情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-DD-76 No.2
2010/7/22

プレゼンテーションスライド情報検索のためのスライドページからの要求関連情報抽出

羽 山 徹 彩[†1] 國 藤 進[†1]

膨大となりつつあるスライドデータの情報アクセス性を高める方法のひとつとして，検索結果の一覧から有用な内容を含むスライドを把握し易くする提示技術が挙げられる．しかしながら，既存の単純な文字列やページ画像に置き換える方法では，図表の情報の欠如や短い語句の理解が困難であったり，視認性を得た画像サイズの確保による一覧性の低下したりすることが問題となる．そこで本研究ではスライド情報探索の効率性を高めるために，スライドページから関連する情報を適切に抽出する手法の開発を行った．提案手法では表示領域を指定することで，その領域に応じたページ画像とそのページ内の要求関連情報を抽出し，レイアウトが持つ関係を保持した提示を行う．評価では提案手法が従来の単純なテキスト提示方法と比較し，小領域の表示であってもスライドの内容をより正確に把握できることを確認した．

Relevant Piece of Information Extraction from Presentation Slide Page for Slide Information Retrieval

Tessai Hayama[†1] and Susumu Kunifuji[†1]

One of useful approaches for better access of increasing slide-data is to provide presentation technique, which supports easily understanding of each slide-page content among search results list. Previous techniques for slide-data processing have converted slide-page data into simple character string or page-image and presented them as search results list-items. However, it is difficult of their methods to understand figure/table information and short phrases, and to view the list of the items to get the image size with legibility. In this paper, we describe a method which extracts relevant piece of information from slide-page information for slide information searching. The proposed method extracts a page-image and relevant piece of information from each slide-data by indicating the presentation region, and then presents them with layout. Our experiment showed that our approach is useful for easily understanding of slide-page information in small region by comparison of simple text presentation method.

1. はじめに

近年の電子化プレゼンテーションの普及により，講義や会議などの多くの場面で電子的なプレゼンテーション資料（スライド）が利用されるようになった．利用されたスライドは遠隔講義資料や Web コンテンツとして逐次的に蓄積され，膨大かつ重要な知識資源となりつつある．そのため，スライドに含まれる情報の利活用性を高める情報アクセス技術が必要となる．

スライドに含まれる情報のアクセス性を高める有用な方法のひとつとして，スライドを検索した結果に対し，ユーザが要求する情報の関連箇所だけを切り出し，提示することが挙げられる．そのためには，スライドに含まれる情報を適切に分類し，構成的に扱うことができるデータ管理と情報提示が必要となる．このように検索結果から要求に関連しない情報を排除できることで，要求に関連する情報を容易に気付き易くなったり，検索結果の一覧から有用なスライドを効率よく取捨選択できたりする．

これまでスライドを扱った検索技術はデータベース分野や教育工学分野において，検索方式やその検索結果の提示方式を中心に取り組まれてきた．Min[5] は論文データベースから検索結果を分かりやすく提示する方式として，論文の節とスライドページの対応付け手法を開発し，論文の検索結果とともにその検索キーワードに対応するスライドページ画像を提示する方法を開発している．Guo ら[2] はスライド検索として，検索キーワードとともにその結果のスライドページ画像を用いた段階的な検索方式を開発しており，その検索結果としてスライドページ画像を提示する方法を採用している．横田ら[7] は教育コンテンツの再利用性を高める研究プロジェクトにおいて，教育コンテンツの自動作成のために，講義映像とその映像で利用されているスライドページ画像を画像検索によって対応付けることを行っている．一方，現状の商用検索システムの多くはスライドデータが扱っているように，検索子と照合したテキストとその周辺テキストを提示することを行っている．以上のように，既存のスライド検索技術ではスライドデータに対し，単純な文字列やページ画像に置き換えて扱ってきた．しかしながら，このような方法では，図表の情報の欠如や不完全な文の理解が困難であったり，視認性を得た画像サイズの確保により一覧性が低下したりすることが問題となる．

そこで本研究ではスライド情報探索の効率性を高めるために，スライドページから関連す

†1 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology

る情報を適切に抽出する手法の開発を行った．提案手法では表示領域を指定することで，その領域に応じて検索要求に関連する情報を抽出し，レイアウトによる関係を保持した提示を行う．

2. アプローチ

本研究では一般的な検索方式であるキーワード探索システムを対象として行う．つまり，キーワードを入力とし，そのキーワードと照合されたスライドを結果として提示する情報検索システムにおいて，スライドページに含まれる情報から検索要求に関連する情報だけを適切に抽出する手法を開発する．

本節では，まず本研究で扱うことのできるスライドに含まれる情報において述べ，次にその情報から検索要求に関連する情報を抽出のための設計指針について述べる．

2.1 スライドに含まれる情報

スライドに含まれる情報にはテキスト，写真，線，及び基本的な図形などのプリミティブなオブジェクトから構成されている．このようなプリミティブなオブジェクトは，タイトル，本文，図，表及び装飾といったスライド内容を理解しやすくする基本表現とする纏まり成している．各スライドには，発表の流れに沿ったそのスライドの内容を表現しているタイトルが付与され，そのスライド内容を説明するための項目や補助資料として，本文，図及び表などの基本表現が利用されている．また，それ以外のスライドに含まれているオブジェクトとしては，特定の内容を協調する記号や関係線，あるいは発表日付などのスライド内容と直接関係のない装飾表現がある．このように，スライドに含まれる情報には内容に関係するタイトル，本文，図及び表の 4 種類の属性と，内容に直接関係しない装飾属性のいずれかに分類することができる．そして，各スライドページには，ページタイトルの内容を説明をするために，それら内容に関係するオブジェクトの纏まりが構成的に表現されている[3]．

2.2 検索要求に関連する情報抽出の設計方針

情報検索システムにおいて結果の一覧性を確保するためには，検索要求に関連するスライドに含まれるオブジェクトを任意の小領域に提示できる必要がある．スライドページに含まれる情報は一般的に大画面を想定したレイアウトや表現を用いて，オブジェクトを配置させており，そのままのレイアウトを小領域に適用することが難しい．また，スライドに含まれる図表に対しては小領域に合わせて縮小を行うと，それに対して視認性が損なわれてしまう．そのため，検索要求に関連するオブジェクトを抽出した際には，小領域に適合するための再構成と提示方法が必要となる．

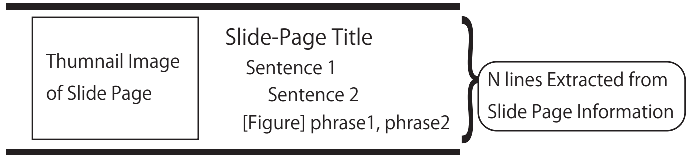


図 1 スライドページの情報提示インタフェース
Fig. 1 Interface for Slide-Page Presentation Information

また，検索結果一覧における 1 つの結果は情報抽出によりいくつかの情報が排除されていたとしても，スライドの内容をより正確に把握できるような提示である必要がある．これまでの検索結果の提示方法ではスライドページから，検索子のキーワードと照合するテキストとその周辺テキストだけを抽出してきた．スライド中の本文や図表に含まれている語句は比較的短く，不完全な文である場合が多い．そのため，それらにキーワード照合したとしても，意味のなり語句の並びが提示されるため，そのままでは理解することが困難となる．そのため，検索子と照合したオブジェクトの所在やそのオブジェクトの理解に役立つようなオブジェクトや手掛かりも提示するための抽出方法が必要となる．

3. 提 案 手 法

本研究ではスライドページに含まれる情報から，任意の表示領域に応じて検索要求に関連する情報を抽出し，提示する手法を提案する．そのための提示インタフェースと情報抽出処理について，それぞれ 3.1 節と 3.2 節で述べる．また提案手法の適用による想定する効果について，3.3 節で述べる．

3.1 スライドページの情報提示インタフェース

本研究で提案するスライドページに含まれる情報を任意の表示領域に提示するためのインタフェースを，図.1 に示す．

本インタフェースの表示領域には，画像表示領域とテキスト表示領域が含まれる．画像表示領域ではスライドページの縮小画像が表示され，その縮小率はテキスト表示領域に表示されるテキストの行数によって決められる．テキスト表示領域ではスライドに含まれる情報が表示され，1 行目にページタイトルが，それ以下の行にその他のページ内の情報が提示される．その各行には 1 つの属性を持つ情報（本文，図，表）が 1 つ割り当てられ，他の行との関係を視覚的に把握し易くするために，字下げを使って表示されている．1 行辺りの文字数が多い場合には，制限文字数以内で領域内に収まるように部分的に抜粋される．また図表

の属性を持つ情報を割り当てる場合には，その領域内で視認性が高い大きさでの表示することが難しいため，“[Figure]”や “[Table]”の属性を先頭に付与し，それに含まれる文字列を並べた表示を行う．

以上のような提示インタフェースを実現するために，スライドに含まれる各情報に対し属性と他の情報との関係を定義し，任意の領域内で収まるための情報抽出を行う必要がある．前者には文献[3] のスライド情報の構造抽出手法を利用した定義付けを行い，後者には次節で述べる提示情報抽出を用いる．

3.2 提示情報抽出

スライドページに含まれる情報をすべて提示すると，任意の小領域に収まりきれない場合がある．本抽出手法では，スライドに含まれる情報のなかで指定した数だけ抽出し，それら関連性を保持して構成的な提示を行う．

情報抽出の処理手順を以下に示す．

0. 前処理として，スライドページに含まれる情報に対し構造抽出を行う．この構造抽出処理では文献[3] の構造抽出手法に用いて，タイトルをルートとし，その他の情報をノードとして関連付けた木構造を生成する．ここでの木構造はルートが最上位となる．
1. 情報の抽出数として N の値を設定する．
2. 検索子が含まれている構造の中で最上位の階層位置を検出する．
3. その同じ階層に検索子を含む情報が複数検出された場合には，右優先でそれら情報を N 以下の数で抽出する．もし，抽出された数が N を満たした場合，処理を終了する．
4. 検出された情報の 1 つ下の階層位置にある情報に着目する．
   4.1 その情報の中に検索子を含む情報が検出された場合，右優先でその情報を N 以下の数で抽出する．もし，抽出された数が N を満たした場合，処理を終了する．
   4.2 その情報の中に検索子が含む情報が検出されない場合，右優先でその情報を N 以下の数で抽出する．もし，抽出された数が N を満たした場合，処理を終了する．
5. 手順 4 を，抽出された数が N を満たすか，対象となる情報がなくなるまで繰り返す．

以上の処理で抽出された情報は手順 2 で初めに抽出された情報をもとに，新たな木構造として表現することができる．

本提示情報抽出の適用例を図.2 に示す．スライドページに含まれる情報に対し，構造抽出手法を適用すると Title をルートとした木構造へ展開される．それに対し，本提示情報抽出ではまず検索子が含まれている情報の検出として，“Title”が抽出される．次に，その“Title”の下の階層にある情報として，“Figure1”が抽出される．“Figure1”と同じ階層に他

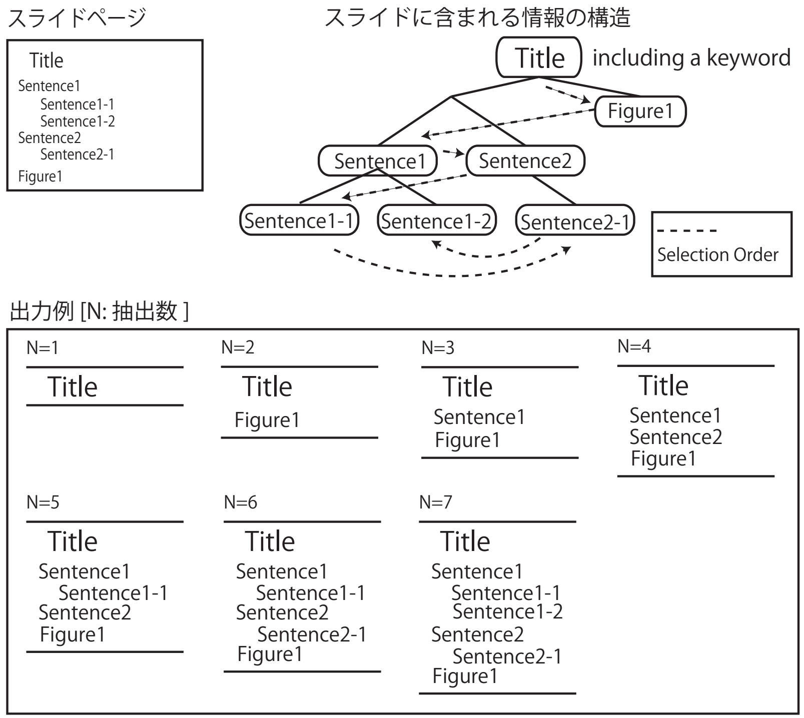


図 2 提示情報抽出手法の例
Fig. 2 An Example of Presentation Information Extraction

の情報がないため，さらに下の階層に着目し，その階層の右から “Sentence1”，“Sentence2”と順に抽出される．さらに下の階層に着目し，1 つ上位ノードから最右にある情報として，“Sentence1-1”，“Sentence2-1”と順に抽出され，最後に “Sentence1-2”が抽出される．以上の抽出順序が優先順位となり，指定された N の数だけ抽出された情報をインタフェースに提示する．その際，各情報の階層関係は字下げを使用し，下位階層の情報の表示位置が左になるように表示される．

3.3 想定する効果

提案手法により，以下の 2 点の効果が期待される．

- 縮小画像を提示することで，言語化できない情報の存在を気付かせたり，どの程度の情

Vol.2010-DD-76 No.2
2010/7/22

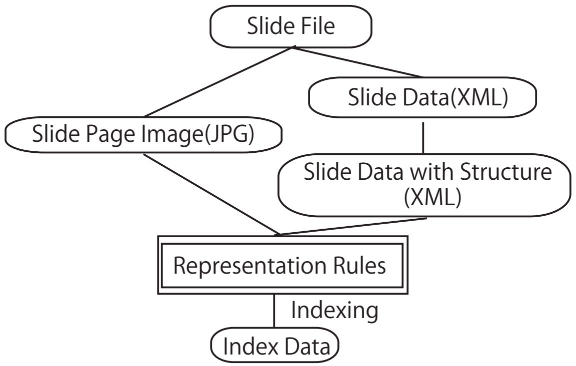


図 3 提案手法の処理手順の概要
Fig. 3 Overview of procedure of a proposed method

報量が含まれているのかスライドに含まれる情報の全容を知ることができる．

- 検索要求に関連するテキスト情報を抽出し，構造関係を保持した提示を行うことで，スライドの内容を局所的な切り取った情報であっても，正確にそのスライドを評価することができる．

4. システム概要

提案手法の処理手順の概要について，図.3 をもとに説明する．

提案手法では，まずフィルタープログラムによって，スライドファイルのページごとに，“テキスト”，“写真”，“線”などのプリミティブなオブジェクトとそれらオブジェクトに対しページ上の縦横位置やフォントサイズなどの情報を付与し，XML 形式のデータとして抽出を行う．そして，それら XML 形式のオブジェクトデータに対し，構造抽出プログラム[3]を使用して，各オブジェクトに”タイトル”，”本文”，”図”，”表”及び”装飾”のいずれかの属性を割り当て，それらを木構造となるような関係を規定した構造情報のタグを付与する．また，スライドファイルからページ画像を抽出しておき，オブジェクトに関する情報が定義されたデータと関係付けて，ページ単位でのデータ管理を行う．提案手法はページ画像データとオブジェクトに関する情報が定義されたデータをもとに，スライドページの情報を小領域に適用可能な提示表現を生成する．

以上のような一連の処理は検索インデックス作成の前処理として，従来の検索システムにそのまま組み込むことができる．

現状のシステムでは，スライドファイルとして Microsoft PowerPoint 形式のファイルに対応しており，データ抽出と画像変換のプログラムは Microsoft Visual Studio C# によって実装されている．また，検索エンジンとその提示インタフェースは Web アプリケーションとして，Java Servlet によって実装されている．

5. 評価実験

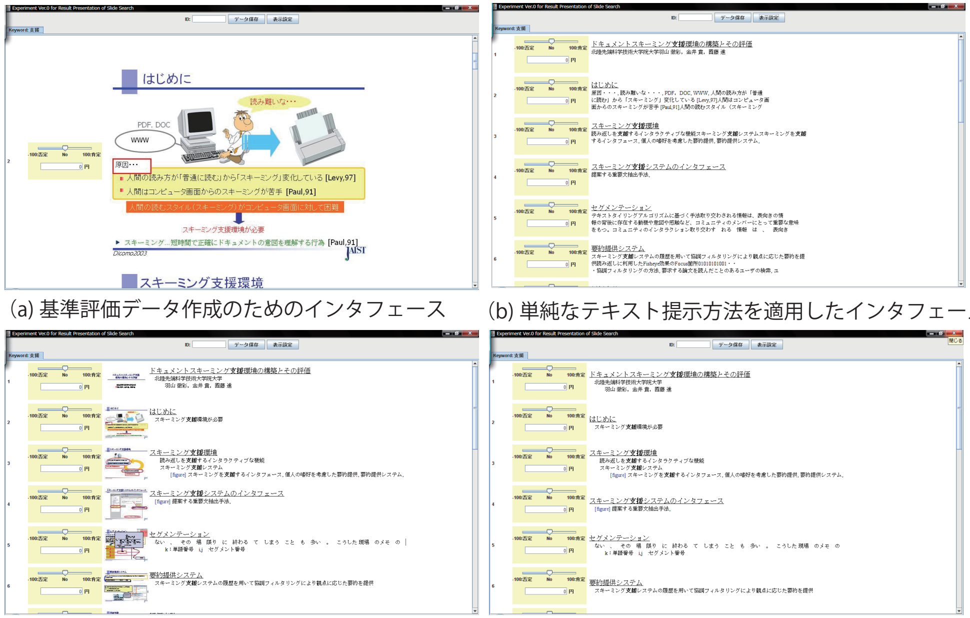
5.1 実験概要

小領域であってもスライドの内容を把握し易いように提示する提案手法の有効性を検証するために，各スライドページの基準評価付けたデータをもとに，提案手法と従来の単純なテキスト提示方法を適用した提示により，どの程度正確に把握できるかの比較した．そのために本実験ではスライド基準評価データの作成と各提示方法を適用した提示による評価データの収集のために，図.4 に示すようなインタフェースを作成し，利用した．また，提案手法のレイアウト表示だけの有効性も確認するために，提案手法に画像が提示されない場合も比較対象として加えた．

各スライドページの評価付けには評価対象にレイアウトやデザインがともなうため，感覚的に評価できる方法が有用であると考えた．そこで，本実験のスライドページの評価付けには，Willingness to Pay(WTP) と Experience Utility(EU) という指標[4] によって，検索結果の Web ページに対して感覚的に有効性と興味をそれぞれ測定可能な Arapakis ら[1] の方法を用いた．その手順としては，まずスライドページごとに WTP 測定のために図.5 に示すような金額を入力するテキストボックスと EU のためにスライド式に値指定可能なスライダーを用意し，各提示方法が適用された各ページを見ながら評価値を付けることを行う．また，WTP の金額の値に対しては，各値が 0 から 1 の値となるように評価者ごとに正規化を行った値を用いる．

評価付けデータは被験者として情報検索に慣れている大学院生 4 名に対して実施された．対象となるスライドデータは Web 上から，一般的な話題として IT 関係の Web ニュース項目から「グリッド」,「クラウド」,「電子書籍」,「YouTube」の 4 種類を検索子として採用し，各 20 個のスライドファイルを収集した．そして，検索子を含まれているスライドページから,「検索子の含まれる位置の違い」,「図表の有無」,「レイアウト (字下げ) の有無」および「テキスト情報量の多少」などのスライド内容の多様性を考慮して，話題ごとにスライド数を 20 枚ずつに絞り込んだ．また，提示情報抽出量は，一般的なスライド検索システムの表示行数である 4 とした．

Vol.2010-DD-76 No.2
2010/7/22



(a) 基準評価データ作成のためのインタフェース

(b) 単純なテキスト提示方法を適用したインタフェース

(c) 提案手法を適用したインタフェース

(c) 提案手法を適用したインタフェース
（スライド画像無し）

図 4　実験で用いたシステムインタフェース
Fig. 4　Interface of a system used in the experiment

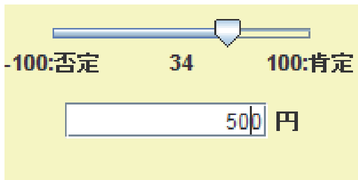


図 5　WTP と EU に基づいた評価付けインタフェース
Fig. 5　Inteface for evaluation based on WTP and EU

表 1 基準評価データと各提示方法での WTP 評価データとの相関係数
Table 1 Correlation between basic evaluated data and data evaluated by using each presentation method

	単純なテキスト提示	レイアウト付きテキスト提示	提案手法による提示
評価基準データとの相関係数	0.44	0.52	0.62

表 2 基準評価データとの評価値 WTP の差が大きいページ数（WTP の値の差が 0.3 以上の場合）
Table 2 Frequency of Slide Page which have difference between WTP values (the value of the difference is more than 0.3.)

	単純なテキスト提示	レイアウト付きテキスト提示	提案手法による提示
図表が含まれている	13	8	3
全体の情報量が抽出する情報量より少ない	10	7	3
字下げが含まれている	12	9	15
全体の情報量が抽出する情報量より多い	12	10	11
検索子がタイトル以外に照合されている	10	10	10
字下げが含まれている ＆ 検索子がタイトル以外に照合されている	3	2	7

被験者は 1 人当たり，3 つの提示手法とスライド基準評価データの作成に対して，それぞれ 20 枚のスライドの評価を行い，合計 80 枚の評価付けデータを作成した．また，話題と提示手法とのカウンターバランスをできるだけ考慮し，被験者ごとに 3 つの提示手法とスライド基準評価データ作成に用いる話題のスライドの組み合わせを変えることを行った．また，被験者属性の調査と実験に関する定性的データを収集するために，事前と事後にアンケート調査を行った．

5.2 結果と考察

スライドの基準評価データと，各提示方法を使って得られた評価値データとの相関係数を表.1 に，その基準評価データとの評価値 WTP の差が大きいページ数をスライド内容ごとに分類したものを表.2 に，それぞれ示す．

スライドの基準評価データと相関が強い順序は表.1 が示すように,「提案手法による提示」,「レイアウト付きテキスト提示」,「単純なテキスト提示」であった．そのため，スライドの内容を限られた小領域でより正確に把握するための提示方法としては，単純にテキストを並べるよりもレイアウト構造を付与させる方が有効であり，またテキストだけでなく，スライドページ画像も提示させることが有効であるといえる．

一方で，基準評価データと大きく異なる評価を行ったスライド内容において,「図表がある場合」と「情報量が少ない場合」に関しては提案手法による提示が単純なテキスト提示に比べ有効であったが,「字下げなどのレイアウトがある場合」,「検索子がタイトル以外に照合している場合」および「情報量が多い場合」に関しては両者の提示方法において，ほとんど差がみられなかった．また,「字下げなどのレイアウトがある場合」かつ「検索子がタイトル以外に照合している場合」には，単純なテキスト提示の方が有用な傾向がみられた．以上から，ページ画像や属性情報の付与により図表の存在を与えることが，スライドの内容をより正確に把握することを促しているといえる．また，スライドに含まれる情報の多くを提示し，正確にレイアウトを与えることが有用であるが，一方では，スライドに含まれる情報の一部を切り出して，レイアウトを付与して提示することは，それほど有効でないことが考えられる．この点の調査に関しては今後の課題とする．

6. おわりに

本研究ではスライド情報探索の効率性を高めるために，スライドページから検索要求に関連する情報の抽出手法を提案した．提案手法は小領域であってもスライドの内容を把握できることを考慮し，任意の表示領域に応じて検索要求に関連する情報を抽出し，テキスト表示領域と画像表示領域を持つ情報提示インタフェースへ表示を行う．テキスト表示領域ではスライドの情報構造にもとづき，検索子が含まれている箇所とその下位階層の情報を指定行数だけ選択し，レイアウト構造が持つ関係を保持した提示を行う．画像表示領域では指定行数に応じて，スライドページ画像が縮小化され，提示される．評価では，提案手法が従来の単純なテキストだけの提示に比べ，スライドの情報をより正確に把握できることを示すとともに，テキストをレイアウト化することの有効性も示した．

今後の課題はさらに実験データを増やし，より有用なスライド情報の抽出手法を開発することが挙げられる．また今回の実験結果を利用し，スライド情報検索のためのランキングアルゴリズムの開発にも着手していきたい．

謝辞 本研究成果の一部は，財団法人電気通信普及財団 平成 22 年度研究調査助成金により実施されたものである．

参考文献

1) Arapakis , I. and Jose , J.M. and Gray, P.D., *Affective feedback: an investigation into the role of emotions in the information seeking process*, Procs. ACM SI-

GIR Conference on Research and Development in Information Retrieval, pp.20–24 (2008).

2) Guo Min Liew, Min-Yen Kan: *Slide image retrieval: a preliminary study*, Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL ’08),pp.359.362 (2008),

3) 羽山徹彩，難波英嗣，國藤進: “プレゼンテーションスライド情報の構造抽出”, 電子情報通信学会論文誌. J92-D(9), pp.1483–1494, (2009).

4) Irene Lopatovska , Hartmut B. Mokros, *Willingness to pay and experienced utility as measures of affective value of information objects: Users’ accounts*, Information Processing and Management: an International Journal, v.44 n.1, p.92-104, January, 2008

5) Min-Yen Kan: *SlideSeer: a digital library of aligned document and presentation pairs*, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL ’07),pp.81.90 (2007).

6) Pia Borlund, *The Concept of Relevance in IR*, Journal of the American Society for Information Science and Technology, Vol. 54(10), pp.913–925 (2003).

7) 横田治夫 (研究代表者)： **教育的コンテンツを対象とした高度情報統合・配信に関する研究**, 科学研究費補助金特定領域研究「情報学」A02 コンテンツの生産・活用に関する研究,(2001 年度から 2005 年度).