B2-3

既知電子文書の検索における特徴量の組合せ手法

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
‡ 国立情報学研究所 ソフトウェア研究系

† 〒 630-0101 奈良県生駒市高山町 8916-5
‡ 〒 101-8430 東京都千代田区一ツ橋 2-1-2

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

あらまし　現在用いられている電子文書は，テキストのみではなく画像などで構成されていることが多い．よって，これら複数のメディアを考慮した電子文書を検索するシステムは必要である．本稿では，複数のメディアで構成された電子文書から複数の特徴量を抽出し，それらを一つに統合して文書の評価値を求める方法を提案する．我々のシステムでは，問合せと各メディアとの類似度をそれぞれ求め，それらを一つに統合して文書の評価値を求める．本研究では，複数の評価値から文書の評価値へ統合する際に二つの問題を考慮した．まず，問合せとメディアとの類似度を正規化するための方法として偏差値を用いた．さらに複数の類似度を統合する際に用いるために最適な評価関数を求めた．その結果，偏差値を用いた手法が有効であったこと，*p*-norm と T-operators のうちの一つの評価関数が良い適合率を示すことが分かった．
キーワード　情報検索，複数のメディア，特徴量の抽出，評価関数，偏差値

Normalization and Integration of Heterogeneous Similarity Values for Multimedia Document Retrieval

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

† Graduate School of Information Science, Nara Institute of Science and Technology
‡ Software Research Division, National Institute of Informatics

†8916-5 Takayama, Ikoma, Nara 630-0101, Japan
‡ 2-1-2 Hitotsubashi, Chiyoda, Tokyo 104-8430, Japan

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

Abstract　Current electronic documents often contain not only text data but also image data. Hence, we need IR systems which take into account the features of different media types. In this paper, we propose an IR system which integrates multiple features of multimedia documents into one feature. In our IR system, similarity values between each medium of a document and query are integrated into a document score. When the similarity values between each medium and query are integrated into document score, we have to consider two issues; one is the normalization of each similarity values, the other is the selection of mathematical function to integrate similarity values. Instead of raw similarity values of each medium, we define the “Deviation Value” of the similarity values to normalize the significance of each medium. Moreover, we performed extensive experiments to find an appropriate mathematical function used to integrate similarity values of each medium into the document score. As a result, we found that the deviation value is useful to normalize multiple features calculated by different schemes. We also found that p-norm and one of T-operators are the best mathematical functions in our experiments.

Key words　Information Retrieval, Multimedia Data, Appearance, Contents

1　はじめに

現在用いられている電子文書はテキスト情報のみではなく画像など複数のメディアで構成されている．利用者が以前見たことのある電子文書を検索する場合，キーワードを Boolean 検索するだけではなく，キーワードのレイアウト情報や画像情報などから検索することは有効である．我々の以前の研究では，テキストと画像を含む電子文書を，利用者のレイアウトに関する記憶を用いて検索する手法を提案した [1]．ところが，利用者は電子文書における記憶としてレイアウトのみではなく内容を記憶していることも多く，それらを利用した検索を行うことは必要である．

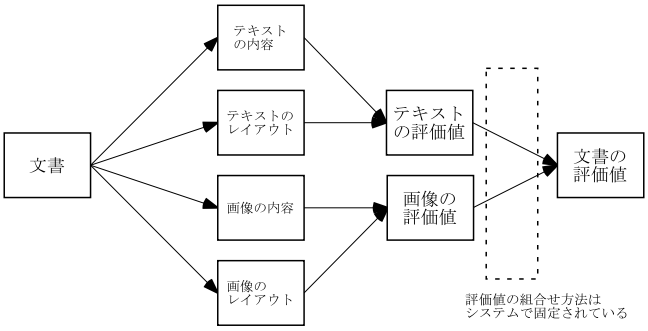

図 1: 本稿での提案手法

本稿では，文書からテキストと画像のレイアウト，内容に関する特徴量を取り出し，利用者が検索する際の手がかりを増やすことによって文書検索の際の適合率の改善を図る．まず，各メディアごとに特徴量を抽出し，それぞれのメディアの評価値を算出する．そして，それらを問合せの記述に基づいて統合を行い，文書の評価値を算出する．

本稿で提案している手法と以前の研究で提案されている手法 [1] との大きな違いは，複数の評価値の統合の手法にある．以前の手法では，4 つの評価関数を用いて統合していた．一方，拡張ブーリアン検索やファジィセットモデルなどにおいて多くの統合方法が提案されており，以前に用いられていた 4 つの関数を含む 29 種類の評価関数が存在することが知られている [3]．本研究では 29 種類の評価関数を比較することによって，より最適な統合方法を発見することを目的としている．

また，問合せと各メディアとの類似度はメディアの種類ごとに異なる方法を用いなければならない．ところが，これらの類似度を単純に関数を用いて統合を行った場合，一つのメディアの類似度のみが文書の類似度のみに反映されてしまうといった問題点がある．例えば，テキスト部分の問合せとメディアとの類似度が 0.5 となる場合と画像部分の類似度が 0.5 となる場合では，同じ数値であっても各メディアが全体としてどのように分布しているかという点を考慮した場合に同じ類似度であるとはいえないと考えられる．このようなメディア間の類似度の分布の相違を解消するために，本研究では偏差値を用いて問合せとメディアとの類似度の正規化を行った．

2　基本的事項

2.1　特徴量の抽出

複数のメディアで構成されている電子文書を，テキストや画像の情報を用いて検索するための手法を提案するためには，まず電子文書に含まれている各メディアの特徴量を抽出しなければならない．以下，各特徴量の表現方法について述べる．

検索される電子文書群 DB は，電子文書 D_i $(i = 1, 2, \cdots, N)$ の集合である．電子文書 D_i はテキストと画像の 2 つのメディアで構成されているため，まず各メディアに分割する．また，文書に含まれるテキストは単語に分割し，文書に含まれる複数の画像は各々の画像に分割する．文書 D_i に含まれる各単語を d_{ij}^T $(j = 1, 2, \cdots, J_i)$，各画像を d_{ik}^I $(k = 1, 2, \cdots, K_i)$ とする．以下，各々の単語，画像をオブジェクトと呼び，それらから特徴量を抽出する方法について述べる．

2.1.1　テキストと画像のレイアウト

PDF や SMIL など，現在用いられている多くの電子文書ではオブジェクトのレイアウトが指定されている．本研究ではオブジェクトのレイアウトを特徴量として抽出する．これらの特徴量は，文献 [1] で定義されているベクトルと同じであり，これを $\boldsymbol{f}^L(d)$ と表現する．レイアウトの特徴量はオブジェクトが占めている領域の右上，左下の座標とする．つまり，あるオブジェクト d $\{d|d_{ij}^T, d_{ij}^I\}$ の占めている領域をの右上の座標を $[x^{left}(d), y^{left}(d)]$，左下の座標を $[x^{right}(d), y^{right}(d)]$ とすると，オブジェクト d の特徴量 $\boldsymbol{f}^L(d)$ は次のようなベクトルとして表現できる．

$$\boldsymbol{f}^L(d) = [x^{left}(d), y^{left}(d), x^{right}(d), y^{right}(d)]$$

2.1.2　テキストの内容

文書には多くの単語があるが，各単語の出現頻度情報から文書の特徴量を抽出できる．本研究では，ベクトル

空間モデルによる検索を行った．まず，検索対象となる文書に含まれる単語を重複なく抜き出し，テキストベクトルの基底 W とする．

$$W = \{w_1, w_2, \cdots\}$$

特徴ベクトルの要素として，本研究では tf/idf 法 [2] で重み付けされた単語の出現頻度情報を用いる．検索対象の文書集合 DB に含まれる文書数を N，単語 w_k を含む文書数を $df(w_k)$，文書 D_i 中の単語 w_k の出現回数を $tf(D_i, w_k)$ とすると，文書 D_i のテキストの内容を表す特徴ベクトル $\boldsymbol{f}^{TC}(D_i)$ は次のように定義される．

$$\boldsymbol{f}^{TC}(D_i) = [f(D_i, w_1), f(D_i, w_2), \cdots]$$

ただし，各要素は次の式によって計算される．

$$f(D_i, w_k) = \frac{tf(D_i, w_k) \cdot \log\left(\frac{N}{df(w_k)}\right)}{\sum_{k=1}^{n} tf(D_i, w_k)}$$

つまり，特徴ベクトルは，tf/idf 法 [2] で重み付けされた単語の出現頻度を用いる．

2.1.3　画像の内容

画像オブジェクトからは，色のヒストグラム情報や模様情報など様々な特徴量を得ることができるが，本研究では色のヒストグラム情報を利用した．そこで，画像オブジェクトの内容をあらわす特徴ベクトル $\boldsymbol{f}^{IC}(d)$ の各要素を，各色の画像全体を占める画素数として定義した．画像オブジェクト d 中の色番号が g ($g = 0, 1, \cdots, g_{max}$：画像の色数) である画素の割合を $c^g(d^I_{ij})$ とおき，画像オブジェクト d の内容をあらわす特徴ベクトル $\boldsymbol{f}^{IC}(d)$ を次のように定義する．

$$\boldsymbol{f}^{IC}(d) = [c^1(d), c^2(d) \cdots]$$

2.2　問合せと類似度

2.2.1　利用者の問合せ

本研究では，次式で定義された問合せを用いて検索を行う．*Layout* を問合せとなる領域，*Term* を問合せ単語，*Image* を問合せ画像とすると，利用者の問合せ *Query* は次のように定義される．

Query ::= (**term** *Term* **on** *Layout*)
| (**image** *Image* **on** *Layout*)
| *Query* , *Query*

ここで，上 2 つの指定はレイアウトを用いた問合せであり，利用者が記憶している文書の見た目に関する情報を問合せとして記述する部分である．

2.2.2　問合せの拡張

前節では，利用者がどのようにシステムへ問合せを行うかについて述べた．本節では，利用者が入力した問合せを処理する方法について述べる．

オブジェクトのレイアウトへの問合せは次のように処理される．まず，テキストや画像のレイアウトに関する問合せのベクトル表現を $\boldsymbol{Q}^L(q)$，単語の内容に関する問合せのベクトル表現を $\boldsymbol{Q}^{TC}(q)$，画像の内容に関する問合せのベクトル表現を $\boldsymbol{Q}^{IC}(q)$ とする．ここで，q は問合せオブジェクトであり，利用者の問合せに含まれる複数の条件のうちの 1 つである．問合せのベクトル表現は，ほぼ文書の特徴量の表現と同一のものを用いる．

問合せ q の指定領域をの右上の座標を $[x^{left}(d), y^{left}(d)]$，左下の座標を $[x^{right}(d), y^{right}(d)]$ とすると，$\boldsymbol{Q}^L(q)$ は次のようなベクトルとして表現される．

$$\boldsymbol{Q}^L(q) = [x^{left}(q), y^{left}(q), x^{right}(q), y^{right}(q)]$$

テキストの内容に関する問合せ q のベクトルを次のように表現する．

$$\boldsymbol{Q}^{TC}(q) = [f'(q, w_1), f'(q, w_2), \cdots]$$

ただし，$f'(q, w_k)$ は問合せ q で指定した単語と w_k が一致するならば 1 なければ 0 となるような関数である．

画像の内容に関する問合せ q のベクトルを次のように表現する．

$$\boldsymbol{f}^{IC}(d^I_{ij}) = [c^1(d^I_{ij}), c^2(d^I_{ij}) \cdots]$$

ただし，問合せ q での指定における色番号が g ($g = 0, 1, \cdots, g_{max}$：画像の色数) である画素の割合を $c^g(d^I_{ij})$ とする．

2.2.3　各メディアの評価値の算出

各メディアの評価値は，各々異なる手法によって計算される．また，文書の特徴量は異なったレベルで抽出されている．つまり，レイアウトや画像の内容の特徴量はテキスト，画像などのオブジェクト 1 つに対して 1 つ抽出されているのに対して，テキストの内容の特徴量は 1 つの文書ごとに 1 つ抽出されている．そこで，本研究では特徴量が取り出されたレベルに応じた評価値を算出する．

レイアウトの類似度 $S^L(q,d)$ を定義する．まず問合せ領域 $\boldsymbol{Q}^L(q)$ とオブジェクト領域 $\boldsymbol{f}^L(d)$ の面積をそれぞれ $|\boldsymbol{Q}^L(q)|$，$|\boldsymbol{f}^L(d)|$，問合せ領域とオブジェクト領域が重なっている領域を $|\boldsymbol{Q}^L(q) \cap \boldsymbol{f}^L(d)|$ とすると，レイアウトの類似度を次のように定義する．

$$S^L(q,d) = \frac{|\boldsymbol{Q}^L(q) \cap \boldsymbol{f}^L(d)|}{|\boldsymbol{Q}^L(q)|}$$

次に，テキストの内容の類似度 $S^{TC}(q,D_i)$，画像の内容の類似度 $S^{IC(q,d)}$ を次のように定義する．

$$S^{TC}(q,D_i) = \boldsymbol{Q}^{TC}(q) \cdot \boldsymbol{f}^{TC}(D_i)$$

$$S^{IC}(q,d) = \boldsymbol{Q}^{IC}(q) \cdot \boldsymbol{f}^{IC}(d)$$

3 特徴量の組合せ手法

我々の目的は，異なるレベルで算出された複数の類似度を統合し，文書の評価値を求めることである．ここで考えなければならない点はテキストの内容の類似度は文書 1 つにつき 1 つしか求められないのに対してその他の類似度はオブジェクトの数だけ存在することである．つまり，もし単純な相加平均を用いてそれらの類似度を統合すると，一番数の多い類似度であるテキストのレイアウトに依存した結果が文書の評価値となってしまう．現在，複数の数値を統合するためには相加平均や乗算などが多く用いられているが，複数のメディアで構成された文書の検索精度を向上させるためには複数の類似度が均等に文書の評価値に反映される必要があると考えられる．

本研究ではまず，文書ごとにテキスト，画像のレイアウトと内容という 4 つの類似度を求めておき，それらを評価関数を用いて統合するという手法をとった．つまり統合の段階が 2 つあり，それぞれの段階について効果があると考えられる関数を求める必要がある．

本研究では評価関数として文献 [3] で提案されている T-operator，平均関数，p-norm の 3 つの種類の合計 29 種類の評価関数を用意した．それぞれの関数を表 1, 2, 3 に示す．以下，これらの関数の任意の 1 つを $\uplus$，これらの関数のうち and を表す関数を $\otimes$，or を表す関数 $\oplus$ と表記する．ただし，平均関数のうち A_1，A_2，A_3 は and と or の区別が無いため，同じ関数を用いる．

3.1 各メディアの評価値

文書の評価値を求めるために，各メディアの評価値を求めなければならない．本研究では各問合せ q ごとに 1 つの評価値を求める．テキスト，画像のレイアウト情報

表 1: T-operators.

	AND	OR
T_1	$\min(x,y)$	$\max(x,y)$
T_2	$x \cdot y$	$x+y-xy$
T_3	$\max(x+y-1,0)$	$\min(x+y,1)$
T_4	$\frac{xy}{x+y-xy}$	$\frac{x+y-2xy}{1-xy}$
T_5	$\begin{cases} x & \text{if } y=1 \\ y & \text{if } x=1 \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} x & \text{if } y=0 \\ y & \text{if } x=0 \\ 1 & \text{otherwise} \end{cases}$
T_6	$\frac{\lambda xy}{1-(1-\lambda)(x+y-xy)}$	$\frac{\lambda(x+y)+xy(1-2\lambda)}{\lambda+xy(1-\lambda)}$
T_7	$\max\left(1-((1-x)^p+(1-y)^p)^{1/p},0\right)$	$\min\left((x^p+y^p)^{1/p},1\right)$
T_8	$\frac{1}{1+\left((\frac{1}{x}-1)^\lambda+(\frac{1}{y}-1)^\lambda\right)^{1/\lambda}}$	$\frac{1}{1+\left((\frac{1}{x}-1)^{1-\lambda}+(\frac{1}{y}-1)^{1-\lambda}\right)^{-1/\lambda}}$
T_9	$\frac{xy}{\max(x,y,\lambda)}$	$1-\frac{(1-x)(1-y)}{\max(1-x,1-y,\lambda)}$
T_{10}	$\max((1+\lambda)(x+y-1)^\lambda xy,0)$	$\min(x+y+\lambda xy,1)$

表 2: Averaging Functions.

A_1	$(1-(1-x)\cdot(1-y))^\gamma \cdot (x\cdot y)^{1-\gamma}$
A_2	$\gamma \cdot \max(x\cdot y) + (1-\gamma)\cdot \min(x\cdot y)$
A_3	$\gamma \cdot (1-(1-x)\cdot(1-y))(1-\gamma)\cdot(x\cdot y)$
$A_4(AND)$	$\gamma \cdot \min(x\cdot y) + \frac{(1-\gamma)(x+y)}{n}$
$A_4(OR)$	$\gamma \cdot \max(x\cdot y) + \frac{(1-\gamma)(x+y)}{n}$

表 3: Paice and p-norms.

Paice (AND)	$\frac{x+ry}{1+r}$,
Paice (OR)	$\frac{x+ry}{1+r}$,
p-norm (AND)	$1-\left(\frac{(1-x)^p+(1-y)^p}{2}\right)^{1/p}$
p-norm (OR)	$\left(\frac{x^p+y^p}{2}\right)^{1/p}$

による評価値 $S^{TL}(q,D_i)$，$S^{IL}(q,D_i)$ は次の式で計算を行う．

$$S^{TL}(q,D_i) = \sum_d S^{TL}(q,d)$$

$$S^{IL}(q,D_i) = \left(1-\prod_d (1-S^{IL}(q,d))\right)$$

3.2 文書の評価値

3.2.1 偏差値を用いた特徴量の統合

統合する 2 つの評価値は，それぞれの値の範囲が評価値の種類によって異なり，もし 0.5 という値がテキストのレイアウト情報とテキストの内容の評価値に付いていたとしてもそれらの意味合いは異なる．このような評価値を合成すると，平均的に高い値をとる評価値の影響が大きくなり，評価値の種類による格差が生じる．この問題は，たとえ 0 から 1 までの数値へ評価値を正規化したとしても解決しない問題である．そこで，本研究ではそれぞれの特徴量の偏差値を用いて統合することにより，特徴量間の格差を解消した．

偏差値を用いた評価値は次式で表現される．

$$S'^{TL}(D_i) = \frac{(S^{TL}(D_i) - \bar{S}^{TL}(DB))}{\sigma^{TL}} \cdot 10 + 50$$
$$S'^{IL}(D_i) = \frac{(S^{IL}(D_i) - \bar{S}^{IL}(DB))}{\sigma^{IL}} \cdot 10 + 50$$

ここで，$\bar{S}^{TL}(DB)$，$\bar{S}^{IL}(DB)$ はテキスト，画像の評価値の平均であり，σ^{TL}，σ^{IL} はテキスト，画像の評価値の標準偏差である．

3.2.2 評価関数

最後に，前節までで述べた 4 つの種類の評価値を統合する．本システムで考えられている問合せは and や or で結合されており，結合された接続詞によって評価関数を変更することによって文書の評価値を求める．

利用者の問合せが “q_1 , q_2” （q_1, q_2 は問合せオブジェクト）で示されている場合，次の式で文書の評価値 $S(D_i)$ を求める．

$$S(D_i) = S(q_1, D_i) \otimes S(q_2, D_i)$$

ただし，$S(q, D_i)$ は問合せ q と文書 D_i の評価値であり，問合せの種類によって $S^{TL}(q, D_i)$，$S^{IL}(q, D_i)$ のいずれかである．

本研究では，評価関数として表記した $\otimes$ の部分に合計 29 種類の評価関数を比較して用いることによって，実験によって最も精度の良い評価関数を求めることが目的である．

4 実験

第 3 章においては，主に以下の 2 点の提案を行った．

- 評価関数の比較
  複数の評価値を統合する方法として，拡張ブーリアン検索モデル，ファジィセットモデルなどの手法が提案されている．本研究ではこれらで用いられている 29 種類の評価関数を比較し，検索性能の高い評価関数を求めた．
- 正規化の手段としての偏差値の利用
  複数の異なる手段で計算されている評価値を統合する際に，それらを同等に扱うためには正規化を行うことが必要である．正規化の手法には最大値で除算を行う方法が主に用いられているが，本研究では偏差値を用いる方法を提案した．

以下，実験では評価関数の比較を行い，検索性能の高い評価関数を求めること，正規化の手段として偏差値を用いる方法が実際に有効であったことの 2 つを示す．

4.1 実験方法

本研究で提案した手法が有効であることを確かめるために，手法を実際に実装し評価関数の比較を行う．実験で扱うデータとして，“2000 Digital Symposium Collection” に含まれる 351 個の PDF 文書を対象とする．以下に実験の手順を示す．

1. あらかじめ問合せとそれらに対する解答集合を用意しておく．
2. あらかじめ用意しておいた問合せをシステムに入力する．
3. システムから出力された結果とあらかじめ用意した解答集合から，再現率-適合率グラフを求める．
4. 再現率-適合率グラフから平均適合率を求め，評価関数の性能評価の指標とする．

利用者の問合せとして次の 3 種類の問合せを用意した．

- `Query (a)`:**テキスト + レイアウト，画像**
  単語 “multimedia” が左上にあり，画面キャプチャ画像がある文書
- `Query (b)`:**テキスト，画像 + レイアウト**
  単語 “multimedia” があり，画面キャプチャ画像が右上にある文書
- `Query (c)`:**テキスト + レイアウト，画像 + レイアウト**
  単語 “multimedia” が左上にあり，画面キャプチャ画像が右上にある文書

実験を行う前に，これらの問合せに対する解答集合をあらかじめ作成した．`Query (a)`, `(b)`, `(c)` に対応する正解集合文書の個数はそれぞれ 18, 10, 2 となった．

本システムに問合せを入力するために，次のような問合せ拡張を行った．

- `Query (a)` **term** multimedia **on** upper-left **,** **image** capture **on** anywhere
- `Query (b)` **term** multimedia **on** anywhere **,** **image** capture **on** upper-right

表 4: 評価関数のパラメータの値

T_{6AND}	1.5	A_1	0.5
T_{6OR}	1.5	A_2	0.4
T_{7AND}	13	A_3	0.1
T_{7OR}	13	A_{4AND}	0.1
T_{8AND}	0.8	A_{4AND}	0.1
T_{8OR}	0.8	$Paice_{AND}$	1.0
T_{9AND}	1.0	$Paice_{OR}$	1.0
T_{9OR}	1.0	$Pnorm_{AND}$	2.0
T_{10AND}	-1.0	$Pnorm_{OR}$	2.0
T_{10OR}	-1.0		

- `Query (c)` **term** multimedia **on** upper-left **,** **image** capture **on** upper-right

ここで，位置の指定の部分で “upper-left”, “upper-right”, “anywhere” と記述された部分についてはシステム上で座標指定を行った．また，画像の指定である “capture” についてもシステム上で実際の画素割合として計算した．つまり，検索対象文書中に存在する画像のうち全ての画像キャプチャ部分を取り出し，それぞれについて画素の割合を求め，それらの平均を求めた結果を画面キャプチャ画像の画素割合として計算した．

評価関数にはパラメータを与える必要があるものがある．本研究では，Lee の実験によって得られた，拡張ブーリアンモデルにおいて最適な値 [3] を用いた．実際の値は表 4 に示す．

4.2 実験結果

実験を行った結果得られた再現率-適合率グラフを図 2 に示す．ところが，これらのグラフでは再現率-適合率曲線が交差している部分が多く，このグラフを用いた評価関数の比較を行うことは困難である．そこで，それぞれの評価関数の平均適合率を求め比較を行った．平均適合率による評価関数の比較を図 5 に示す．

これらの結果から，平均適合率の高い評価関数は T_8(*and*) と p-norm (*and*) であることが分かった．拡張ブーリアンモデルにおいて p-norm が有効であることが Lee [3] によって示されているが，拡張ブーリアンモデルにおける評価値の統合の場合のような同じ計算方法で計算されている評価値を統合する場合のみではなく，本研究における手法のように異なる計算方法で計算されている評価値

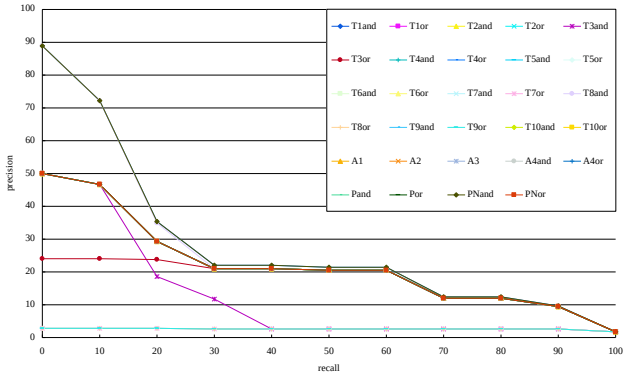

図 2: 29 種類の評価関数における再現率 – 適合率グラフ

表 5: 平均適合率による評価関数の比較

	重み付き		重みなし	
	and	or	and	or
T_1	22.19	22.19	11.11	11.11
T_2	22.19	22.19	18.79	5.644
T_3	13.12	17.26	11.11	10.10
T_4	22.19	29.05	5.461	10.29
T_5	2.582	2.582	2.582	10.10
T_6	22.19	22.19	2.592	10.29
T_7	22.19	2.582	11.11	2.582
T_8	28.55	2.584	21.29	2.592
T_9	22.19	2.592	23.96	18.88
T_{10}	22.19	22.19	18.79	5.468
A_1	22.19		2.593	
A_1	22.19		11.11	
A_1	22.19		23.21	
A_4	22.19	22.19	11.11	11.11
Paice	22.19	22.19	11.11	11.11
p-norm	29.05	22.19	2.692	11.11

を統合する場合にも p-norm による評価値の統合が有効な計算方法であることが分かった．また，多くの場面で用いられている統合方法として相加平均があるが，実験結果から p-norm などの検索精度と比較して適合率が低下することが分かった．

偏差値を用いることにより検索精度が上昇していることを確認するために，偏差値を用いた場合と用いない場合における検索精度の変化を調べた．偏差値を用いた正規化を行わずに評価値の統合を行い，平均適合精度を求めた結果を表 5 の “重みなし” の部分に示す．また，偏差値を用いた場合と用いなかった場合の適合率の変化を図 3 に示す．

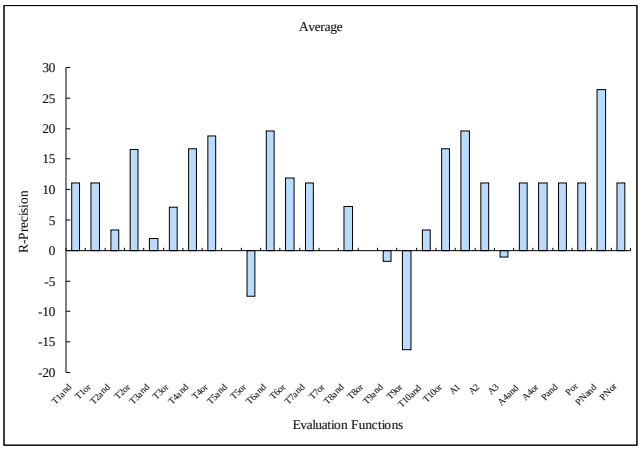

図 3: 偏差値を用いる場合と用いない場合の平均適合率の比較

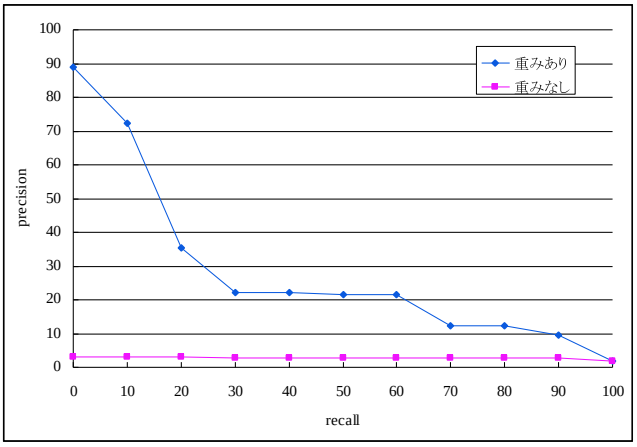

図 4: *p*-norm において偏差値を用いる場合と用いない場合の再現率-適合率グラフ

図 3 は偏差値を用いた場合の平均適合率と偏差値を用いなかった場合の平均適合率の差を表しており，差が負の値の場合は偏差値を用いなかったほうが高い適合率であったことを示している．これらの結果からは，偏差値を用いることによって適合率が上昇した関数が多いことを示している．ところが，全ての関数で適合率が上昇したわけではなく，適合率が低下したものもあった．つまり，T_{6AND} や T_{10AND} など，偏差値を用いることによって適合率が低下する場合があることが分かった．

また，図 4 は，偏差値を用いた場合に最も良かった関数である *p*-norm に注目した再現率-適合率グラフである．これらのグラフから，偏差値を用いることによって適合率が上昇することが実際に確かめられた．

これらの結果から，多くの場合に偏差値を用いる方法が有効であることが分かった．ところが，T_{6AND} や T_{10AND} など，偏差値を用いることによって適合率が低下する場合があることが分かった．

5　関連研究

従来の情報検索では，画像や映像など一つのメディアから多くの情報を抽出し，それを検索に利用することで検索精度を高める手法が主流であった．特に，画像の類似検索の研究では複数の種類の特徴量として色情報やテクスチャ情報（模様情報），形状情報などを組み合わせて検索する方法が数多くある [4]．これらの研究は，検索対象が一つのメディアに限定されているため，複数のメディアを検索対象としている本研究とは，扱うメディアの種類の点で異なっている．また，Web 上の画像を検索する際に Web 文書に含まれるテキスト情報を用いて検索を行う手法がある [5]．この手法では，まずテキストによる問合せでいくつかの画像の候補を検索し，それらの画像に類似した画像を画像検索手法で検索する方法である．この手法は画像の検索を容易に行うための手法であるが，利用者が一度画像の候補を絞る必要がある点，レイアウトによる問合せを行うことが非常に困難である点が本研究と異なる点である．

既存の文書からテキストだけではなく画像などを取り出し，検索に応用しようとしている研究がある [6]．これらの研究と本研究との相違点として，これらの研究ではレイアウトやテキストを扱う特別な手法であり，音声や映像などを扱うにはまた別の手法が必要であるのに対して，本研究で示した手法はそのまま他のメディアに適用できる点が挙げられる．

また，文字列や画像のレイアウトを検索の手がかりとする研究も行われている．例えば World Wide Web (WWW) で用いられている HTML 文書に含まれるテキストや画像を SQL を用いて検索する WebSSQL [7] では，SQL では定義されていなかった画像やレイアウトの問合せ方法を定義することで，問合せの表現方法の拡張を行っている．しかし，WebSSQL は SQL の改良であり，問合せ条件に適合した文書を全て取り出すのに対して，本研究では適合する文書をランク付けして表示する点が異なる．さらに，複数のメディアの評価値を統合するための評価関数は，WebSSQL では最小値しか用いられていない．評価関数は検索精度に影響を与えるものであり，WebSSQL のように一意に定めるのは困難である．また，WebSSQL は HTML 文書を対象としているため，レイアウトが関

覧環境によって異なるといった問題もある．本研究ではこれらの問題点を考慮して，閲覧環境に依存しない電子文書形式である PDF を用いて実験を行い，29 種類の評価関数を用いて実験を行っている．

6 おわりに

本稿では，複数のメディアで構成された電子文書から複数の特徴量を抽出し，問合せと特徴量の類似度を求め，偏差値による正規化を行った後に評価関数による統合を行うことによって，電子文書の評価値を求める方法を提案した．また，評価関数として 29 種類の関数を比較し，検索精度が高いと考えられる関数を評価実験によって求めた．さらに偏差値による正規化が有効であることを実験によって示した．

ただ，評価関数の精度は利用者による正解集合の選択によって異なると考えられる．つまり，利用者がどのような目的で検索を行うのかによって最適な評価関数は変化するものと考えられる．また，偏差値による正規化についても行ったほうが良いかどうか利用者によって変化するものと考えられる．そこで，利用者はどのような文書を適合文書として判断するか，といった分析を行う必要があると考えられる．同様に，評価関数がどのような文書を適合文書として判断しやすいかについて分析する必要があると考えられる．今後は，これらの分析を基に評価関数をシステムによって動的に決定することによって適合率を上昇させる方法について考える．

謝辞 本研究の一部は，文部科学省科学技術研究費基盤研究（課題番号：11480088，12680417，12780309），ならびに科学技術振興事業団（JST）の戦略的基礎研究推進事業（CREST）「高度メディア社会の生活情報技術」プログラムの支援によるものである．ここに記して感謝を表す．

参考文献

[1] 鈴木優, 波多野賢治, 吉川正俊, 植村俊亮．複数のメディアで構成された電子文書の検索手法. 情報処理学会論文誌:データベース, Vol. 42, No. SIG10(TOD 11), Oct. 2001.

[2] G. Salton. *Automatic Text Processing: The Transformational, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1988.

[3] Joon Ho Lee. Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval. Technical Report TR95-1501, Cornell University, 1995.

[4] 串間和彦, 赤間浩樹, 紺谷精一, 山室雅司. 色や形状等の表層的特徴量にもとづく画像内容検索技術. 情報処理学会論文誌, Vol. 40, No. SIG3(TOD1), pp. 171 – 184, February 1999.

[5] Marco La Cascia, Saratendu Sethi, and Stan Sclaroff. Combining Texual and Visual Cues for Content-based Image Retrieval on the World Wide Web. In *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 24 – 28, 1998.

[6] H Fujisawa, Y Shima, M Koga, and T Murakami. Automatically Organizing Document Bases Using Document Understanding Techniques. In *Proceedings of the 2nd Far-East Workshop on Future Database systems*, pp. 244 – 253, 1992.

[7] C. Zhang, W. Meng, Z. Zhang, and Z. Wu. WebSSQL – A Query Language for Multimedia Web Documents. In *Proceedings IEEE Advances in Digital Libraries 2000 (ADL2000)*, pp. 58 – 67, 2000.

[8] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.