データベースシステム　122－17
(2000. 7. 26)

複数のメディアで構成された電子文書の検索手法

鈴木 優　波多野 賢治　吉川 正俊　植村 俊亮

奈良先端科学技術大学院大学 情報科学研究科

〒 630–0101 奈良県生駒市高山町 8916–5

{yu-su, hatano, yosikawa, uemura}@is.aist-nara.ac.jp

あらまし　本稿では複数のメディアで構成された電子文書を，利用者の興味に応じて検索する手法を提案する．従来の電子文書検索に関する研究は単語の頻度情報のみを特徴量とした手法であり，HTML 文書や SMIL 文書のように静止画像や動画像といった複数のメディアを含む文書に対する検索を行うことは困難であった．本研究では，電子文書から複数の特徴量を取り出すことによって複数の特徴部分空間を生成し，利用者が興味を持つ事柄に対する特徴部分空間へ電子文書の特徴量を射影することにより電子文書の検索精度を向上させることを目的とする．

キーワード　情報検索，複数メディアの特徴量，PDF

A Retrieval Method for Multimedia Documents

Yu Suzuki, Kenji Hatano, Masatoshi Yoshikawa, and Shunsuke Uemura

Graduate School of Information Science,
Nara Institute of Science and Technology

8916–5 Takayama, Ikoma, Nara 630–0101, Japan

{yu-su, hatano, yosikawa, uemura}@is.aist-nara.ac.jp

Abstract　In this paper, we present an approach to improve a retrieval method for multimedia electric document. In the past, document features are based on only term frequency, there is no retrieval method for image and movie and et al. In this study, we extract many features from many media including documents.

Key words　Information Retrieval, Multimedia Data, Feature Extraction, PDF

1 はじめに

現在，ネットワークを用いた電子文書の流通が盛んであり，多くの電子文書が複数の計算機上で共有されている．また，電子文書を作成する環境も整備されつつあり，多くのアプリケーションが公開されている．電子文書フォーマットとして良く使われるものとして PDF (Portable Document Format) [1] が挙げられ，雑誌の版下や論文の原稿などに利用されている．また，WWW (World Wide Web) では HTML (HyperText Markup Language) [2] が用いられており，WWW 上での電子文書形式として利用されている．これら電子文書では画像データをはじめとして多くのメディアを扱うことができることが特徴の一つとして挙げられる．さらに複数のメディアを時系列で扱うフォーマットとして SMIL (Synchronized Multimedia Integration Language)[3] が提案されており，音声や映像の放送に利用されている．．このように，今後多くの複数メディアで構成された文書が増加することが予想される．したがって，今後の情報検索では文字情報のみに焦点を当てるのではなく，より多種の情報を統一的に扱う必要があると考えられる．

ところで，従来の文字情報の検索では，問合せとしてキーワードを入力する検索方式が主であるが，利用者によって文書の持つ意味が異なるため，同じキーワードで問合せを行っても興味に応じて結果の文書集合が変化する方がより望ましいと言える．つまりキーワードとそれを特徴付ける要素を利用者が指定することによって，利用者はより興味のある文書を検索することができる．

そこで本稿では，特徴ベクトルを生成する要素として，文書の文字列部分から単語の頻度情報や漢字の出現割合，画像からは色のヒストグラム，メタ情報からはタイトルや著者名，レイアウト情報からは図や文字列がページを占める割合などの様々な特徴量を併用した検索をサポートすることで，より利用者に興味のある情報を抽出することができると考える[1]．

本稿では，電子文書から複数の特徴量をベクトル要素として抽出し，それらを利用者の興味に合わせた部分空間へ射影することによって，検索精度の高い検索を行う方法についての提案を行う．本研究では，現在広く使用されている電子文書である PDF 文書を対象とした実験を行う．

本研究では PDF を対象とした実験を行うが，HTML や XML など WWW 上で流通している文書に対して本研究を適用することによって，より便利な検索エンジンを作成することが可能であると思われる．

[1] 各々の特徴量を表現する特徴空間はそれぞれ独立していると考え，例えば漢字の出現割合と単語の出現頻度情報には相関関係が無いとする．

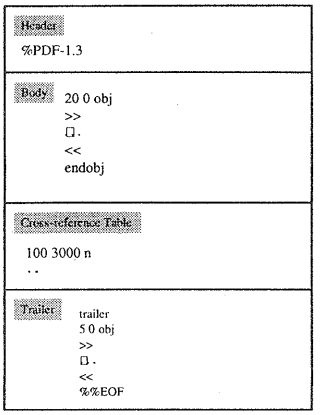

図 1: PDF の内部構造

2 基本的事項

2.1 PDF

PDF が使われる以前，電子文書フォーマットとして PostScript が使われていたが，次のような短所がありその保管や転送には不向きであった．

- 非常に大きなサイズのファイルとなることがある
- 編集ができない
- ファイル内の文字列を検索することはほぼ不可能である

そこで，それらを解決するフォーマットとして PDF が Adobe 社で考案された．PDF は画像や文字列だけでなく映像，音声を扱うこともできる．また，文書から画像のみを取り出すというように，文書の一部分のみを取り出すことが容易にできる．

こうした特徴から計算機上には大量の PDF 文書が存在しているが，それらへの検索は，文字情報のみを利用した検索のみが行われているのが現状である．

以下では PDF の内部構造について説明する．PDF は大きく分けて 2 つの部分から成っている．

- ファイル構造 (File Structure)
- 書類構造 (Document Structure)

PDF では，文書をテキスト部，画像部などの部分 (オブジェクト) に分け，それぞれの関係をファイル構造に記述する．

2.1.1 ファイル構造

PDF は，図 1 のような内部構造をもつ．

- **Header**
  %PDF-<Version Number> という記述

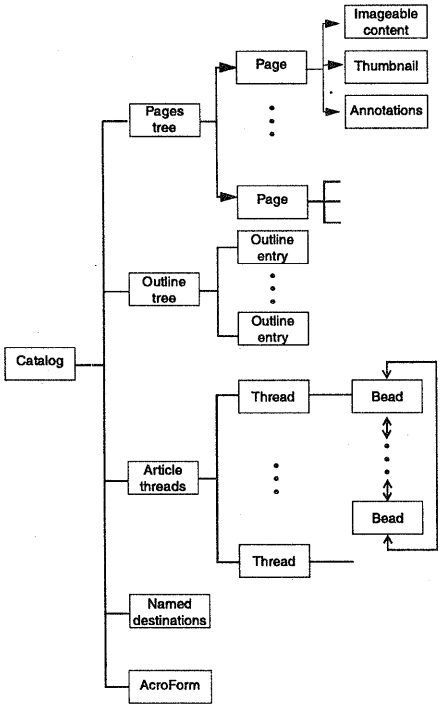

図 2: PDF のオブジェクト構造

- **Body**
  書類構造．2.1.2 節で詳しく説明する．
- **Cross-reference Table**
  オブジェクトのバイト位置情報を保持したもの．PDF ではオブジェクトを番号で示すため，番号でバイトの位置情報を参照するために用いる．
- **Trailer**
  trailer から %%EOF までの文字列．文書構造の root となるオブジェクト情報などを持っている．

2.1.2 書類構造

オブジェクトは図 2 のような木構造になっている．

Body オブジェクトの例として，図 2 の Catalog オブジェクトの記述を示す．

```
1 0 obj
<<
/Type /Catalog
/Pages 2 0 R
/Outlines 3 0 R
/PageMode /UseOutlines
>>
endobj
```

図 3: オブジェクトの内部構造

"<<" から ">>" までがオブジェクトの内容であり，1 行目はヘッダ部分である．ヘッダ部分の先頭の文字 "1" はこのオブジェクトの番号であり，次の文字 "0" はこのオブジェクトが有効であることを示す．オブジェクトの内容は / で始まるコマンド列である．例えば，/Type /Catalog は，このオブジェクトのタイプは Catalog (図 2 の root となるオブジェクト) であるという意味のコマンドである．

Imageable content には，PostScript や JPEG 形式で書かれたデータが入っているが，これらは圧縮・暗号化されている．圧縮アルゴリズムには LZW 法 [4]，暗号化アルゴリズムには 40 ビットの RC4 暗号化方式 [5] が採用されている．圧縮・暗号化されている部分はオブジェクトに含まれている文字情報，画像情報のみであり，レイアウト情報やタイトル，作成者などのメタ情報は暗号化されない．

図 2 における Page の部分ではページ内のレイアウト情報が定義されている．多くの PDF ファイルでは Catalog, Page tree, Page, Imageable content のみで構成されているため，本研究で扱う PDF 文書はこれらのオブジェクトのみを扱う．

2.2 多次元索引

本研究では，レイアウト情報に関する索引として多次元索引を用いることによりより高速に検索を行うことができると考えられる．R-tree[6] は多次元空間データベースにおける索引の方法として知られているため，本研究ではこの手法を用いる．また，論文のような形式のレイアウトではオブジェクトが相互に重なり合うことがないという特徴があることから，この特徴を利用して R-tree を改良することも考えられる．

2.3 関連研究

本研究における手法と関連したものとして，画像の内容検索における複数の特徴量を扱う方法が挙げられる [7]．画像内容検索に用いられる特徴量には，色情報，模様（テクスチャ）情報，形状情報，オブジェクトの位置情報が主に使われている．従来の画像検索ではそれらが統合されて利用されていなかったが，本研究ではそれらの特徴量を統合して検索を行う点が大きく異なる．

複数特徴部分空間を併用する方法として文献 [8]，画像データベクトルを印象語を用いて意味的な射影を行うことによって検索を行う手法が文献 [9] で既に提案されている．これらの検索では特徴部分空間が問合せ毎に同じものが使われるのに対して，本研究では問合せ毎に異なる特徴部分空間を扱う点が大きく異なる．

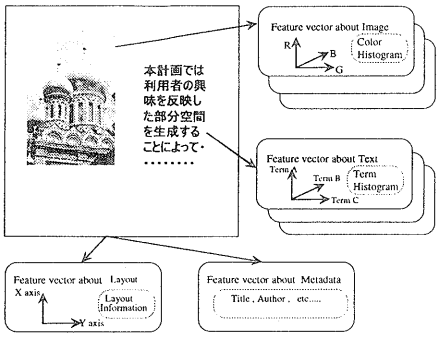

図 4: 本研究で電子文書から抽出する情報

3 電子文書の特徴量の抽出と検索への利用

本研究では，複数のメディアからなる電子文書から特徴量を抽出し，それらをベクトル表現する．本稿では PDF 文書からレイアウト情報などの特徴ベクトルを生成することによってより利用者に興味のある情報を抽出することができると考えている．PDF では文字列のみではなく映像，音声などを扱うことも可能となるため，それらを考慮した検索なども考えられるが，簡単のため本稿では文字列，静止画像のみが出現する PDF 文書のみを対象とする．以下では複数のメディアから構成された文書からの特徴量の抽出について説明する．

3.1 特徴量の抽出

本研究では電子文書から抽出する情報について図 4 に示した通り，大きく分けて四つの情報（文字情報，画像情報，レイアウト情報，メタ情報）を抽出する．

まず，文字情報から次のような特徴量を抽出することができる．

- 単語の出現頻度
- 文の長さ

単語の出現頻度情報は従来から特徴抽出として用いられてきた方法だが，本研究では，さらに文書の見た目の難しさを抽出する目的で，文の長さなどの特徴量を抽出する．また，対象言語を日本語に限定した場合更に次のような特徴量も抽出する．

- 文末（です，だ）
- 漢字，ひらがなの出現割合

次に，画像から得られる特徴量として，次のようなものを抽出することが考えられる．

- 色のヒストグラム
- 模様（テクスチャ）
- 形状

これらの特徴量は，現在行われている画像検索において用いられてきたものである．これらから，例えば「桃色が多く含まれる画像」などの検索ができる．

また，レイアウト情報の特徴量として，次のようなものを考える．

- 文字の大きさ
- 図の出現位置，ページに占める割合
- 文字の出現位置，ページに占める割合

この特徴量から,「右上に大きな図のある文書」のような検索ができる．

更に，メタ情報として次のような特徴量を抽出することができる．

- タイトル
- 著者情報
- 最終更新時刻
- 作成時刻
- 作成者
- キーワード

レイアウト情報は図 2 の Page オブジェクトに，メタ情報は Info directory オブジェクトにそれぞれ記述されているため，それらのオブジェクトの内容を抽出することによって容易に得ることができる．

本稿では，これらの特徴量からレイアウト情報として文字，画像の出現位置，また文字列情報として単語の出現頻度，更に画像情報として色のヒストグラムを用いた．

3.2 特徴部分空間と特徴量からの検索

複数の特徴量を併用する手段として，利用者の興味を持った特徴量へ射影する方法を考える必要がある．つまり，利用者の興味を表現した特徴量空間を容易に作る手法を考えなければならない．

本研究で扱う特徴量は複数あり，それぞれ

- 本質的に意味の異なる情報を持っている
- 特徴ベクトルの次元数と特徴空間上での分布状態がそれぞれの特徴量空間毎に異なっている

などの性質を持っている．本節ではこれらの情報を併用することによって評価値を求める手法について考える．

3.3 索引

文字情報と画像情報をレイアウトを用いて効率良く検索するために，レイアウト特徴ベクトルを用いた文字情報，画像情報索引を用いる．索引には B-tree や B+-tree など様々なものが存在するが，レイアウトによる索引には，2次元空間を効率良く検索することのできる R-tree[6] を用いる．なぜなら，レイアウト情報を用いてシステムに問合せを行う場合，問合せとなる領域の近傍のみを検索対象とするため検索速度が向上するからである．

3.4 評価値の計算

利用者が選択した各特徴量空間へ PDF 文書の特徴量を射影し，それらから評価値を計算する．評価値計算にベクトルの長さの総和を用いる方法などが考えられるが，利用者の要求に応じて自動的にそれぞれの特徴量空間へ重みを付けをした結果の総和を用いる方法が良いと思われる．本研究では複数の特徴量を実験で導出した係数で重み付けする方法を用いた．

4 予備実験

本実験は，検索の際に電子文書の持つ多くの特徴量に注目することによって適合率と再現率が向上することを検証するために行う．ここでは従来用いられた単語の頻度情報を用いた方法と比較し，その有効性を確認する．

4.1 実験方法

実験に使うデータには，*"ACM SIGMOD Anthology"* の Disk1 に含まれているものから 93 年度のもののみ 81 個の PDF 文書を使用した．これらの PDF 文書には，ACM SIGMOD RECORD の内容が論文ごとに分けて収録されている．以下に本実験の手順を示す．

1. 利用者の問合せを考える．例えば，「ビデオデータベースシステムに関する論文で右上に画面キャプチャの図がある論文」など．
2. 問合せに対して，予め人手で答えを用意しておく．
3. システムで 2 種類の方法で答えを求める．
   (a) 単語の頻度情報のみを用いた検索
   (b) 利用者によって選択した特徴量を用いた検索
4. 予め用意した答え集合を基にして，再現率・適合率を求める．

検索は PDF ファイル単位で行われる．つまり，オブジェクト毎に取り出された特徴量を PDF 文書毎に統合することにより比較を行う．システムの全体図については，図 5 に示す．

4.2 特徴量の抽出

特徴量を取り出す部分は次の 3 つの部分に分割できる．全ての特徴量は文書毎に作成される．以下，対象となる PDF 文書群を $\boldsymbol{D}_i : (i = 1, 2, 3, \cdots, n : n$ は総文書数) とする．また，$\boldsymbol{D}_i$ 中には j 個のオブジェクトが存在し，$\boldsymbol{x}_{ij} (j = 1, 2, \cdots m : m$ は $\boldsymbol{D}_i$ における総オブジェクト数) とし，$\boldsymbol{x}_{ij}$ の特徴ベクトルを $\boldsymbol{f}_{ij}$ とする．

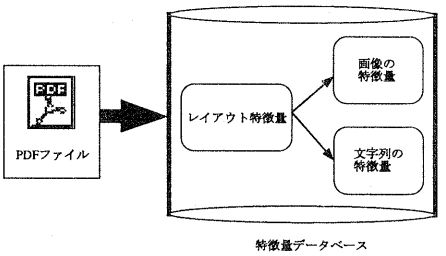

図 5: システムの構成

- **レイアウト情報の特徴量の抽出**
  オブジェクト位置情報は，画像や文書によらずオブジェクトの左上と右下の座標で与えられる．L_x,L_y, R_x,R_y はそれぞれオブジェクトの左上の x, y 座標，右下の x, y 座標とすると，レイアウトベクトル $\boldsymbol{f}_{ij}^{layout}$ は

$$\boldsymbol{f}_{ij}^{layout} = [L_x, L_y, R_x, R_y]$$

  と定義される．

- **文字情報の特徴量の抽出**
  オブジェクトの内容記述部分を取り出し，LZW 法により解凍を行うことにより文字列の内容を得た上で，次のようなベクトルを構築する．オブジェクト $\boldsymbol{x}_{ij}$ に単語 w_k 出現した回数を $t_{ijk} (k = 1, 2, \cdots, N :$ N は全文書に出現する単語の種類) とすると，

$$\boldsymbol{f}_{ij}^{term} = [\frac{t_{ij1}}{t_{ij}}, \frac{t_{ij2}}{t_{ij}}, \cdots \frac{t_{ijN}}{t_{ij}}] \quad (\text{ただし } t_{ij} = \sum_{k=1}^{N} t_{ijk})$$

  と定義される．

- **画像情報の特徴量の抽出**
  今回扱う画像は全て白黒画像であるため，画像の濃淡をヒストグラムにすることによってベクトル表現する．対象とする PDF 文書中の画像が全て 2 階調であることから，全ての画像を 2 階調画像として扱うことにする．オブジェクトに含まれる画像のヒストグラムを取ることによって特徴ベクトルを構築する．オブジェクト $\boldsymbol{x}_{ij}$ に含まれる色番号 l の画素数を r_{ijl} 個とし，

$$\boldsymbol{f}_{ij}^{pix} = [\frac{r_{ij0}}{r_{ij}}, \frac{r_{ij1}}{r_{ij}}] \ (\text{ただし } r_{ij} = \sum_{l=0}^{1} r_{ijl})$$

  と定義する．

全てのオブジェクトに対して $\boldsymbol{f}_{ij}^{layout}, \boldsymbol{f}_{ij}^{term}, \boldsymbol{f}_{ij}^{pix}$ が定義される．以上から $\boldsymbol{f}_{ij}$ は

$$\boldsymbol{f}_{ij} = [\boldsymbol{f}_{ij}^{layout}, \boldsymbol{f}_{ij}^{term}, \boldsymbol{f}_{ij}^{pix}]$$

4.3 索引

本実験では，取り扱う文書のオブジェクト数がそれほど多くないため，索引を用いることによる効率よりも計算量の方が大きくなってしまうと思われる．そこで，本実験では索引による近接探索を行わず，利用者の問合せ特徴ベクトルと文書の特徴ベクトル全てを逐次比較する．

4.4 問合せ特徴ベクトルの作成

利用者の問合せを基にして，次の手順で問合せ特徴ベクトル $\boldsymbol{q}$ を作成する．

1. 利用者により，問合せ Q を入力する．
2. Q からレイアウトに関する部分を抽出し，$\boldsymbol{q}^{layout}$ の要素とする．
3. Q からキーワードとなる単語を切り出し，$\boldsymbol{q}^{term}$ の要素とする．
4. Q から画像の濃度に対応する部分を抽出し，$\boldsymbol{q}^{pix}$ の要素とする．

4.1節で述べたように，本実験では利用者によって特徴量を選択する方法（問合せ法1）と単語の出現頻度情報のみを用いた方法（問合せ法2）の対照実験を行う．問合せ法1の場合は，「右上に黒っぽい画像のある」「データベースシステムという単語が出現している」などのように明示的に指定するものとする．この場合，2つの異なる問合せ要素（「右上に黒っぽい画像がある」と「データベースという単語が出現している」）に対して問合せを行っていることになり，$\boldsymbol{q}^{term}$ と $\boldsymbol{q}^{layout}$ の組，もしくは $\boldsymbol{q}^{pix}$ と $\boldsymbol{q}^{layout}$ の組の2組の問合せ特徴ベクトルが生成されることになる．

これに対して問合せ法2の場合，単語の出現頻度のみからなる問合せ特徴ベクトルは $\boldsymbol{q}^{term}$ のみを用いる．

本稿では，実験における問合せとしてレイアウト情報を用いなかった．対照実験を行う上で問合せ法2ではレイアウト情報の指定を行うことができず，結果に大きな差があらわれてしまうことが明らかだからである．

4.5 評価値の作成

評価値の作成は次のような手法で求める．まず，オブジェクトに対する評価値 F_{ij} を求める．

1. 問合せ特徴ベクトルのうちの1つを考える．
2. 問合せ $\boldsymbol{q}^{layout}$ が指定されていた場合，3.3節で作成した索引を用いて，$\boldsymbol{q}^{layout}$ に含まれるオブジェクトを対象とする．指定されていない場合は，全てのオブジェクトを対象とする．
3. 問合せの対象となるオブジェクトの類似度 $sim(\boldsymbol{q}^{term}, \boldsymbol{f}_{ij}^{term})$ と $sim(\boldsymbol{q}^{pix}, \boldsymbol{f}_{ij}^{pix})$ を求める．ここで，類似度は画像の特徴ベクトル，文字列の特徴ベクトルのコサイン相関値であるとする．また，文字列の類似度は tf/idf 法で重みを付けるものとし，検索語の出現頻度を t_k，検索語が出現する文書数を w_k，検索対象となる PDF の数を $|\boldsymbol{x}|$ とすると，

$$sim(\boldsymbol{q}^{term}, \boldsymbol{f}_{ij}^{term}) = \frac{\boldsymbol{q}^{term} \cdot \boldsymbol{f}_{ij}^{term}}{|\boldsymbol{q}^{term}||\boldsymbol{f}_{ij}^{term}|} \cdot \log t_k \frac{|\boldsymbol{x}|}{w_k}$$

$$sim(\boldsymbol{q}^{pix}, \boldsymbol{f}_{ij}^{pix}) = \frac{\boldsymbol{q}^{pix} \cdot \boldsymbol{f}_{ij}^{pix}}{|\boldsymbol{q}^{pix}||\boldsymbol{f}_{ij}^{pix}|}$$

4. オブジェクトの評価値を計算する．$\boldsymbol{q}_{ij}$ に対する $\boldsymbol{f}_{ij}$ の評価値を $\boldsymbol{F}_{ij}$ とすると，

$$\boldsymbol{F}_{ij} = [F_{ij}^{term}, F_{ij}^{pix}]$$

5. 他の問合せ特徴ベクトルについても同様の操作を行う．

以上のように，オブジェクトに対する評価値を求めていく．最終的に文書に対する評価値 X_j は次のように求める．

1. 文字列部分は，次のような評価値 X_j^{term} とする．

$$X_j^{term} = \sum_j F_{ij}^{term}$$

とする．つまり，tf/idf 法を用いて重みを付け，和を求める．

2. 画像部分は次のような評価値 X_i^{pix} とする．

$$X_i^{pix} = 1 - \prod_{i=1}^{n}(1 - F_{ij}^{pix})$$

つまり，類似度の高い画像が多く含まれる文書の評価値が高く，類似度の低い画像が多く含まれる文書の評価値が低くなるような評価関数を用いた．

3. 文書の特徴ベクトルは，文字列部分と画像部分の評価値を合わせたものであり，次式で与えられる．

$$X_j = \sum_j X_j^{term} + k \sum_j X_j^{pix}$$

ここで k は係数であり，実験によって求める必要がある．本実験では $k = 100$ として評価値を求めた．

4.6 評価実験の結果・考察

本研究で提案した手法が実際に有効であるかどうかを確かめるため，本研究で提案した手法と従来手法との比較実験を行った．比較対象としては，単語の出現頻度による手法を用いた．実験をするための計算機として，Sun Microsystems 社の ULTRA Enterprise 2(CPU

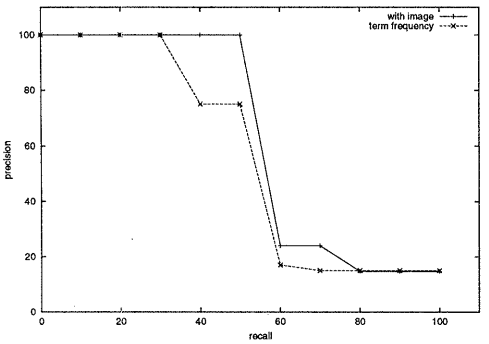

図 6: 適合率–再現率グラフ

UltraSPARC–II 200MHz x 2 , 主記憶 512MBytes) を用いた.

問合せとして「SQL に関する論文で，グラフが入っているもの」を用いた．問合せをシステムに入力する段階で 4.4 節の変換を行った結果，単語の出現頻度のみを用いたシステムへの問合せは “SQL ∧ graph” となり，本研究で提案するシステムへの問い合わせは “SQL ∧ 白っぽい画像を含む ” となった．実験結果は図 6 に示す.

単語の出現頻度を用いたものよりも画像に対する指定を行った場合のほうが適合率が上がることが確認できた．原因としては，利用者が単語以外にも画像情報などの情報を指定したことがあげられる．つまり，画像の特徴をあらわすような情報を単語で表現するよりも画像の特徴量そのものを問合せとしたほうが適合率が高くなることが示された．動画像や音声などの場合も，同様の結果が得られると推測される.

5 あとがき

本稿では，複数の特徴量から検索する手法を提案し，電子文書の検索へ応用した．本手法により電子文書から複数の特徴量を取り出すことによって複数の特徴部分空間を生成し，利用者が興味を持つ事柄に対する特徴部分空間へ電子文書の特徴量を射影することにより電子文書の検索精度が向上することが確認できた.

今後の課題を以下に示す.

1. 本実験における特徴ベクトルの扱いは，全て同等であった．だが，特徴量の分布によって，また特徴量の性質によって特徴ベクトルの扱いを変えることによって，より精度の高い検索を行うことができる．今後は特徴ベクトルの扱い方について再び考える必要があると思われる.

謝辞

本研究の一部は，文部省科学研究費基盤研究 (B)(2)「言語横断型知識発掘システムに関する研究」(課題番号: 11480088)，基盤研究 (C)(2)「開放型高機能 XML サーチエンジンに関する研究」(課題番号: 12680417) ならびに奨励研究 (A)「XML で表現されるマルチメディアデータの効果的検索法に関する研究」(課題番号: 12780309) による．ここに記して謝意を表す.

参考文献

[1] Adobe Systems Incorporated, Reading, Massachusetts. *Portable Document Format Reference Manual*, March 1999.

[2] Dave Raggett, Arnaud Le Hors, and Ian Jacobs. Html 4.01 specification. http://www.w3.org/TR/html4/, December 1999.

[3] Jeff Ayars et al. Synchronized Multimedia Integration Language (SMIL) Boston Specification. http://www.w3.org/TR/smil-boston.

[4] Uninys Corporation. *Data Compression and decompression system with immediate directory updating interleaved with string search*, January 1999. US Patent No.753871.

[5] RSA Security Incorporated. *http://www.rsasecurity.com/*.

[6] Guttman A. R-tree: A dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD*, pages 44 – 57, 1984.

[7] 串間 和彦, 赤間 浩樹, 紺谷 精一, and 山室 雅司. 色や形状等の表層的特長量にもとづく画像内容検索技術. 情報処理学会論文誌, 40(SIG3(TOD1)):171 – 184, February 1999.

[8] 島田 恭宏, 大倉 充, 塩野 充, and 橋本 禮治. 複数特徴部分空間法による手書き類似漢字識別. 電子情報通信学会論文誌 D–II, J78–D–II(10):1460 – 1468, October 1995.

[9] Yasushi Kiyoki, Takashi Kitagawa, and Takanari Hayama. A metadatabase system for semantic image search by a mathematical model of meaning. In *ACM SIGMOD RECORD*, volume 23 of *4*, December 1994.