京都アート・データベースのための高度情報検索方法の研究プロジェクト

マルチメディアデータベース構築のための効率的な類似検索手法の研究

川越 恭二
理工学研究科

概要

現在，絵画などの歴史文化財などを効率良く検索するための手法が注目されている．特に絵画等の画像情報，古文書やそれらに対する文献等のテキスト情報は重要であり，これらの検索を容易に，高速に行うことによって，歴史文化財に対する分析をより容易に行うことができると考えられる．そこで本研究プロジェクトでは，画像情報やテキスト情報など複数のメディアで構成された電子文書，つまりマルチメディア電子文書を検索するための手法について述べる．本報告では，複数の検索システムを統合したマルチメディア電子文書検索手法について，および利用者に対して直感的に理解可能な検索結果を生成するための手法を提案する．評価の結果，検索システムの精度が向上すること，および直感的な検索結果を生成可能であることが明らかとなった．一方，京都など歴史文化財が点在する都市を活性化させるためには，都市の観光客の行動履歴の把握を行うことが重要であるといえる．ところが，大量の観光客の行動履歴を把握するためには，計算機による膨大な計算時間が考えられる．そこで，観光客の行動履歴を高速に把握するために，計算機の処理時間を効率的に削減する手法を提案する．

Research on Advanced Information Searching for Kyoto Art Databases

Research on Effective Retrieval Method for Multimedia Databases

Kyoji Kawagoe
Graduate School of Science and Engineering

Abstract

Currently, many researchers have been focused on developing methods of archiving cultural assets into databases. Especially, archiving and retrieving historical pictures, textual data is important to analyze these assets. In this project, we propose a method for retrieving multimedia electronic documents which consist of graphical, textual, and the other data. In this report, we describe the combination method of multiple mono-media retrieval system to retrieve multimedia electronic documents. Using our proposed retrieval method, users can retrieve information the users need, and the users also show these retrieval results clearly. On the other hand, we describe the method to do similarity search of users' behaviors, the walking path in the historical city like Kyoto. However, it takes hours of time to calculate these similarity values. To resolve this problem, we propose "Through-Area", a method to reduce calculation time of time-series data.

1　検索結果を統合するためのスコア統合関数選択手法

現在，コンピュータネットワークの普及と共に大量のテキストや画像などが流通しているため，それ

らを検索するためのシステムが数多く提案されている．例えば，利用者がテキスト中の単語の出現頻度を基準とした検索を行う必要がある場合，利用者の目的に合致した検索システムを用いることによって，最も精度の高い検索を行うことができると考えられる．ところが，利用者が文の語尾に着目した検索を行う必要がある場合には，単語の出現頻度による検索システムでは十分な精度を得ることができない．また，利用者の検索目的が明確で無い場合や，利用者が全ての検索システムの特性を把握していない場合，利用者の必要な情報を得ることができないと考えられる．

そこで，利用者の検索目的に応じた検索を容易に行うために，複数の異なる検索システムを統合する方法を考える．この検索システムでは，あらかじめ利用者にとって必要であると考えられる複数の検索システムを検索サブシステムとして用意し，各々の検索サブシステムから出力された検索結果を統合する方法である．利用者の検索目的に適合した検索サブシステムが検索システムに含まれている場合，有効となる検索サブシステムが出力した検索結果が統合後の検索結果に反映されるため，検索精度が上昇すると考えられる．

本研究では，最適な統合関数を自動的に選択するための方法として，シャノンの情報量の概念 [1] を援用した尺度によって，検索結果だけから統合関数が最適であるかどうかの度合いを測定する．検索結果だけから統合関数が最適であるかどうかを実際に測定するためには，利用者による検索対象への適合，不適合の判断が必要である．ところが，これらの作業は利用者にとって大きな手間となると考えられる．この手間を軽減するためには，検索結果だけから検索システムによってその精度を推定する必要がある．一方，各々の検索結果に含まれる，検索対象へのスコア分布に注目することによって，統合関数が最適かどうかを推定することができるのではないかと考えた．そこで本研究では，シャノンの情報量の概念を用いてスコア分布から精度を測定し，統合関数の自動選択に用いた．

1.1 基本的な考え方

本稿で提案する方法は，統合検索結果におけるスコアの分布から検索結果を出力する方法である．そこで，まず検索精度が高いと考えられる場合のスコア分布，つまり理想的なスコア分布について考える．

一般に，利用者が入力する問合せに適合する検索対象，つまり正解検索対象集合の数は，検索対象集合全体の数と比較すると非常に少ないと考えられる．例えば，TREC [2] や NTCIR [3] 等に代表される，情報検索システムを評価するためのテストコレクションには，問合せとそれに対する正解検索対象集合が示されている．これらのテストコレクションにおいても，やはり正解検索対象集合は検索対象集合全体の数と比べて非常に少ない．

このような条件下では，ある検索システムによって高いスコアを付与された検索対象の数が小さいほうが，その検索システムは十分に検索対象群から正解検索対象を見つけ出すことができているのではないかと考えられる．つまり，検索システムが十分少ない検索対象集合を抽出することができた場合には，その検索システムは検索対象集合のうちの一部分に利用者の検索意図と合致した特徴を発見できていると推定できる．この場合，その検索システムの検索精度が高いことが推定できる．

以上の考え方を，スコアの分布から推定する方法に応用する．ここで，二つの統合関数 F_1，F_2 を用いた検索システム $R(F_1)$，$R(F_2)$ を考える．それぞれの検索システムは，統合関数の違いを除いて全く同一のシステムであるとする．さらに，ある問合せをそれぞれの検索システムに入力した場合のスコアの分布は図 2 (a), (b) に示すように表現できたとする．これらの図の x 軸はスコアの値を表しており，y 軸は x 軸で表示されているスコアの値と 1 との間のスコアが付与されている検索対象の数を表している．例えば，x 軸が 0.1 の部分の y 軸の値は，検索システムによって付与されたスコアの値が 0.1 以上 1 以下である検索対象の数を表している．

これらの図から，$R(F_1)$ におけるスコア分布では高いスコアが付与された検索対象の数が $R(F_2)$ の場

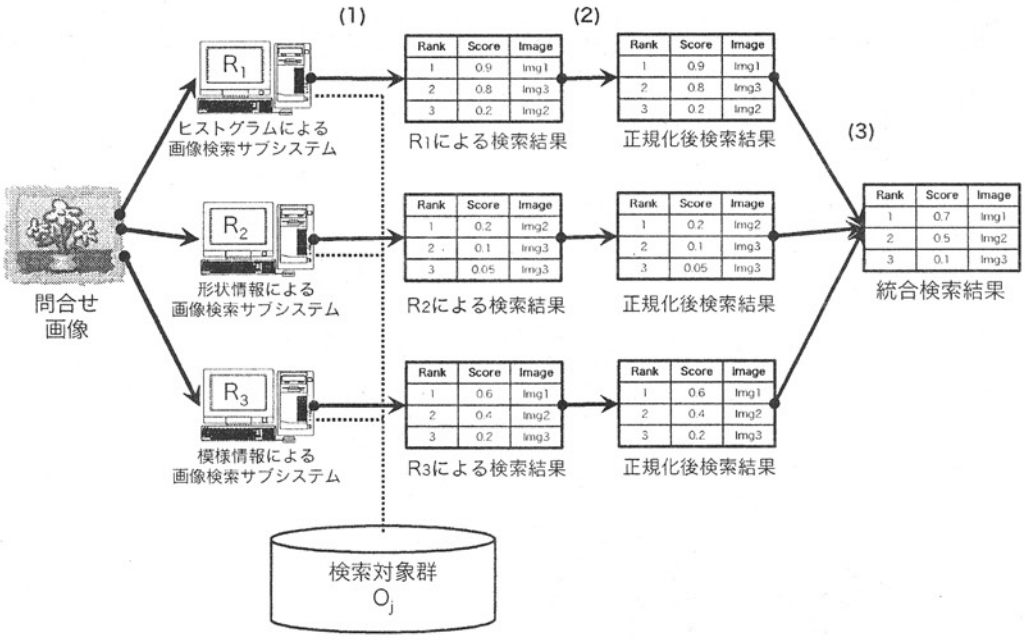
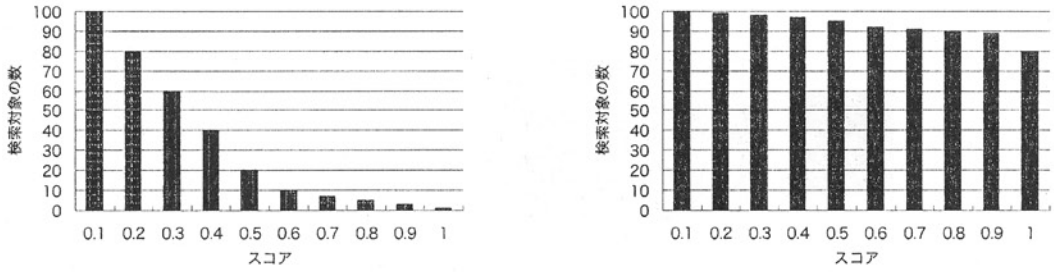


図 1　複数の検索サブシステムを組み合わせた検索システム



(a) 検索システム $R(F_1)$　　(b) 検索システム $R(F_2)$

図 2　検索システム $R(F_1)$ と $R(F_2)$ を用いた場合のスコア分布

合よりも少ないことが分かる．つまり，前節で述べた仮定から，$R(F_1)$ は利用者の検索意図と合致した検索対象だけ対して高いスコアを付与している可能性が高いと考えられる．以上の議論から，$R(F_1)$ は $R(F_2)$ よりも良い検索システムであると考えられるため，検索システムは統合関数 F_1 を選択する．

1.2　実験結果および考察

図 3 に，各々の問合せを用いた場合の適合率を示す．左から順に，三つの統合関数を用いた検索システムの平均適合率のうち，最も高いもの，平均，最も低いもの，提案手法によって選択されたものをそれぞれ表している．

これらの実験結果から，11 個の問合せにおいて最も平均適合率が高い統合関数を選択していることが分かる．

本実験で用いた問合せでは，形状情報に注目した問合せが多い．このような場合，形状情報を用いた検索サブシステムによる検索結果が統合検索結果に反映されている統合関数が良い傾向があることが予期できる．我々は，これらの問合せの傾向から CombMNZ を用いることが最も利用者の検索意図に適していると考えていた．そのため，もし提案手法を用いて CombMNZ を用いた場合には検索精度が向上すると予想された．実験の結果，9 個の問合

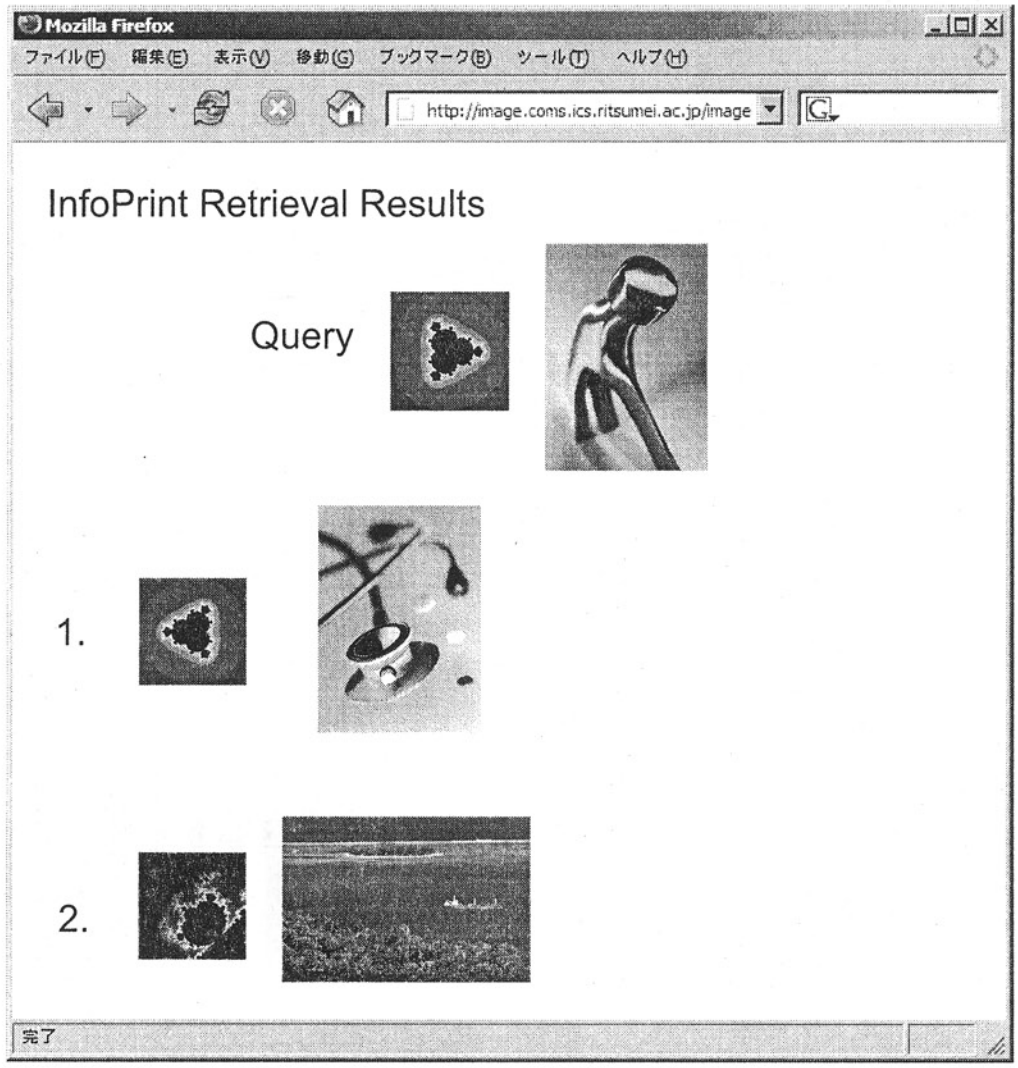


図 3　重み付き平均適合率の比較

せにおいて CombMNZ が選択されていることが分かった．つまり，利用者の検索意図に合致した統合関数が選択されている可能性が高いことが分かる．

以上の結果から，確かに本提案手法によって検索精度が高いと考えられる統合関数を選択できることが分かった．

2　InfoPrint: 直感的な検索結果の生成

近年，多くの研究者によって画像検索システムが提案されているが，それらのシステムが出力する検索結果は，順位付きリストであることが多い．検索システムは，RSV (Retrieval Status Value) と呼ばれる，利用者が入力した問合せ画像とと検索対象画像との類似度を基に検索結果リストを生成する．利用者は，検索結果に含まれる RSV を検索対象画像と共に閲覧することによって，検索対象画像がどの程度問合せ画像に類似しているか判断することができる．そのため，利用者は最小限の検索対象画像だけを閲覧することができるため，多くの画像を閲覧する負担を軽減することができる．つまり，RSV は画像検索などマルチメディア検索システムにおいて重要な情報であるといえる．ところが，利用者は RSV を閲覧しないことが多いため，利用者によって RSV は十分活用されていない情報であるといえる．

本研究では，検索対象画像群と問合せ画像を *InfoPrint* と呼ばれる画像に変換し，利用者に提示す

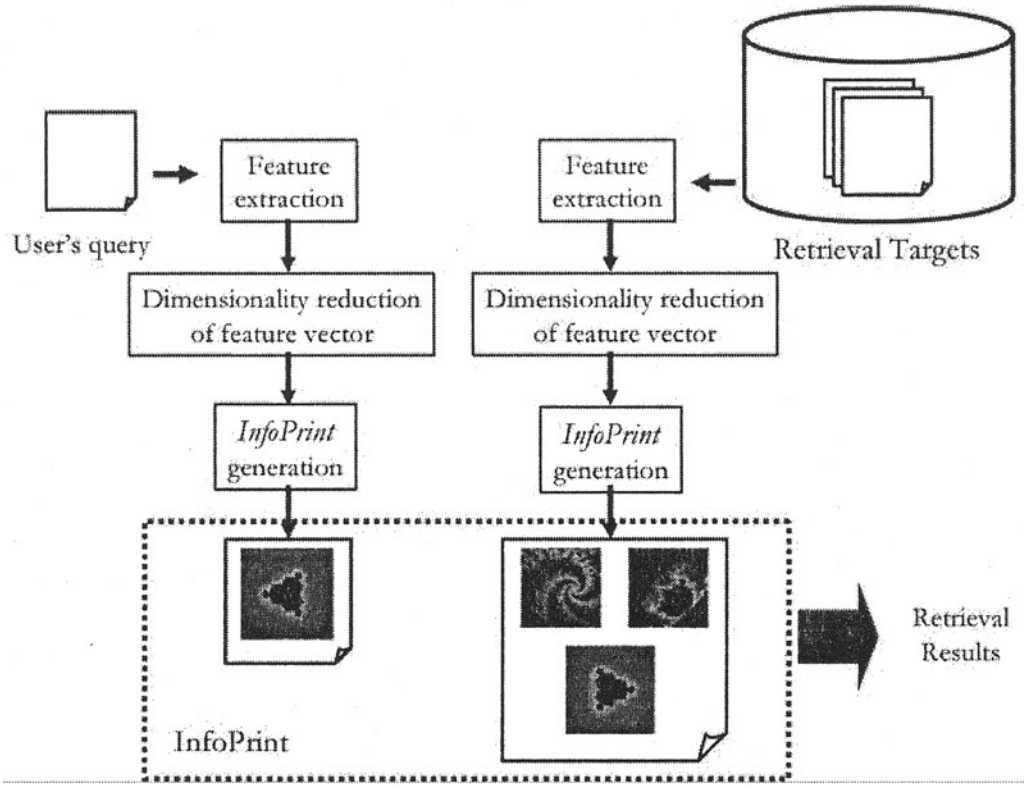


図 4　*InfoPrint* を用いた画像検索システムの結果提示

る手法についての提案を行う．*InfoPrint* とは，画像から抽出された複数の特徴量を画像として表現したものである．つまり，問合せ画像と検索対象画像の *InfoPrint* がそれぞれ類似している場合には，問合せと検索対象画像は類似している．利用者は *InfoPrint* を検索対象画像と同時に閲覧することによって，利用者は直感的に複数画像間の類似度を認識することができる．

2.1　InfoPrint の生成手法

InfoPrint は図 4 に示すように，以下四つの手順によって生成される．

1. 検索システムによって画像から特徴量を抽出する．ここで特徴量は 2000〜3000 次元程度の高次元のベクトルとして表現される．
2. 高次元の特徴ベクトルを 10 次元程度の低次元ベクトルへ削減する．
3. 特徴ベクトルの各要素をパラメータとして，マンデルブロー集合図形（*InfoPrint*）を作成する．
4. 検索対象と共に *InfoPrint* を利用者に提示する．

3　観光客の行動履歴データに対する類似検索高速化手法

GPS によって得られる利用者の位置情報を用いて，初めて行く観光地での現在地に関連した情報を提供するサービス．過去にその場所を訪れたことのある人を参考にし，これから訪れる場所へのルートや次にどちらに進むことがおすすめかを掲示するといった利用者の行動支援サービス等が可能となる．また，モバイル機器，携帯電話などの情報通信機器や，ブロードバンド，無線 LAN などのネットワーク・サービスが急速に普及するなかでは，多種多量の情報をやり取りできる環境が広がっている．そのために，利用者の国籍，性別，年齢，目的，その地域への訪問回数などによる多くの情報を加えて用いることによって，利用者の嗜好に合わせたサービスを提供することも可能となる．こうしたユビキタス

環境においては，様々な自分の知りたい情報を，知りたいときに知ることが可能となり，GPS を用いた利用者の現在地に関する位置情報を生かしたサービスへのニーズが高くなると考えられる．また，そういった状況下では過去の移動データからの効率的な検索システムへのニーズは高くなると考える．しかし，行動履歴データへの検索処理は，データの持つ特性から非常に時間のかかる処理となる．また，検索処理精度も低下するという問題点がある．

そこで，観光においての行動支援サービスを想定し，利用者の観光ルートにおける行動履歴データと特徴が類似している行動履歴データを大量のデータから検索するために，スルーエリアを用いた類似検索手法を提案する．行動履歴データは利用者の位置情報を時系列データとして表現したものであるため，本論文では行動履歴データを時系列データとして扱う．また，スルーエリアとは実際の観光ルートにおいて多く訪れるポイントをエリアとして定義するものである．問い合わせデータに含まれるスルーエリアを探しだし，検索対象データからそのスルーエリアの含まれているものを絞込む．この絞込みを行うことによって，検索対象データのデータ数の削減を行うことができる．削減を行った検索対象データから，類似度を求め類似検索処理を行う．スルーエリアを用いることによって行動履歴データへの類似検索処理の検索処理精度の低下させることなく，処理時間を短縮し，検索処理効率を改善することができる．

4 おわりに

本稿では，京都における歴史文化財のアーカイブに最も必要な，主に画像を中心としたマルチメディアデータを高精度に検索するための手法についての説明を行った．さらに，京都観光における観光客の行動支援を行うために，行動履歴の高速な検索手法についての提案を行った．

今後は，画像以外の他メディアに関するデータベースへの検索手法に関する研究，および高精度化，高速化についての研究を行う予定である．現在，計算機上でデータを扱うための標準的なフォーマットである XML (Extensible Markup Language) が，データベース上や WWW (World Wide Web) を中心に普及している．つまり，歴史文化財のアーカイブを XML 形式によって保存しておくことによって，より多くの利用者によって利用することができると考えられる．ところが，現在 XML は主にテキストを格納するために用いられており，マルチメディア XML については十分に研究されているとはいえない．そこで，今後はマルチメディア XML の検索手法について研究を行う予定である．

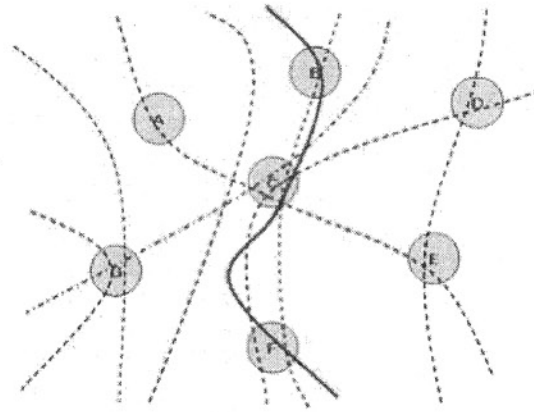


図 5　観光客の行動履歴に対する高速な類似検索

参考文献

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, Vol. 27, pp. 379 – 423, 1948.

[2] National Institute of Standards and Technology. Text retrieval conference (trec). http://trec.nist.go.jp/.

[3] 国立情報学研究所. Ntcir 情報検索システム評価用テストコレクション構築プロジェクト. http://research.nii.ac.jp/ntcir/.