社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－GN－51　(16)
2004／3／19

デジタル文書の共有・活用支援システム

～マルチメディア検索システム MIRACLES による実現と評価～

遠藤進　馬場孝之　椎谷秀一　上原祐介　増本大器　長田茂美

オフィス内では，プレゼンテーション用ツールやワードプロセッサで作成されたデジタル文書を共有することが広く行われている．そこで，それらのデジタル文書を有効に活用するために，大量の文書の中から必要とする文書を効率よく検索する手段が必要になってきている．本稿では，各デジタル文書から抽出した図を分類配置して表示し，それらを眺めながら目的の文書を直感的かつ効率的に探すことができるデジタル文書共有・活用支援システムについて述べる．さらに，文書内のテキストを対象とした全文検索システムと本システムとの比較を行い，文書の再利用を行う場面を想定した実験において，全文検索システムより３倍から５倍高速に検索できることを確認した．

Digital document sharing system
-Application of MIRACLES to digital document retrieval and its evaluation-

Susumu Endo, Takayuki Baba, Shuichi Shiitani, Yusuke Uehara, Daiki Masumoto, and Shigemi Nagata

In offices, digital documents such as presentation documents and word processor documents are shared. To reuse these digital documents effectively, we want to retrieve the desired document from a large number of documents. In this paper, we propose digital documents sharing system. This system arranges figures in documents in virtual 3D space, and user can retrieve a digital document by walking through the space. In experiments, we compared retrieval time taken by this system with one taken by a full text search system, and we found that by this system users could retrieve the desired document three or five times faster than the full text search system.

１．はじめに

オフィス内では，プレゼンテーション用ツールやワードプロセッサで作成されたデジタル文書を蓄積し，それを共有することが広く行われている．それにより，他人が作成したデジタル文書を一部修正して利用したり，図を部分的にコピーして新しいデジタル文書を作成したり，といった再利用が可能になる．しかし，蓄積される文書の数が増えるにつれ，大量の文書の中から必要とする文書を効率よく検索する手段が必要になってきている．

従来デジタル文書の検索には，各デジタル文書内で使用されているテキストを対象として全文検索を行う手法が広く用いられてきた．

しかし，デジタル文書は通常，図とテキストから構成されており，テキストだけでなく図を手がかりに探したい場合もある．一般的に図は記憶に残りやすく，文書を特定の図と関連付けて記憶していることも多い．また，図を作成するには時間がかかることが多いため，図を再利用する利点は大きい．

[1]（株）富士通研究所　IT メディア研究所
INFORMATION TECHNOLOGY MEDIA LAB.,
FUJITSU LABORATORIES LTD.

一方，我々はこれまでに，画像を一覧表示し，それをユーザが眺めて目的の画像を探し出すマルチメディア情報検索システム MIRACLES (Multimedia Information RetrievAl, CLassification, and Exploration System)の研究開発を行ってきた[1][2]．MIRACLES では，画像から色や構図などの特徴量を抽出し，その特徴量が似ている画像同士が近くに集まるように仮想三次元空間に分類配置する．ユーザはその三次元空間を動き回り，目的の画像に色や構図が似ているものが集まっている付近を重点的に探索することで，目的の画像を直感的かつ効率的に探すことができる．この検索方法では，人間の検索能力を利用し，計算機でそのサポートをすることで，計算機だけ，あるいは人間だけでは困難な検索を可能としている．

本稿では，この MIRACLES の検索方法をデジタル文書に対応させ，デジタル文書を直感的かつ効率的に検索できるデジタル文書共有・活用支援システムについて説明する．さらに，本システムと全文検索システムとの比較を行うため，検索にかかる時間を測定する評価実験を行い，本システムの有効性を確認した．

２．デジタル文書共有・活用支援システムの構成

デジタル文書共有・活用支援システムのベースとなる MIRACLES は，テキストによる意味的検索と画像による視覚的検索とを兼ね備えたマルチメディア情報検索システムであり，1）クローラによる情報収集，2）情報の類似性に基づく配置，3）インタラクティブな情報検索といった機能を持つ．デジタル文書共有・活用支援システムでは，これらの機能をデジタル文書向けに拡張して使用している．

以下では，デジタル文書共有・活用支援システムにおける各機能について説明する．

２．１．クローラによる文書収集

本システムでは，あらかじめ検索対象のデジタル文書を解析し，文書に含まれている図とテキスト，さらにその特徴量の情報をセットとして抽出しておく．

検索対象となるデジタル文書の登録方法として下記の２種類の方法が考えられる．本システムでは両方の方法に対応している．

一つは，新しい文書や更新した文書をユーザが明示的に登録する方法である．登録時にユーザに分類情報やその文書を使用した会議名などのメタデータを入力させることが可能であり，それらのメタデータを利用した検索を行うことができる．しかし，デジタル文書を新規作成したり編集したりするたびに登録しなおすのは面倒であるため，ユーザが必要な文書の登録を怠る場合も多い．

もう一つの方法として，コンピュータ上でユーザが文書を置いておく場所を決め，システムがその場所にあるデジタル文書を自動的に登録する方法が考えられる．分類情報やその文書を使用した会議名などのメタデータを使用した検索はできなくなるが，ユーザは新しく文書を作成したり，編集したりする際に明示的に登録する必要がない．図１にこの方法における収集処理を示す．システムは各クライアントのマシン上で文書が保存してある場所を定期的に参照してデジタル文書を探し，見つかったデジタル文書から，文書内

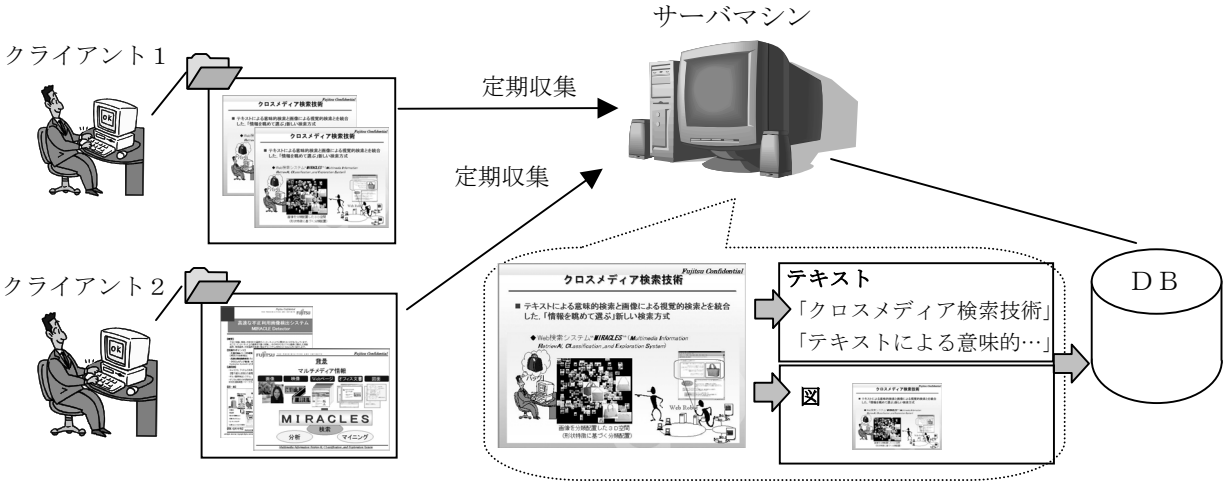

図１　デジタル文書共有・活用支援システムにおける文書収集

にある図とテキストをペアにして収集する．

プレゼンテーション用文書では，スライドのレイアウトが検索の際に重要になるため，スライドのサムネイルを図として使用する．また，スライド中で使用されているテキストやスライドに関連付けられたメモをペアとして収集する．ワードプロセッサで作成された文書では，テキストが主となり，ところどころ図が挿入されている場合が多いため，図とその周りにあるテキストをペアとして収集する．

さらに収集した図から，その色や構図に関する特徴量を抽出する．図の色の特徴量として，各画素の RGB 値を HSI 色座標値に変換し HSI 空間を格子状のブロックに分割して各ブロックに含まれる画素数をカウントした HSI ヒストグラム特徴量を用いている[3]．また，図の構図の特徴量として，輝度値を Wavelet 変換した特徴量を用いている[4]．

また，テキストからは意味内容を表す特徴量を抽出する．テキストの特徴量として，特定の単語がテキスト中に出現した頻度をベクトルで表したものである単語頻度特徴量を用いている[5]．

２．２．文書内容の類似性に基づく配置

ユーザが，探している情報に関連したキーワードを入力すると，本システムは，キーワードを含むテキストとペアになっている図だけを選び出す．長い文書でもキーワードに関係した図だけが選び出されるため，文書全てを参照する必要が無くなる．

次に，本システムはそれらの図を特徴量が似ているもの同士が近くに集まるように平面に分類配置する．この配置には自己組織化マップを用いている[6]．自己組織化マップでは，高次元の特徴ベクトル空間を低次元空間に写像する際に，高次元空間における分布の状態を低次元空間においても，できるだけ保存するように配置することができる．図２にプレゼンテーション用文書のスライドを Wavelet 特徴量に基づいて配置した画面例を示す．配置の結果，構図が似ているスライドが集まって配置されている．自分が作成したスライドを探す場合などはある程度構図を覚えているため，記憶にある構図に似たスライドが集まっている付近を探せばよく，直感的かつ効率的に探し出すことができる．

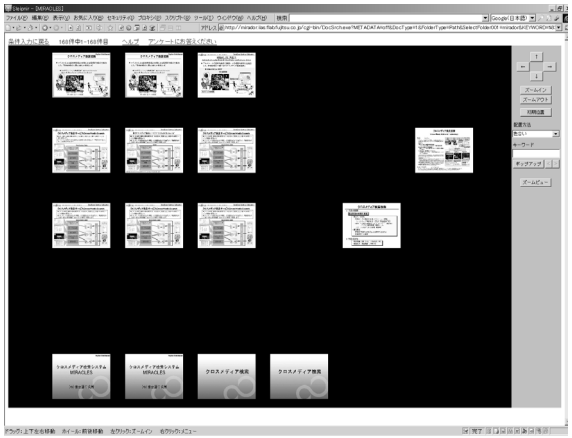

図２　プレゼンテーション用文書の各スライドを Wavelet 特徴量に基づいて配置した画面例

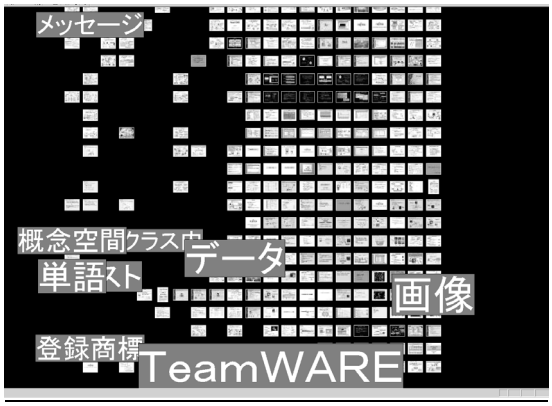

図３　プレゼンテーション用文書の各スライドを単語頻度特徴量に基づいて配置した画面例

また，単語頻度特徴量で配置することで，テキストの意味内容が類似したもの同士が近くに集まるように配置することができる（図３）．テキストの内容によって配置を行った場合は，色や構図の特徴量での配置を行った場合とは異なり，見た目の類似性を利用して探すことが困難である．そこで，どのような内容の図がどのあたりに配置されているのかを示すランドマークとして，テキストから自動抽出されたキーワードをラベルとして表示する[5]．

２．３．インタラクティブな文書検索

本システムは，ユーザが入力したキーワードにヒットした図を，仮想的な三次元空間上に配置して表示する．最初の視点は全ての図を見渡せる位置に設定される．ユーザはこの三次元空間内をフライスルーすることによって図に近づいて，その図が探している図かどうかを確認することができる．さらに，配置の基準となる特徴量を変更することで，色や構図，テキストの意味内容などさまざまな観点から図を探すことも可能である．

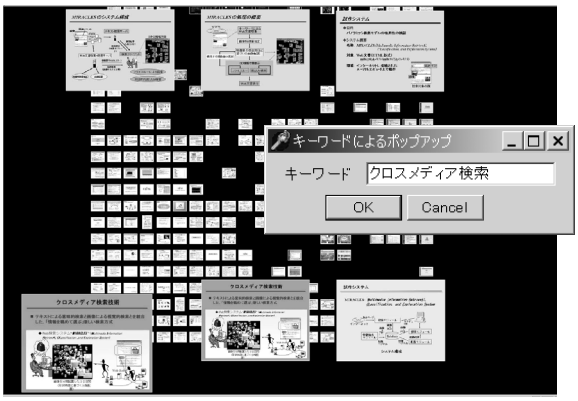

図４　ポップアップした画面例

また, ユーザはこれらの画面からキーワードを入力することで検索対象を絞り込むこともできる. キーワードを入力すると, そのキーワードを含むテキストとペアとなっている図だけがポップアップして拡大表示される. ユーザはポップアップした図だけを対象に探せばよいので効率的である（図４）[2].

さらに, デジタル文書では図の一部が変更されて再利用されることが多いため, 各図の違いが明確にわかるようにするために, マウスカーソル位置にある図を別ウィンドウに拡大表示することができる(図５). マウスを移動させることで次々とウィンドウ内の図が切り替わり, 違いを容易に把握できる.

以上のような操作によって探し出した図をユーザが選択すると, その図を含むデジタル文書を再利用できる. 具体的には, その文書作成に利用したアプリケーションが立ち上がり, 文書を編集したり, 一部をコピーして新しい文書を作成したりすることができる.

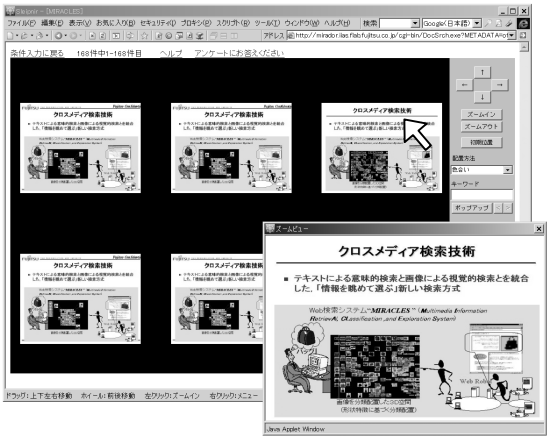

図５　類似したスライドを比較している画面例

２．４．システムの実現方法

本システムはクライアントサーバ方式で実現されており, サーバ側でキーワードによる絞込み処理を行い, その結果をクライアントに転送して表示している. クライアント側のプログラムはJava 言語で記述されているため一般的な Web ブラウザで利用することができる.

３．評価実験

本システムを評価するため, 全文検索システムとの比較実験を行った. 比較方法として本システムを使用してスライド単位で検索した場合と, 全文検索システムで文書のタイトルや概要を一覧表示して文書単位で検索した場合とで検索にかかった時間を計測した. 時間には表示する図や文書の転送・表示時間を含んでいる. 検索対象の文書は 213 件の Microsoft PowerPoint 文書(以下, PowerPoint 文書と呼ぶ）であり, スライドの総合計は 5147 枚である. また, 本システムを初心者でも使用できるかどうかを検証するため, 筆者が所属する研究所の研究員で, 本システムをあまり利用したことがない 10 人を被験者とした.

今回は, 典型的な再利用の例である３つの場合で実験を行った. それぞれ, キーワードだけで十分絞り込める場合, キーワードで件数は絞り込めるが図だけが変更された類似文書がたくさんあり結局は一つ一つ眺める必要がある場合, 見た目の印象だけしか覚えていなくキーワードでの絞込みが困難な場合である.

実験１　キーワードだけで十分に絞り込める場合

『我々の研究テーマのひとつである「イメージマイニング」のデモの操作手順を説明している文書を探すように』とだけ被験者に指示した. 「イメージマイニング」「デモ」の論理積で検索すると 5 件の文書に絞り込まれる. また, その 5 件の文書についても, 文書のタイトルから内容が推定できる.

実験２　図が一部変更された文書がたくさんある場合

図６の PowerPoint 文書を提示し,『このような文書を使用したいと思ったが, 下側のシステム構成図で使用されているキャラクタ画像を別のキャラクタ画像に置き換えた版の文書があるはずであり, その方が望ましいため, その文書を探すように』と被験者に指示した. この場合, 図だ

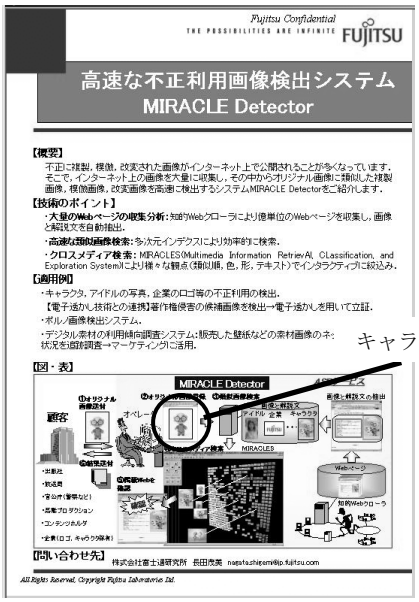

図６　実験２で使用した PowerPoint 文書

けが異なっている文書が多数あり，タイトルである「高速な不正利用画像検出システム MIRACLE Detector」をそのまま入力して検索しても 13 件の文書，14 枚のスライドがヒットする．テキストの内容は全て同じであり，それ以上のキーワードでの絞り込みは困難である．

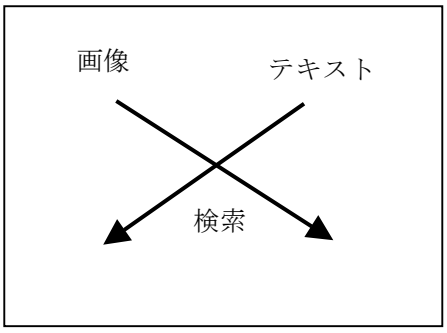

図７　実験３で被験者に提示した図

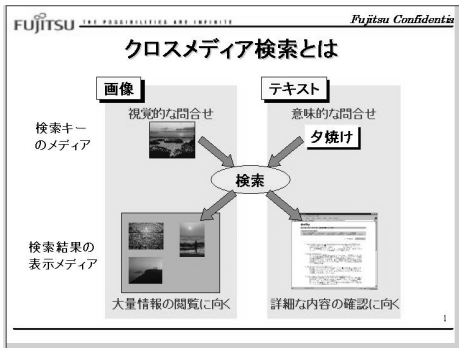

図８　実験３の検索対象の図

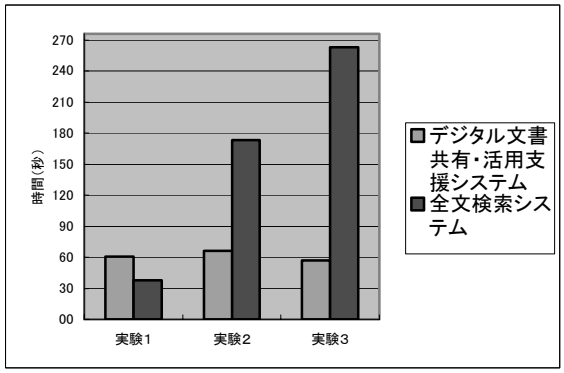

図９　実験結果

実験３　見た目の印象だけを覚えている場合

過去に見たことがあるスライドを探しているが，見た目の印象だけしか覚えていないような場合を想定した実験である．被験者に図７を提示し，『我々の研究テーマの一つである「クロスメディア検索」を説明したスライドで，内容は正確には覚えていないが，画像とテキストをクロスして検索する図を探すように』と被験者に指示した．図８が実際の検索対象の図である．キーワードとして「クロスメディア検索」「画像」「テキスト」の論理積を入力しても 43 件の文書がヒットする．スライド数では 500 枚程度あり，そのうち 110 枚のスライドに上記のキーワードが含まれる．

図９が実験結果である．３つの実験で，それぞれ提案するデジタル文書共有・活用支援システムを使用した場合と全文検索システムで検索した場合とで，被験者 10 人が検索に要した時間の平均をグラフとして表現している．

実験１は適切なキーワードを入力すれば，それだけで十分絞り込まれるものであり全文検索システムの方が速い．

実験２は，キーワードだけでは絞込みが十分できない場合であり，全文検索システムでは，結局 13 件の文書を一つ一つ見ていくしかないためかなり時間がかかっている．デジタル文書共有・活用支援システムでは，全ての文書内のスライド 14 枚をざっと眺めて絞り込むことができる．そのため，検索にかかった時間は全文検索システムの 1/3 になっている．

実験３では，実験２よりさらに絞込みが難しくなる．全文検索システムのみで検索するには，40 件以上の文書を一つ一つチェックする必要がある．結果的に 10 人中 3 人の被験者が検索をあき

らめた．デジタル文書共有・活用支援システムでは，表示されるスライド数は 110 枚に増えたが実験２とあまり変わらない時間で終了しており，検索にかかった時間は全文検索システムの 1/5 程度である．なお，検索をあきらめた被験者については，あきらめて検索を中断した時点での時間を結果として使用しているため，実際にはもう少し差が開くと思われる．

本システムに慣れていない被験者を対象とした実験だが，それでも実験２，３のような場面で全文検索システムより 3 倍以上高速に検索することができ，本システムの有効性が確認できた．また，本システムが初心者にも使えるものであることが示された．

ただし，各被験者がデジタル文書共有・活用支援システムで検索する手順を観察した結果では，全体を概観しながらフライスルーする方法しか使用しない被験者が多く，それ以外のキーワードでポップアップする機能やマウス位置の図を拡大して表示する機能などはほとんど使われず，これらの機能の有効性については十分に検証できなかった．

４．まとめ

マルチメディア検索システム MIRACLES の応用として，デジタル文書を検索するデジタル文書共有・活用支援システムを提案した．本システムでは，図から色や構図といった特徴量を，テキストからテキスト特徴量を抽出し，その特徴量に基づいて図を仮想三次元空間に分類配置する．ユーザはその三次元空間を動き回りながら，配置された図を手がかりに目的の文書を直感的かつ効率的に探すことができる．

さらに文書内のテキストを対象とした全文検索システムとの比較実験を行い，本システムが再利用を想定した実験において，3 倍から 5 倍以上高速に検索できることを確認した．

今後は，今回検証が十分に出来なかった本システムの各機能の有効性に関する検証や配置方式を変更した場合における検索効率の違いの検証などを行う予定である．

参考文献

[1] 遠藤進，指田直毅，増本大器，長田茂美，棚橋純一：画像情報とテキスト情報を統合的に利用したインタラクティブな Web 検索システム．第 5 回知能情報メディアシンポジウム予稿論文集．電子情報通信学会，pp.163-170, 1999.

[2] 上原祐介，遠藤進，指田直毅，増本大器，長田茂美，棚橋純一：MIRACLES：マルチメディア情報のパノラミック検索システムーWeb 検索への応用ー．技術研究報告データ工学研究会 DE2000-3, 電子情報通信学会, pp.17-24, 2000.

[3] 高木幹雄，下田陽久監修：“画像解析ハンドブック”，東京大学出版会，1991.

[4] 村尾晃平，安藤淳禎：画像をキーとする類似画像検索システム，1998 年電子情報通信学会 情報・システムソサイエティ大会， D-11-60，pp.175，1998.

[5] 遠藤進，椎谷秀一，上原祐介，増本大器，長田茂美：テキストによる意味的な検索と画像による視覚的な検索を統合したマルチメディア検索システム MIRACLES，DBWeb2001，IPSJ Symposium Series Vol.2001, No.17, pp.249-256, 2001.

[6] T.コホネン著，徳高平蔵，岸田悟，藤村喜久郎訳：“自己組織化マップ”，シュプリンガー・フェアラーク東京，1996.