特別研究報告書

教材スライド間の類似性に基づく講義の構造分析

指導教官　　美濃 導彦 教授

京都大学工学部情報学科

三木 健司

平成16年2月2日

教材スライド間の類似性に基づく
講義の構造分析

三木 健司

内容梗概

近年、講義の映像化のために自動撮影システムの研究が行われてきており、これらの研究により、人手による負担を少なくし、講義状況を記録できるようになってきた。また、それらの映像コンテンツをいつでも自由に利用できるようになってきた。これらの研究においては講師および生徒の動きや状態に基づいた状況理解を行い、撮影の方法を決定する手法や、講義資料を記録する手法が提案されている。しかしながら、講義中の重要な状況の推定に関する情報を記録する研究はなされていない。講義状況の撮影や講義資料の記録に加えて、重要な状況がどの部分であるか推定が可能となれば、映像コンテンツをより有効利用することができるようになる。そこで、本稿では講義中の重要な状況の推定に関する情報を記録するための、講義構造の分析を目的とする。

講義状況には教材の使い方や時間などといった講義を構成する要素の構造と関連があるとあると考え、講義の要素としての教材に注目する。講義において利用される教材として板書や配布資料、パワーポイントスライドなどが挙げられるが、このような教材は講師によって説明される内容の一部あるいは、内容を要約したものである。本手法においては、教材としてパワーポイントスライドを用いる講義を対象とする。これは、パワーポイントスライドを用いる講義では次のような特徴が挙げられるためである。

- 関連するスライドを提示するなどの内容に合わせたスライドの切り替えが容易であるため、講義における内容の推移がスライドの推移として表れる
- 電子的に記録することが容易である

具体的な手法としては次のようになる。スライドの推移から構造分析を行っていくため、スライドそのものとスライドの切り替えに注目する。各スライドが異なる内容を説明している場合、全てのスライド間の切り替えが示しているのは内容の切り替わりである。しかしながら、何枚かのスライドである内容を説明している場合はそのスライド間における切り替えが示すのは内容の切り替わりではない。そこでスライド間の関係からスライドの内容的なグループ化を

行うことで講義の内容的な区切りを検出することを目指す。まず、このようなスライド間の関係を得るためには、スライドそのものに注目し、それぞれのスライドを特徴付ける必要がある。そこで、スライドに含まれるキーワードを形態素解析により抽出し、キーワードに対して TF ＊ IDF 法を用いて、スライドの特徴を求めた。次に、この特徴を利用し、スライド間の関係としてスライド間の類似度を定義し、それを利用することで類似した内容を説明しているスライドの内容的なグループ化を行う。これにより各スライドが所属するグループについての情報が得られるため、講義におけるスライドの提示順序にしたがってスライドの所属グループの推移を調べることで、内容的な区切りを検出する。スライドの推移において、内容的な区切りの分布に注目することで講義の構造分析を行い、重要区間の推定を行う。また、キーワードに基づいた各スライドの得点を定義し、スライドの重要度とした。

以上で述べた本手法の有効性を示すため、実際に教材としてパワーポイントスライドを利用する講義に対して実験を行った。内容的な区切りの分布による構造分析を行った場合に重要であると推定された区間に含まれるスライドに対して、スライドの得点を調べたところ、その講義において高得点の部類に含まれる事が分った。これにより、重要な部分の候補となるスライドの提示区間を推定するための、講義の構造分析が出来ることを確認した。

an analysis of a lecture's structure based on the similarity between the slides used at the lecture

Kenji MIKI

Abstract

In recent years, research of automatic shooting systems is done to get video contents of lectures.By these researches, the burden by the help is lessened and a lecture's situation can be recorded now. Moreover, those video contents can be used now freely . In these researches, the situation understanding based on a motion and condition of a lecturer and a student is performed, and the technique of determining the method of shooting and the technique of recording lecture data are proposed. However, the research which records the information about presumption of the important situation in the lecture is not made. In addition to shooting a lecture situation, or record of the data used in the lecture ,if the presumption of an important situation becomes possible, such video contents can be used more effectively. Then, it aims at the analysis of lecture structure for recording the information about presumption of the important situation under lecture in this paper.

we thinks that the situation of a lecture is related to the structure of an element which constitute the lecture, such as how to use teaching materials for a lecture situation and time. So we notes the teaching materials as an element of a lecture. Although blackbord, distribution data, the PowerPoint slide, etc. are mentioned as teaching materials used in a lecture, such teaching materials summarize some contents explained by the lecturer or the contents. In this techniques,we regard the lecture using the PowerPoint slide as teaching materials as an object. This is because the following features are mentioned at the lecture which uses the PowerPoint slide.

- Since we can easily change the slide for a related slide acording to the transition of the contents, transition of the contents in a lecture appears as transition of a slide.
- we can easily reecord the slide's data of the lecture electronically .

The concrete technique is as follows. In order to perform analysis of lecture's structure from transition of slides, we note the transition of the slide and slide

itself. when each slide explains difference contents , all of the changes of showing slide mean the change of contents. However, when the contents which are explained by using more than one slide , we can't easily deciede the change of contents from the change of slide. Then, it aims at detecting a contents-pause of a lecture by grouping the slides based on the relation between the slides. First, in order to obtain the relation between such slides, each slide needs to be characterized by using the feature of the slide itself. Then, the morphological analysis extracted the keyword contained in a slide, and the feature of a slide was given by using the TF*IDF method . Next, by using this feature, the degree of similarity between slides is defined as a relation between slides, and we group the slides explaining the contents which were similar by using it. Since the information about a group that each slide belongs is acquired, a contents-pause is detected by investigating transition of the group of a slide according to the presentation order of the slide in a lecture. In transition of a slide, the analysis of a lecture's structure is performed by using the distribution of a contents-pause, and the important section is presumed. Moreover, the score of each slide based on the keyword was defined, and it considered as the importance of a slide.

In order to show the validity of this technique described above, it experimented to the lecture which actually uses the PowerPoint slide as teaching materials. When the analysis of the lecture' structure based on the distribution of a contents-pause was performed and the score of a slide was investigated to the slide included in the section presumed to be important, it turns out that it is contained in the category of a high score in the lecture. It checked that the analysis of a lecture's structure for presuming the presentation section of a slide which serves as a candidate of an important portion could be performed.

教材スライド間の類似性に基づく講義の構造分析

目次

第1章　はじめに

近年、講義の映像化のために自動撮影システムの研究が行われてきており、これらの研究により、人手による負担を少なくし、講義状況を記録できるようになってきた。[1] また、それらの映像コンテンツをいつでも自由に利用できるようになってきた。これらの研究においては講師および生徒の動きや状態に基づいた状況理解を行い、撮影の方法を決定する手法や、講義資料を記録する手法が提案されている。しかしながら、講議中の重要な状況の推定に関する情報を記録する研究はなされていない。ここで、講義状況の撮影や講義資料の記録に加えて、重要な状況がどの部分であるかを推定することが可能となれば、その映像コンテンツを利用するメリットとして以下のようなことが挙げられる。利用者側としては、どの部分が重要度であるかの判断基準が得られるため、復習の際に講義状況の記録すべてに同じような時間をかけて目を通す必要がなくなり、重点的に重要な部分を参照することが可能となる。講師側としては、自分の講義方法の流れがどのようなものであるかという特徴が分析できる。この特徴に対するアンケートなどを利用者側にすることで、利用者側にとって講義をよりよいものに出来るようになる。このように、講義中の重要な状況の推定ができれば、その推定をもとに映像コンテンツをより有効利用することができるようになる。

そこで、本稿では講義中の重要な状況の推定に関する情報を記録するための、講義構造の分析を目指す。また、提案手法を利用し、上述のように利用者側がより有効利用できるように、既存のアーカイブシステムとは異なる利用方法を提供するシステムのプロトタイプを作成した。

概要は次のようになる。まず、講義状況には教材の使い方や時間などといった講義を構成する要素の構造と関連があり、講義の要素としての教材に注目する。講義において利用される教材として板書や配布資料、パワーポイントスライドなどが挙げられるが、このような教材は講師によって説明される内容の一部あるいは、内容を要約したものであり、講義の構造分析に用いることは可能である。本手法においては、教材としてパワーポイントスライドを用いる講義を対象としている。これは、板書を用いる講義に比べ、パワーポイントスライドを用いる講義では、

- 関連するスライドを提示するなどの内容に合わせたスライドの切り替えが

容易であるため、講義における内容の推移がスライドの推移として表れる

- 電子的に記録することが容易である

といった利点があるためである。

講義の構造分析のためにパワーポイントスライドを用いる場合、

- スライドの切り替え
- スライドそのもの

の2点に注目することでいくつかの情報が得られる。

まず、スライドが切り替えについて注目する。スライドが切り替わった場合、内容も切り替わったと考えることが出来るが、類似する内容を説明している場合は大きな概念を説明しているという点からは内容的には切り替わっていないと考えることが出来る場合もある。そこで、スライド間の関係に注目し、グループ化を行う。これにより、あるグループに含まれるスライドは同じ概念を説明しているものであると判断するのである。これにより、各スライドがどのグループに属すかの情報が得られるため、実際の講義におけるスライドの提示順序に従ってスライドの移り変わりを調べることで、内容的な切れ目を知ることが出来る。グループ化の際に用いるスライド間の関係としてはスライド間の類似度を定義し、これを利用する。

このようなスライド間の関係を得るためには、スライドそのものに注目し、それぞれのスライドを特徴付ける必要がある。そこで、スライドに含まれるキーワードを用いた特徴付けを行う。また、キーワードに基づいた各スライドの得点を定義し、その得点分布を調べた。得点帯によるグループ分けを行い、各グループに含まれるスライドについての分析を行い、同様に講義の構造分析に利用する。

本稿は以下のような構成になっている。まず第2章においてスライド間の関係と講義構造について述べ、第3章において講義の構造分析手法について述べる。第4章では本手法を実際に適用し、本手法の妥当性について考察する。最後に第5章で結論を述べる。

手法の概要は次のようになる。まず、スライド間の関係を出すため各スライドのテキスト部分に注目する。そのなかから、出現頻度が高くかつ遍在する語句をキーワードととして抽出する。得られたキーワードの出現回数と全スライドにおけるキーワード数を利用し、各スライドに特徴ベクトルを与える。この特徴ベクトルからスライド間の距離を求め、一定の閾値以下のものは同じグル

ープとし、グループ化を行う。このようにグループ化を行っていった場合に、ある値で初めて同じグループに含まれるようになった時、これを許容距離と定義し、この許容距離を各スライド毎に調べていく。次に、各スライドにキーワードに基づいて得点を与え、これを重要度とする。得られた重要度と許容距離に注目することで、どのような講義構造であるのかを分析を行う。

第2章　対象とする講義およびその構造とスライドの関係

2.1　対象とする講義形態と従来研究との比較

大学における講義の形態としては、グループディスカッションや最近ではインターネット上での講義配信や遠隔講義など様々な形態のものがあるが、今回対象とする講義は複数の生徒が講義室に集まり、講師がそれに対面する形で行われる講義である。このような講義においては板書や配布資料やパワーポイントスライドなどの教材が利用されており、それに基づきながら順次説明が行われていく。最近ではアニメーション効果など視覚に訴え、印象に残る資料作りが可能であるという点からパワーポイントスライドを用いた講義が増えてきている。また、通常の板書や配布資料に比べ、パワーポイントスライドは電子的であり、講義状況を撮影するような研究においては同時にその講義において利用された電子的な資料を記録することが可能となっている。特に、パワーポイントスライドを用いる講義では、板書のみを用いる講義に比べると、以前説明したスライドを容易に提示することが出来るという利点がある。つまり、

講義における内容の推移がスライドの推移として表れる

という大きな特徴があるのである。

以上より、本研究ではパワーポイントスライドを用いる講義を対象とし、スライドに注目する。そこで、スライド内のテキスト部分から重要な語句を抽出することを考えた場合、従来より、インデックスの付加や新聞記事要約のためにテキストの分析を行う様々な研究が行われている。[2][3][4] このような研究においては、重要文抽出や重要語句の抽出の際に、構文解析を行ったり、重要度の付与として，主要語、高頻度の名詞、位置情報などの種々の情報を用いている。スライド内においてはこれらと同様な処理を行うことで重要語句を抽出するが、本研究の特徴として、その他にスライド間の関係に基づいたグループ化を行うことによって得られる内容の区切りの分布を利用することが挙げられる。また、従来の講義の自動撮影についての研究 [1] などでは、講義の内容についての状況の推定に関する情報を記録する研究は行われていないため、内容の区切りの分布に基づいて講義の構造分析を行い講義内容の重要区間の推定に関する情報を得ることを目指す。このような情報が得られることにより、従来の研究によって得られている映像コンテンツがより有効に利用できるようになる。

2.2　講義構造とスライドの提示順序の関係

講義の構造分析にあたり、今回対象としている講義は主にパワーポイントスライドを用いて進めていく講義であるため、スライドの提示方法に注目する。前節で述べたように、講義内容の移り変わりに合わせて、提示されるスライドが順次変わっていくため、講義の内容に区切りがあれば、スライド間にも内容の区切りがあると考えられる。

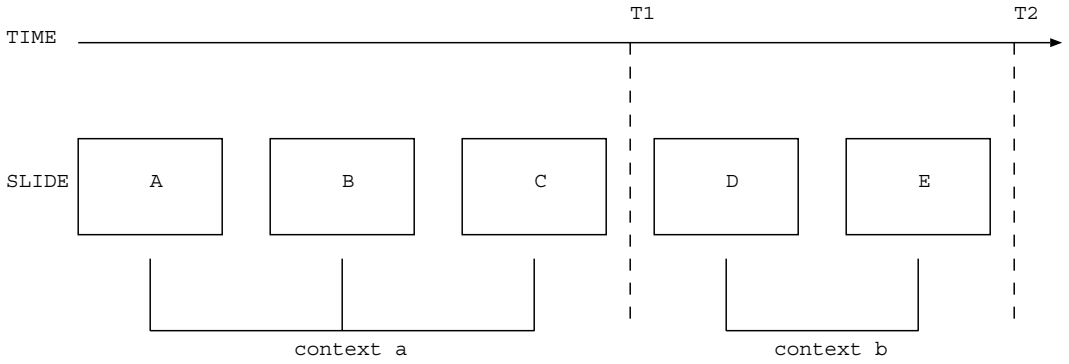

図 2.1: 講義の流れ

例として、図 2.1 のように、講義内容にあわせてスライドが A から E へ移り変わっていく場合を考える。このような場合に

- スライド A、B、C がある内容 a を説明している
- スライド D、E がある内容 b を説明している
- 内容 a と内容 b は異なる

といった情報が得られれば、時間 T1 において内容的な区切りが存在することが判断できると考えられる。そこで、これらの情報を得るために、スライド間にどのような関係があるのかを調べていく。

まず、現在注目しているスライドの前後に提示されたスライドとの関係について考える。講義の流れを考えた場合、ある内容が説明されている時点の時間的前後ではそれに類似した内容を説明している可能性が高い。しかしながら、前後に提示されたスライドについて考えると、スライドが切り替わる場合として、異なる内容の説明が始まった場合も考えられるためスライドの前後関係のみで判断する事は難しい。

次に、スライドの提示方法として、復習などで一度提示されたスライドを再

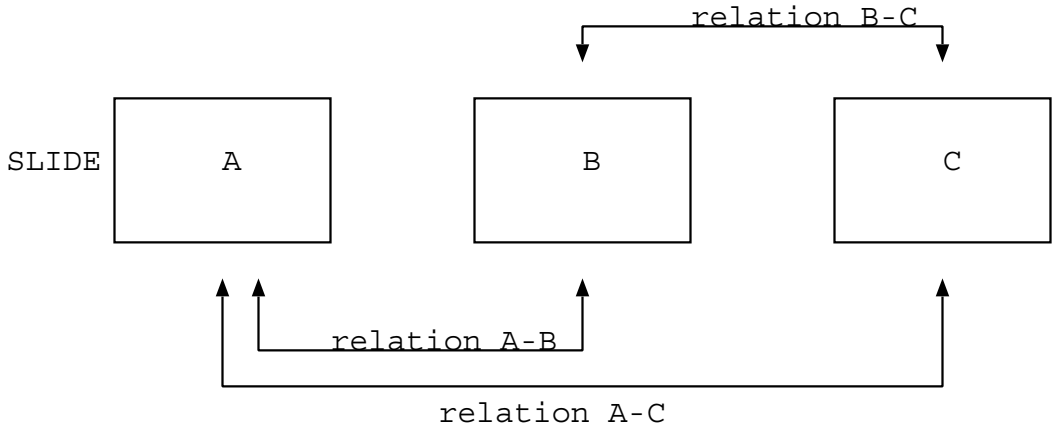

図 2.2: スライド間の関係

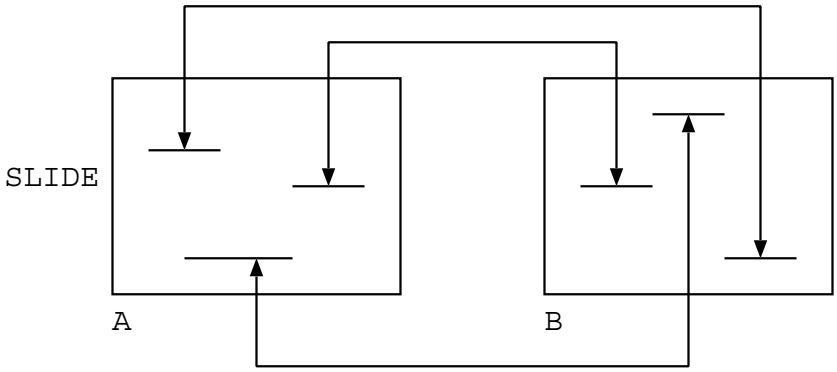

図 2.3: スライド情報による比較

び提示する場合について考える。この場合は何かしらの関係があるために再び提示しているため、内容的に類似している可能性が高いと判断できるが、そのような提示がいつ起こったのかの判断が難しい。

したがって、図 2.2 のように全てのスライド間との関係を調べていく。

実際にこのようなスライド間の関係を調べるためには図 2.3 のように、スライドの内容を用いて比較を行うことが必要になる。

このようなスライドから得られる情報を比較することで、それぞれのスライド間の関係性を示す値を求め、その値を閾値で区別することにより、内容的に類似したスライドであるのかどうかの判断をしていくのである。このようなスライド間の関係が得られれば、講義におけるスライドの提示順序にしたがって、スライドの関係を見ていくことで講義内容の推移に関する情報が得られるはずである。

ここで、スライド間の関係と講義内容の移り変わりについて考えると、ある講義において大きな概念を説明しているような部分では、内容的に類似しているスライドが連続し、順に細かな概念を説明している部分では、何枚か内容的に類似しているスライドが連続し、内容の切り替わりが生じ、その後また内容的に類似しているスライドが連続していくと考えられる。また、講義によっては大きな概念を説明しているような部分がなく、細かな概念を説明している部分のみで構成されている場合もあると考えられる。そこで、このような 2 種類の講義を講義構造の例として以下のように呼ぶ。

演繹的な講義　大きな概念をまず説明し、順に細かな概念のものを説明していく講義 (図 2.4)

帰納的な講義　細かな概念を説明し、その組合せで大きな概念を説明していく講義 (図 2.5)

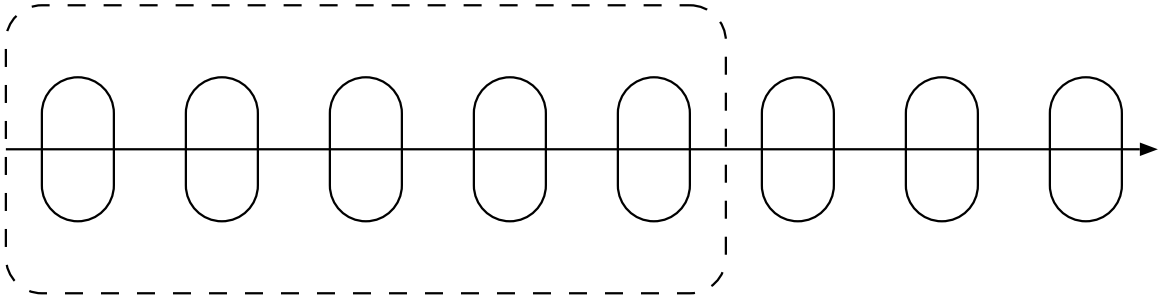

図 2.4: 演繹的

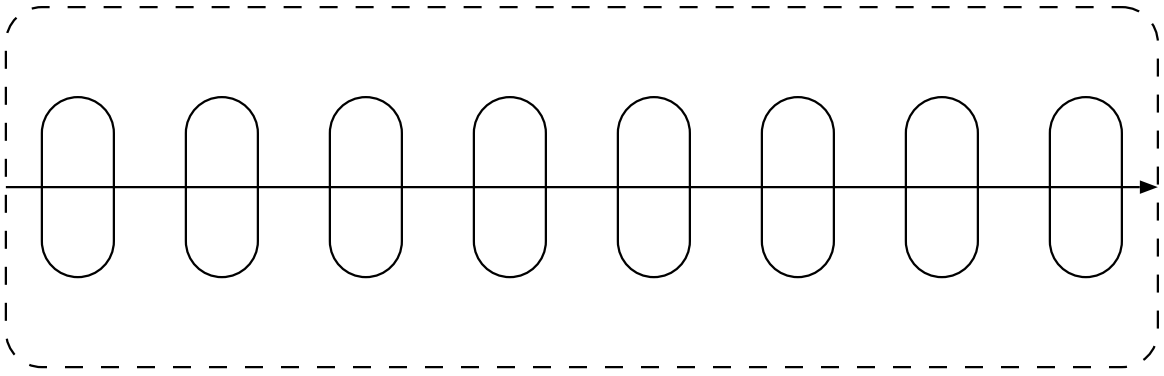

図 2.5: 帰納的

例として上記の 2 種類の講義をあげたが、講義内容の移り変わりに注目することで、対象となるの講義がどのような構造をもった講義であるのかを調べていく。

第 3 章　講義の構造分析手法

3.1　スライド間の類似性

講義の構造分析を行うためにまず、スライドのテキスト部分に注目し、各スライドの特徴ベクトルを求めていく。

講義の構造分析を行うためにスライド間の関係に注目するのであるが、まずそのスライドそのものの特徴を求めなければならない。一般に、パワーポイントスライドは講師が口頭で説明を行う際に補助的に用いられるものであり、口頭での説明が最も情報を含むものであるのに対して、スライドは口頭での説明の際に用いられる用語や概念をまとめたものであると考えることができる。これは、講師が内容を分析した上で、説明が必要な内容がタイトルや箇条書きの項目となるようした上で、生徒が理解しやすくなるようにスライドを作成する場合が多いことからも明らかである。しかしながら、内容一つ一つが細かく説明されているような講義においては、用いられるスライド数が多くなってしまうため、どの部分が重要なのか、どの部分が同じような概念について説明している部分なのかがわかりにくくなってしまう場合がある。このような場合、スライド間においていくつかの関係性が見つかれば、その関係性によりスライドをグループ化することにより、グループ毎にある概念を説明していると判断することが出来ると考えられる。そこで、同じような概念を説明している部分がどの部分であるかを導き出すために、スライド間の関係に考える。今回、このスライド間の関係として、スライド間の類似性を利用する。

スライド間の類似性を考える際にまず始めに考えられるのが

- スライド内の全文書 (含まれる字句すべて) を比較し、その類似性を調べる

という方法である。1 枚のスライドの特徴を含まれる字句すべてであらわすこの方法の場合、文書が全く同じであれば正確にそれを検出できる利点があるが、助詞や接続詞などの内容的に意味をもたないとみなせる部分までもが比較の対象となり、計算コストも高くなってしまう。今回の場合は、全く同じものであるかという点よりも、スライド同士がどの程度類似しているのかが重要なの

である。そこで、スライドからキーワードを抽出し、そのキーワードに基づいたスライドの特徴からスライド間の類似性を求めていく。ここでのキーワードとしては

- 名詞あるいは名詞がいくつか連接した語句

と定義する。このようなキーワードを定義した理由は、助詞などの内容についての情報としての意味をあまり持たないものは考えなくともそれほど影響を与えないと考えられるからである。また、数式等は内容的には意味を持つものであると考えられるが、その部分はより具体的で細かな説明をしていると考えられるためキーワードに含まないのである。スライドの文章からこのキーワードを抽出するために形態素解析をおこなう。

次に、スライド間の類似度を求める際に、比較の方法として、

- スライドに含まれるキーワードの種類で比較する（そのキーワードを含む、含まない）
- キーワードの種類に加え、出現頻度をも考慮して比較する

の 2 種類が考えられる。

キーワードの種類のみで比較する場合、その語句がその講義における普遍的な意味を持つ語句であり頻繁に出てくるようなものであれば多くのスライドにおいて、類似度が高くなってしまう（例として、パターン認識の講義における「認識」というキーワード）。そのような語句が存在すると、目的としているスライド間の関連性とは異なる関連のスライドまでもが同じ関連性というグループに含まれてしまうという問題が生じる。

そこで、そのキーワードの出現頻度も考慮に入れてスライド間の類似度を考えていく。出現頻度を加えることにより、キーワード毎に重みを与えるのである。これにより、先の問題は解決されるが、新たな問題として、実際には異なるスライドではあるがキーワードとして選ばれた語句毎の出現頻度が同じである場合に、類似度が高くなる可能性があるが、このような場合が生じるのは低確率であるため考慮しない。

以上のようにして、出現頻度による重み付けを行ったキーワードを用いてスライドを特徴付け、スライド間の比較を行うことにより、スライド間の類似性を求めていく。具体的なキーワード重み付けの方法（ TF*IDF 法）について、次節で説明する。

3.2　キーワードの重み付け（TF*IDF 法）

キーワードの重み付けの際には、どのような性質のキーワードがそのスライドを特徴付けるのかが重要である。どのようなキーワードがそのスライドの特徴として重要であるかは以下の 2 点から考えることが出来る。

point1　そのスライドで何度も繰り返して出現するキーワードはそのスライドで重要な概念を説明している

point2　そのスライド以外の多くのスライドにおいても何度も出現するようなキーワードはより一般的であり、そのスライドのみを特徴付ける性質を持つものではない。

この 2 つの観点から重み付けを行う。この際に用いるのが TF*IDF 法である。TF*IDF 法は TF（term frequency）と IDF（inverse document frequency）を利用したものである。TF が point1、IDF が point2 を考慮に入れたものである。

TF（term frequency）とは、ある文書（スライド）d におけるキーワード t の生起頻度であり、

tf(i,j):文章（スライド）D_i におけるキーワード j の出現頻度

と表す。

IDF（inverse document frequency）は文書数（スライド数）N と、キーワード t が一回以上生起する文書（スライド）の数 df(document frequency) によって次のように定義される。

$$idf(t) = log(\frac{N}{dft})$$

dft:キーワード t が一回以上生起する文書（スライド）数

この TF、IDF を用い、文書（スライド）D_i における語句 j の重みは

$$w_{ij} = tf_{ij} * idf(j)$$

として表す。このようにして、各スライドにおける、キーワード毎の重みを得ることが出来る。

3.3　スライド間の類似度

次に、前節によって得られたキーワード毎の重みを利用し、スライドの特徴ベクトルを表し、それを用いて具体的にスライド間の類似度を求めていく。

スライド D_i の特徴ベクトルを

$$D_i = (w_{i1}, w_{i2}, ..., w_{in})$$

n:文書集合におけるすべての異なるキーワード数

として表し、これを用い、以下のような式でスライド間の類似度を求める。

$$\text{類似度}\,(D_i, D_j) = \frac{w_{i1}w_{j1}+...+w_{in}w_{jn}}{\sqrt{w_{i1}^2+...+w_{in}^2}\sqrt{w_{j1}^2+...+w_{jn}^2}}\ ,\ (i \neq j)$$

このようにして得られたスライド間の類似度を、以下では利用していく。

3.4 スライドのグループ化

前節において得られたスライド間の類似度をどのように用いていくかについてここでは述べていく。講義の構造を分析するにあたり、内容の移り変わりに注目するのであるが、その方法として、

- 切り替えの前後のスライドに注目し、そのスライド間の類似度を調べる

という方法が考えられる。その場合は前後のスライドが異なるあるいは類似度が低いといった情報しか得られない。ここで求めたいのは、スライドがある内容を説明するものから違う内容を説明するものに切り替わった場合である。そのためにはスライドごとにグループ化を行っておき、どのグループからどのグループへの変化であるかという情報を得る必要がある。このスライドのグループ化を、前節で示した類似度に基づいて行っていくのである。

次のような方法によりグループ化を行う。あらかじめ全てのスライド間の類似度を計算しておき、スライドを提示された順番に見ていく。

1. 一枚目のスライドに対してグループ番号を与え、それ以降に提示されたスライドとの間の類似度が一定の閾値以上のものである場合に同じグループ番号を与える。同時にその類似度も保持しておく。
2. ２枚目以降のスライドに対しても同様に、それ以降に提示されたスライドとの間の類似度について調べていく。まず、現在注目しているスライドがどのグループにも含まれていなければ新たなグループ番号を与える。次に、比較対象となるスライドがすでにグループに属しているかどうかにより、処理を変える。
   (a) どのグループにも属していない場合は、一定の閾値以上の類似度であれば同じグループ番号を与える。同時にその類似度も保持しておく。

(b)　既にグループに属している場合は、得られた類似度と、保持されている類似度の値を比較しより近いと判断されればグループ番号を書き換え、そうでなければ何もしない。

この手順の例を図 3.1 に示す。

類似度が 0.5 以上のものを同じグループに含む場合、スライド C は一度グループ 1 に含まれるが、スライド B との類似度がスライド A との類似度より高いため、最終的にはグループ 2 に含まれるのである。

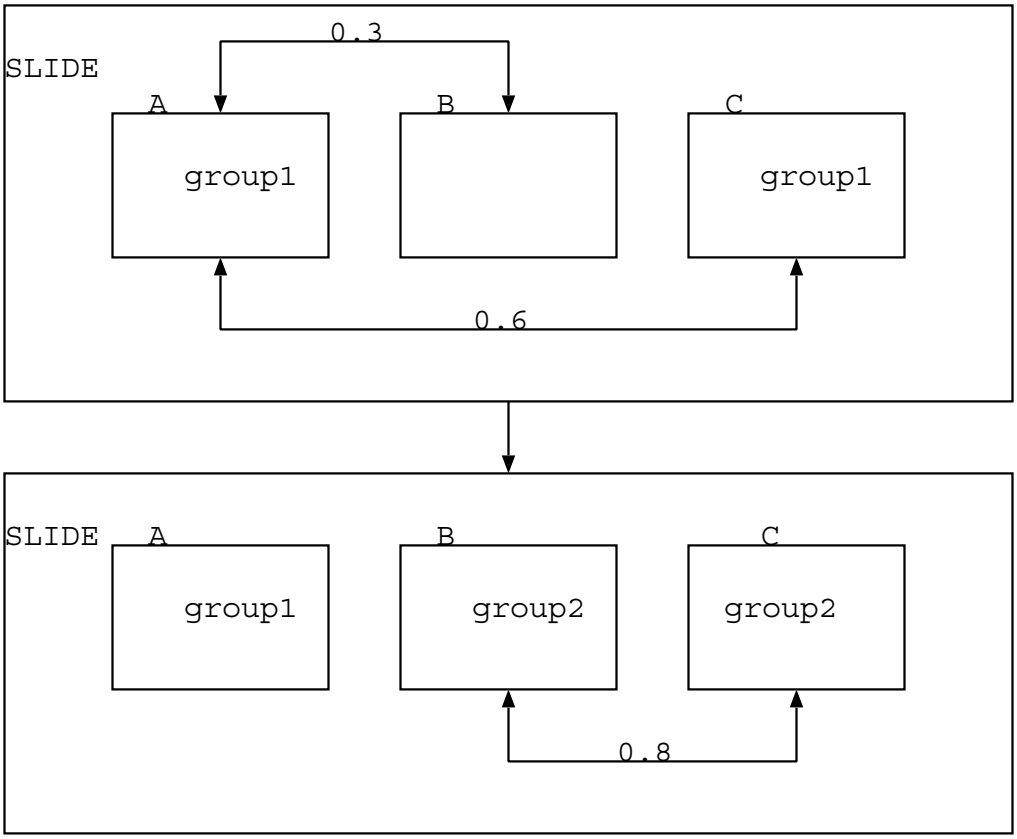

図 3.1: スライドのグループ化

このようにしてグループ化を行っていくのであるが、ここで同じグループに含まれるスライドは、ある内容を説明しているスライド群であると考えられる。そこで、このようなスライドをあるコンテキストを表すスライドと呼ぶ。また、スライドの切り替えが行われたとき、切り替え前のスライドと切り替え後のスライドにおいて、表すコンテキストが異なる場合、これをコンテキストカットと呼ぶ。つまり、コンテキストカットが表すのは内容が切り替わった点と

いえる。コンテキストとコンテキストカットの関係は図 3.2 のようになる。

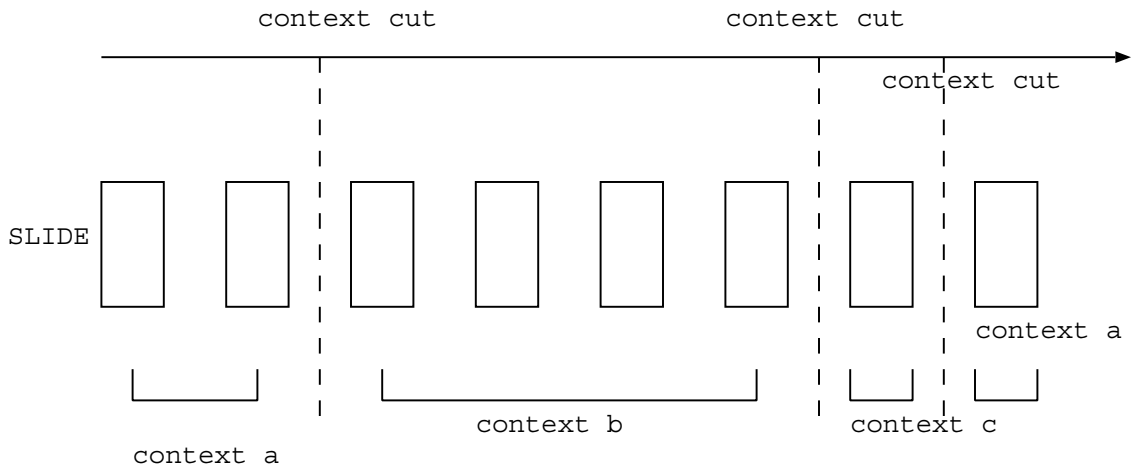

図 3.2: コンテキストとコンテキストカットの関係

このコンテキストカットに注目し、講義の流れを見ていくと、それぞれの講義がコンテキストカットの分布で特徴付けることができる。図 3.3 を参考にコンテキストカットの出現頻度の持つ意味を考えていく。コンテキストカットの

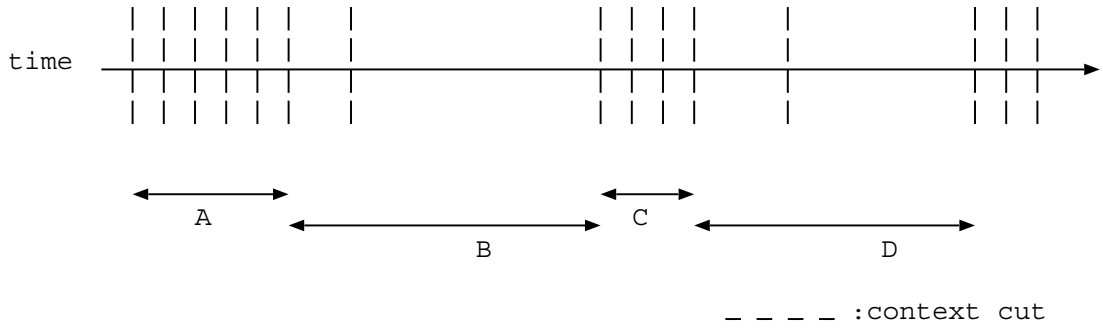

図 3.3: コンテキストカットの分布例

出現頻度の高い A や C の部分では、出現頻度の低い B や D の部分に比べ、次々と異なる内容について説明を行っていると考えられ、逆に、コンテキストカットの出現頻度の低い B や D の部分は、出現頻度の高い A や C の部分に比べ、

ある内容について説明するスライド数が多いと考えることができる。次々に異なる内容を説明している部分では、細かな概念を順次説明していると考え、ある内容について説明するスライド数が多い部分では、大きな概念を説明している、またはその部分が重要であることを示していると考えることにより講義の分析に利用する。

このような分析を多数の講義に対して行うことによって、2章2節で説明したように、講義構造からどの部分が大きな概念を表しているかが判断できるような構造の分類が出来るようなると考えられる。

ただし、ここで注意しなければならないのがグループ化を行うかどうかの判断基準である閾値の値である。適切な閾値が選択出来なければ、本来異なるグループに含まれるべきスライドが同じグループに含まれることなどが考えられ、構造の分析の結果にまで影響を与えることになる。しかしながら、スライドが類似しているかどうかは主観的な判断によるものであるため、アンケートなどを行うことによって妥当であると思われる値は決定出来たとしても、それが正しい値であるという保証はない。

以上のように、スライド間の類似性に基づいて講義の構造についての分析が可能となると考えられるが、ここまではスライド間の評価を用いたのみであり、スライドそのものの評価を利用していない。そこでスライドそのものの評価を導入し、講義の構造分析に利用していく方法を次節で述べる。

3.5　スライドの得点を用いた選択

前節までに、スライド間の類似性に基づいた評価は利用しているが、スライドそのものについての評価を利用していないため、ここではスライドそのものの評価として、スライド毎の得点を定義する。

スライドの得点　あるスライドにおいて出現するキーワードの全スライドにおける出現回数を、キーワード毎に加えたもの

このような得点付けを行った場合、

- 得点が高くなると考えられるスライド
  頻繁に出現するキーワードつまり、複数のスライドで何度も説明されるような重要あるいは大きな概念をあらわすキーワードを多く含むスライド
- 得点が低くなると考えられるスライド
  そのスライドのみで出現するキーワードつまり、そのスライドを特徴付け

てはいるが、他のスライドと関連性の無いキーワードを含むスライドとなると予想される。また、多くのスライドにおいて出現するような、一般的ではあるが大きな概念をあらわしてはいないキーワードを含むスライドでは得点が高くなってしまうことが考えられるが、多くのスライドにおいて出現するならば全体的に得点があがるだけなのでこれは問題とならない。

以上により、得点が高いスライドを含むグループほどより重要あるいはより大きな概念を説明しており、得点が低いスライドを含むグループほどより細かな概念を説明していることが推定できる。この推定が正しければ、前節までに得られたスライド間の関係としてのコンテキストカットの出現頻度とスライドの得点を用いることにより、概念毎のグループ分けとその重要度が得られることになる。これにより、スライドそのものとスライド間の評価に基づいた、講義の構造分析が可能となる。

第4章　実験と考察

4.1　実験

本研究では、スライドを用いた講義を対象としているため、実際にそのような形態の複数の講義に対して実験を行った。講義は一回あたり一時間半で行われる。用いるスライドのデータはあらかじめ講義状況の撮影のための講義アーカイブ実験により得られているため、そのデータを用いた。パワーポイントのデータから不必要な部分をあらかじめ除いておき、キーワードを抽出するために形態素解析システムとして茶筌 [5] を利用した。

キーワード抽出の例として、以下のような内容のスライドからキーワードを抽出することを考える

識別能力と頑健性
識別能力, 与えられた学習パターンを高精度に分類するか, 頑健性（汎化能力）, 学習サンプル以外のパターンに対応できるか, 過学習, 識別能力を高めすぎると頑健性を損なう, 特に、訓練時と異なった環境, (例)　スポーツ（球場)、試験、文字（活字)、音声（方言）, できるだけ多くのデータで学習しておく, システムを必要以上に複雑にしない（単純に）

上記の内容のスライドを用いた場合、このスライドにおけるキーワードとそれに関連した情報が表 4.1 のような形で得られる。

キーワード	キーワードの重み	そのスライドでの出現回数	全スライドでの出現回数
活字	0.731857	1	1
データ	0.331298	1	7
試験	0.589175	1	2
頑健性	1.46371	3	3
複雑	0.0534196	1	27
精度	0.731857	1	1
訓練時	0.731857	1	1
システム	0.0534196	1	27
学習サンプル以外	0.731857	1	1
多く	0.589175	1	2
音声	0.731857	1	1
パターン	0.446493	1	5
高	0.731857	1	1
識別能力	1.46371	3	3
文字	0.589175	1	2
球場	0.731857	1	1
例	0.363029	1	8
必要以上	0.0534196	1	27
対応	0.731857	1	1
環境	0.731857	1	1
分類	0.400559	1	5
方言	0.731857	1	1
汎化能力	0.505711	1	3
スポーツ	0.589175	1	2
過	0.731857	1	1
単純	0.0534196	1	33

学習パターン	0.505711	1	3
学習	0.0728027	2	33

表 4.1: キーワードについての情報

次に、得られたキーワードを利用してスライド間の類似性を求めていく。例として次のスライド A に対して類似度が高いスライドと低いスライドは以下のようなものである。

- スライド A
  誤り訂正学習法　（4.3 節）
  現在の識別関数の出力が正しくなければ、重みを修正,W ’=W+cY (クラス 1 の Y に対して正でないとき）,W ’=W-cY (クラス 2 の Y に対して負でないとき）,c：修正増分（正）, ベクトル Y は超平面 Y・W=0 に直交するので、超平面の反対方向への移動に対応, 出力が正しいときは修正しない, 全パターン集合に対して正しい出力となるまで繰返し…図 4.2 の例

- 類似度 0.321355 のスライド
  パーセプトロン収束定理
  パターン集合が線形分離可能なとき,c>0, or 0< λ <2 の修正方法はいずれも, 有限回の学習で解に到達する, 図 4.2... パターン 1,2,3,4.. の順に提示して c=1 で 5 回で収束,(cf)　数値解析手法 Y は超平面 Y・W=0 に直交するので、超平面の反対方向への移動に対応, 出力が正しいときは修正しない, 全パターン集合に対して正しい出力となるまで繰返し…図 4.2 の例

- 類似度 0.831972 のスライド
  誤り訂正の方法
  固定増分　c　(c>0),c が小さすぎると小刻みすぎる,c が大きすぎると振動する, 絶対修正,W ’Y=(W+cY)Y>0 → c>—WY—/YY, 部分修正,cY= λ —WY—/Y 順に提示して c=1 で 5 回で収束,(cf)　数値解析手法 Y は超平面 Y・W=0 に直交するので、超平面の反対方向への移動に対応, 出力が正しいときは修正しない, 全パターン集合に対して正しい出力となるまで繰返し…図 4.2 の例

次に、スライドそのものの得点を求め、その得点分布について調べる。ここでは

- 講義一回のみの場合の分布
- 同一講義で日付が異なる講義七回分についての分布

の 2 種類の分布について調べたデータをそれぞれ図 4.1 、図 4.2 に記載する。

次に、構造分析に用いる情報の例として、ある一回の講義に対して、スライド間の類似度の閾値を 0.3 としてグループ化した結果得られた、コンテキストの推移、コンテキストカット、スライドの得点とその順位について、表 4.2 に示す。

スライド番号	グループ	スライドの得点	スライドの得点順位	コンテキストカットの有無
slide1	class-1	15	22	無
slide2	class-2	81	21	有
slide3	class-3	203	20	有
slide4	class-4	257	17	有
slide5	class-5	219	19	有
slide6	class-6	238	18	有
slide7	class-7	300	16	有
slide8	class-8	358	14	有
slide9	class-9	0	23	有
slide10	class-10	419	12	有
slide11	class-11	411	13	有
slide12	class-10	545	6	有
slide13	class-12	449	8	有
slide14	class-10	615	1	有
slide15	class-12	449	8	有
slide16	class-10	615	1	有
slide17	class-12	449	8	有

slide18	class-13	425	11	有
slide19	class-4	442	9	有
slide20	class-14	542	7	有
slide21	class-15	610	2	有
slide22	class-14	581	5	有
slide23	class-15	610	2	有
slide24	class-15	582	4	無
slide25	class-15	610	2	無
slide26	class-15	582	4	無
slide27	class-15	610	2	無
slide28	class-15	582	4	無
slide29	class-16	0	23	有
slide30	class-17	609	3	有
slide31	class-18	0	23	有
slide32	class-17	609	3	有
slide33	class-19	328	15	有
slide34	class-17	428	10	有
slide35	class-20	0	23	有

表 4.2: スライドについての情報

4.2 考察

元スライドと表 4.1 を見る限り、キーワードそのものは正しく切り出すことができている。ここで注目べき点はキーワードの重みが妥当であるかという点である。ここで得られたキーワードの重みを用いて、スライド間の類似度を計算しているためである。表のデータでは想定どうりにそのスライドでの出現回数が高く、全スライドの中でそのキーワードの出現するスライド数が低いものほど高くなっている。

次に、スライド間の類似度の妥当性を検討する。実験データとして記載した類似度が約 0.8 のスライドと約 0.3 スライドを元のスライドと比較すると、類似度が約 0.8 のスライドは元のスライドに類似していると判断できる。類似度が

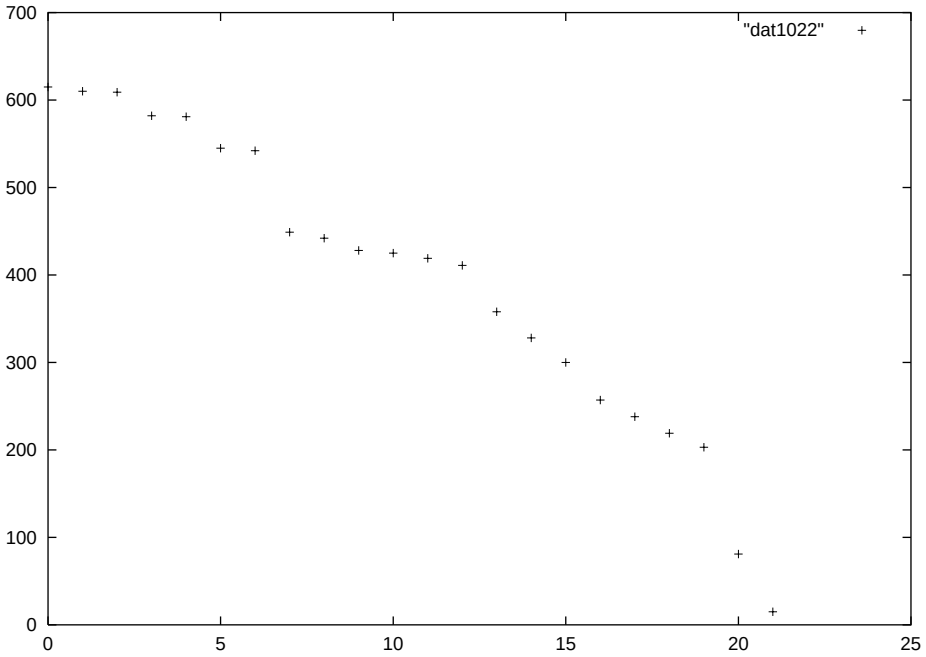

図 4.1: 講義一回における得点分布

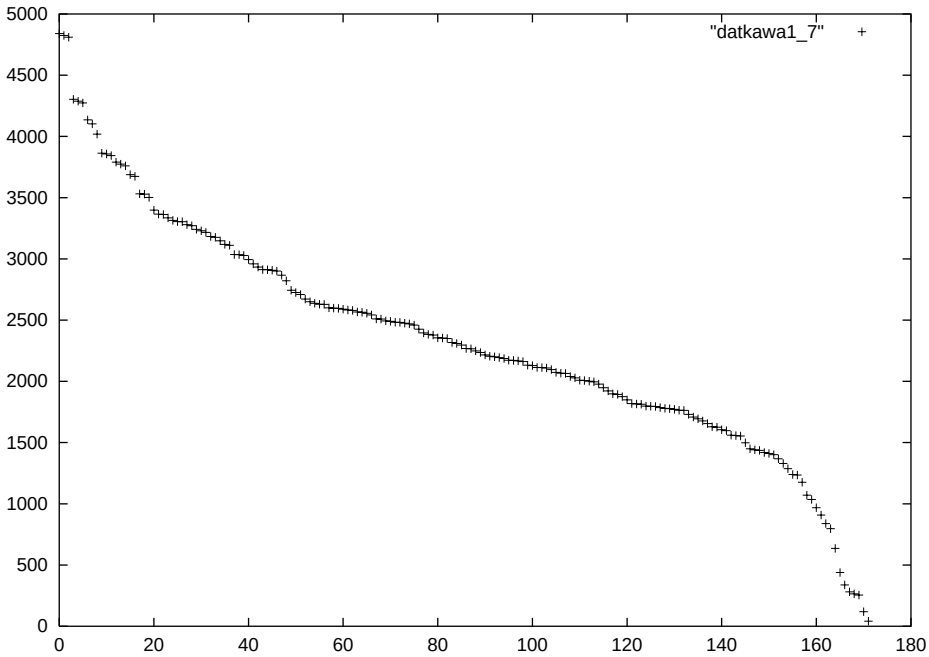

図 4.2: 講義七回分における得点分布

約 0.3 のスライドについても、一部で同じ文章が使われているところを考慮にいれると、類似しているスライドとしてみなして良い。スライド間の類似度を用いてグループ化を行っているため、適切な閾値でグループ化を行えば、得られたグループは妥当である。スライドに内容的な偏りがなければ、この閾値を高くすればより内容的に類似したスライドのグループが得られるが、グループ数が多くなり、逆に、低すぎる場合は内容的に類似していないと見なせるスライドをも含むグループとなるが、グループ数は少なくなる。ここでの目的はスライド間の関係から構造を分析することであるため、適切に閾値を選ばなければならない。

ここで、第 3 章 4 節で述べたように、正しい閾値を選ぶことは困難であるため、類似度が約 0.3 のスライドは類似スライドとして見なせるとし、ある講義において閾値を 0.3 としてグループ化を行った。その結果が表 4.2 である。この表を元にこの講義の構造を分析していく。表のグループの項目に注目すと、スライド 35 枚に対してグループ数は 20 となっている。ただし、スライド 9 、29 、31 、35 は空スライドであったため、場外して考える。すると、実質 31 回の切り変えに対してグループ数は 16 となる。また、コンテキストカットは 24 回出現しており、その全てがスライド 23 からスライド 28 の間以外のところで出現している。また、グループと得点から同一スライドと思われるスライドが複数出てきているが、同一スライドは主に近いスライド番号で出現している。また、スライド 19 ではグループ番号が 4 であり、かなり以前のスライドで説明された内容について再び説明したことが分る。このような特徴から講義の構造としては、序盤では次々と異なる細かな概念について説明を行っていき、中盤では同一スライドを交互に説明し、序盤に説明した部分について一度説明を加えたあと、スライド 24 から 28 の間で少し大きな概念あるいは重要な部分を説明したものと推定できる。

次に、スライドの得点の妥当性について検討していく。表 4.2 にはスライドの得点と得点順位の項目がある。スライドの得点が高ければ順位が高くなる。ここで、スライドの得点を用いずに構造分析を行った結果、大きな概念あるいは重要な部分を説明していると判定されたスライド 24 から 28 について注目する。ただし、スライド 24 と 26 と 28 、23 と 25 と 27 は同じスライドである。これらのスライドの得点順位は 2 位と 4 位である。得点順位が最下位であるものが 23 位であることから、この講義では実際は 23 枚のスライドを用いており、様々な

提示をした結果、35 回の切り替えを行っている。重要だと判断された 24 から 28 のスライドの得点順位が全 23 種類のスライドにおいて、2 位と 4 位であることから、ほぼ妥当な結果と言える。しかしながら、重要と判断されたスライドよりも順位が高いスライドが存在する。1 位のスライドであるスライド 14、16（この二つは同じスライド）と、3 位のスライドであるスライド 30、32（この二つは同じスライド）について考察を行う。3 位のスライドについては空のスライドである 29、31 を除いた場合、重要と判断されたスライドの次に説明されている内容である。それに対して 1 位のスライドについては、スライドの得点を用いずに構造分析を行った際に、交互に同一スライドを説明していると判断された部分に含まれる。しかしながら、これ以上に得点の妥当性を決定付けるようなものは得られなかった。

このようなスライドの得点による評価が高いスライドについても重要であると判断し、コンテキストカットの出現頻度に基づいた構造分析を補完し、重要である部分の候補として推定することが出来た。

ここで取り上げた講義とは別の日付の同じ講義について同じ様に実験を行った場合にも、同じように分析することができた。

第 5 章　おわりに

本稿では、講義中の重要な状況の推定のために、教材スライドに注目し、スライド間の類似性に基づいた講義構造の分析する手法を提案した。講義内容の推移が講義スライドの推移に表れることを利用し、講義スライドに注目し、内容の移り変わりをスライド間の関係性から求めた。スライドから形態素解析によりキーワードを抽出し、そのキーワードを元に、スライド間の類似度を得る。この類似度を適切な閾値で区切ることで類似したスライドをグループ化し、スライドの切り替えが行われる際にそのスライドの含まれるグループに注目することで、内容の切れ目であるコンテキストカットを得る。講義におけるコンテキストカットの出現頻度から講義の構造を分析する。また、スライドそのものの得点を導入することで、スライド間の類似性に基づいた講義の構造分析を補完する。

本手法の有効性を確認するため、実際にスライドを主に利用する講義に対して構造の分析を行った。コンテキストカットから得られた分析結果とスライド

の得点から得られた分析結果から、重要な部分の候補となるスライドを推定することができた。

今後の課題として次のものが考えられる。キーワードとして、現在は名詞や名詞が連接したものを選んでいるが、カタカナ語や新しい語は未知語として出る場合が多いので、このような語もキーワードに加えることでより正確なスライド間の関係が得られるようになり、さらに様々な講義に対応することが可能となると考えられる。

また、講義における内容の推移がスライドの推移として表れると考えているため、以前提示したスライドを遡って提示する際に通過するだけで説明の行われないスライドも有効な切り替えとしているが、提示時間が短いものを除外することで、行き先のスライドのみ有効にすることができ、より正確なコンテキストカットの出現頻度を利用した構造分析が可能となると考えられる。

さらに、同一講義で日付の異なるものにおいても同様の構造が出れば、それは講師ごとの特徴と見なすことができるため、同一講師で別内容の講義についての検討の余地が残る。

また、追加実験として、図 4.1 や図 4.2 で示した得点の分布を調べ、得点が高いスライド、中程度のスライド、低いスライドのグループに分け、それぞれのグループについて代表的なスライドについて比較実験を行うことにより得点の高低によって重要度が異なるのか調べることで得点の妥当性についての別の側面からの評価を行う必要がある。スライドの提示時間を得点に加えた場合についても同じような方法で評価を行う。

最後に、構造分析を行って得られた結果が妥当であるか（重要な部分あるいは大きな概念を説明していると判断された部分が妥当であるか）アンケートにより客観的に調べる必要がある。

謝辞

本研究を進めるにあたり、ご指導を賜りました美濃導彦教授、角所考助教授、日頃より熱心な御指導と本報告書の作成において多くの御助言を頂きました西口敏司助手、八木啓介助手に心より感謝いたします。最後に研究あるいは、本報告書作成に関して、多くの御助言を下さいました美濃研究室の皆様方に深く感謝致します。

参考文献

[1] 遠隔講義における多様なセンサを用いた話者状況の映像化東 和秀京都大学大学院情報学研究科修士論文 2001

[2] Salton, Gerald (1970) ”Automatic text analysis” Science, Vol.168, p.335-343

[3] 重要文抽出，自由作成要約に対応した新聞記事要約システム YELLOW 大竹 清敬, 岡本 大吾，児玉 充，増山 繁情報処理学会論文誌「データベース」Vol.43, No.SIG 2(TOD 13), pp.37-47 (2002) .

[4] 文章内構造を複合的に利用した論説文要約システム GREEN 山本 和英, 増山 繁, 内藤 昭三自然言語処理, Vol. 2, No.1, pp. 39-55 (1995)

[5] 松本祐治、北内啓、山下達雄、平野善隆、松田寛、浅原正幸：日本語形態素解析システム 茶筅、奈良先端大学院大学テクニカルレポート、NAIST-IS-TR99012 (1999)