about "news subject, generation status, and conclusion status in news structure pattern" related to "news structure pattern" in portions P_7, P_3, and P_4 from slides 7, 3, and 4. Our method based on WordNet did not determine that "subject" or "status" has a part-of relationship with "structure"; this was one of the reasons why the recall was low. However, the precision of almost all snippets in the present experiments was high. These results indicate that our method generates snippets of the relevant portions of slides on the basis of the relationships between slides and that the method can be successfully applied to slide retrieval based on the use of snippets containing the portions that contain detailed explanations of the content related to the query. However, the precision of snippet C was low since it does not explain the content related to the query in detail. Figure 7 shows snippet C for slide 9 for the query keyword "group"; our method determined that slide 10 has a detailed relationship with slide 9 in terms of "group," but the portion P_{10} about " direction of motion of the group" did not explain "group" in detail in the portion P_9 by participants' descriptions. We believe that one reason for obtaining this result is that the participants did not recognize the relationships between slides from which certain relevant portions were present in the snippets. In addition, many of the relevant portions were not strongly related to the portion of the retrieved slide, and this reduced the precision. It can be seen that our method can generate snippets of the relevant portions of slides related via the query by effectively using the relationships between slides. The results of this experiment suggest that we need to improve the snippet-generating algorithm using the relationships between slides and extracting the relevant portions of slides.

VI. Concluding Remarks

We have proposed Slide KWIC using conceptual relationships and presentational structure. Our snippet-generation method is used to retrieve desired slides and generate snippets of retrieved slides on the basis of the relationships between slides with regard to a user query. The type of relationship is based on the conceptual relationships between keywords and the presentational structure of indents in the slide text. In several experiments, our method was able to extract relevant portions by using conceptual relationships and presentational structure, and the portions cannot be extracted using the matching keywords method. Moreover, we evaluated the validity of generating snippets; however, we must be able to evaluate the efficacy of browsing snippets when users browse slides with their snippets.

In the future, we plan to finish developing a prototype system and evaluate it with a large set of actual lecture data. Our method can enhance retrieval techniques if a user proposes a query that includes two or more keywords; the relationship between the keywords in the query needs to be determined to retrieve the user's desired slides by analyzing the relevance of the queried keywords. We can also extend this approach to use not only an is-a relationship but also a has-a relationship between keywords; further, we can also use the presentational structure to determine the relationships between slides.

Acknowledgment

This research was supported in part by a Grant-in-Aid for Scientific Research (B)(2) 20300039 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

[1] H. Yokota, T. Kobayashi, H. Okamoto, and W. Nakano, "Unified contents retrieval from an academic repository," in *Proc. of International Symposium on Large-scale Knowledge Resources (LKR 2006)*, March 2006, pp. 41–46.

[2] T. Kobayashi, W. Nakano, H. Yokota, K. Shinoda, and S. Furui, "Presentation scene retrieval exploiting features in videos including pointing and speech information," in *Proc. of International Symposium on Large-scale Knowledge Resources (LKR 2007)*, March 2007, pp. 95–100.

[3] H.-H. Le, T. Lertrusdachakul, T. Watanabe, and H. Yokoda, "Automatic digest generation by extracting important scenes from the content of presentations," in *Proc. of the 19th International Conference on Database and Expert Systems Application (DEXA 2008)*, Septemper 2008, pp. 590–594.

[4] A. Kushki, M. Ajmal, and K. N. Plataniotis, "Hierarchical fuzzy feature similarity combination for presentation slide retrieval," *EURASIP Journal on Advances in Signal Processing archive*, vol. 2008, no. 188, January 2008.

[5] D. Kitayama, A. Otani, and K. Sumiya, "A scene extracting method based on structural and semantic analysis of presentation content archives," in *Proc. of the 7th International Conference on Creating, Connecting and Collaborating through Computing (C5 2009)*, January 2009, pp. 275–280.

[6] Y. Wang, D. Kitayama, R. Lee, and K. Sumiya, "Automatic generation of learning channels by using semantic relations among lecture slides and recorded videos for self-learning systems," in *Proc. of the 11th IEEE International Symposium on Multimedia (ISM 2009)*, December 2009, pp. 275–280.

[7] J. Parapar, J. López-Castro, and A. Barreiro, "Blog snippets: a comments-biased approach," in *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, July 2010, pp. 711–712.

[8] T. Penin, H. Wang, T. Tran, and Y. Yu, "Snippet generation for semantic web search engines," in *Proc. of the 3rd Asian Semantic Web Conference on The Semantic Web (ASWC 2008)*, December 2008, pp. 493–507.

[9] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[10] G. A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM (CACM)*, vol. 38, no. 11, pp. 39–41, November 1995.

[11] Y. Wang and K. Sumiya, "Semantic ranking of lecture slides based on conceptual relationship and presentational structure," in *Proc. of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010)*, September 2010, pp. 2801–2810.