# 複数のメディアで構成された電子文書の検索における評価値の統合手法

6W－8

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

## 1 はじめに

近年，電子文書はテキストデータだけでなく，画像，映像など複数のメディアで構成されていることが多くなった．我々の以前の研究 [1] では電子文書を構成している各メディアと問合せとの類似度を統合することによって，複数のメディアで構成された電子文書を検索する手法の提案を行った．しかし，この手法ではテキスト部分の類似度と画像部分の類似度の数値の分布には差があるため，平等な比較を行っているとはいえず，検索性能低下の一つの要因となっていた．本稿では各類似度に対し重み付けを行うことでこの問題点を克服する手法を提案し，実験によって検索性能が向上したことを示す．

## 2 評価関数を用いた評価値の統合手法

我々の以前の研究では，テキストと画像が混在している電子文書 $D_i$ と問合せ $Q$ の評価値を求める際に，テキスト部分とそのレイアウトの評価値 $X_i^{term}$ $(0 \leq X_i^{term} \leq 1)$ と，画像部分とそのレイアウトの評価値 $X_i^{image}$ $(0 \leq X_i^{image} \leq 1)$ という 2 つの評価値を求め，それらを評価関数を用いて統合するという手法をとっていた．

### 2.1 評価関数の選定とその特徴

本手法では，平均，p-norm，確率という 3 つの考え方を基にして選ばれた以下の 6 つの関数を，テキストの評価値と画像の評価値を合成するために用いている．

- 相加平均

$$X_i = \frac{X_i^{term} + X_i^{image}}{2} \tag{1}$$

- 相乗平均

$$X_i = \sqrt{X_i^{term} \cdot X_i^{image}} \tag{2}$$

**An Integrated Method of Scores Calculated with Multimedia Documents**

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

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

**‡Software Research Division, National Institute of Informatics**

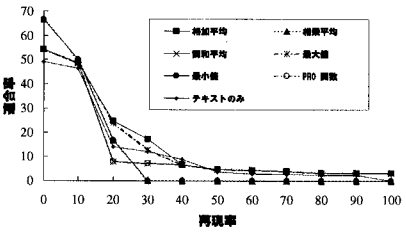


図 1: 各評価関数を用いた場合の再現率 - 適合率グラフ

- 調和平均

$$X_i = \frac{2}{\frac{1}{X_i^{term}} + \frac{1}{X_i^{image}}} \tag{3}$$

- 最大値

$$X_i = \max\{X_i^{term}, X_i^{image}\} \tag{4}$$

- 最小値

$$X_i = \min\{X_i^{term}, X_i^{image}\} \tag{5}$$

- PRO 関数

$$X_i = 1 - ((1 - X_i^{term}) \cdot (1 - X_i^{image})) \tag{6}$$

### 2.2 評価実験

実験で扱うデータとして “ACM SIGMOD Digital Symposium Collection Volume 1” に含まれる PDF 文書を用いた．文書数は 351 個である．3 つの異なる問合せを行い，再現率 – 適合率グラフ描き，平均したものを図 1 に示す．

実験の結果，相加平均を評価関数に用いた場合に適合率が一番高いことが示された．これに対し，相乗平均や調和平均を用いた場合，相加平均を用いた場合と比較して適合率が低下した．これは，テキストの評価値が 0 で画像の評価値が高い場合，相乗平均を用いた文書の評価値は 0，調和平均を用いた文書の評価値は非常に小さな値となってしまい，正解文書が検索されないからである．同様の理由で最小値は非常に適合率が低い．つまり，各メディアの評価値全てがある程度文書の評価値に反映されたほうが，あるメディアの評価値が文書の評価値に反映されるよりも良いことがわかった．

## 3 評価値への重みを考慮した各メディアの評価値の統合手法

### 3.1 重みの定義

テキストや画像の評価値は，それぞれのメディアに適した評価の方法で計算されているため，評価値の分布は各メディアによって異なる．2.1 節で述べた実験では，テキストの評価値が 0 もしくは 1 に近い値であるのに対して，画像の評価値が 0 付近に集中していた．これらのような，分布が異なる評価値から文書の評価値を計算した場合，文書の評価値にほとんど画像もしくはテキストの評価値が反映されない場合がある．本研究では，これらの問題を解決するために重みを導入した．重みを考慮したテキストの評価値，画像の評価値 $X_i^{'term}$，$X_i^{'image}$ をそれぞれ以下のように定義する．

$$X_i^{'term} = \Theta_{term} \cdot X_i^{term}$$
$$X_i^{'image} = \Theta_{image} \cdot X_i^{image}$$

重み $\Theta_{term}$，$\Theta_{image}$ は，複数の問合せに対するのテキストと画像の評価値の分布を均等に並べるのが目的である．そこで本稿では，いくつかの問合せを行った結果計算されたテキスト，画像の評価値の中からそれぞれ最も高い評価値であったものを 1 とした場合の相対的な評価値を用いることとした．つまり，問合せ $Q_t(t = 1,2,\cdots$：問合せの数) に対する文書 $D_i(i = 1,2,\cdots$：検索対象となる文書数) のテキストの評価値を $X_{it}^{term}$，画像の評価値を $X_{it}^{image}$ とした場合に，$\Theta_{term}$，$\Theta_{image}$ を次のように定義する．

$$\Theta_{term} = \frac{1}{\max_{i,t} X_{it}^{term}}$$
$$\Theta_{image} = \frac{1}{\max_{i,t} X_{it}^{image}}$$

したがって，2.1 節で述べた 6 つの評価関数を $f$ とすると，文書の評価値 $X_i$ は次のように再定義される．

$$X_i = f(X_i^{'term}, X_i^{'image})$$

### 3.2 評価実験

2.1 節で述べた 6 つの評価関数を用いて文書の評価値を求める際に，重みを用いた場合の実験結果を図 2 に示す．ここで，一番適合率の高かった関数である相加平均を用いて，重みを用いた場合と用いなかった場合の比較を行ったところ，図 3 に示す結果となった．

この結果から，重みを用いて評価値の分布の差の軽減を図ることで適合率の向上が認められ，本手法の有効性を実証することができたことが分かる．

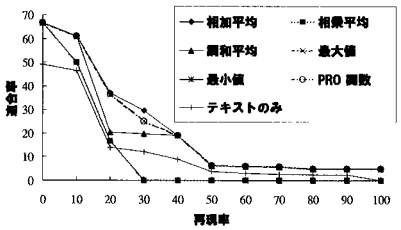


図 2: 重みを用いた場合の再現率 - 適合率グラフ

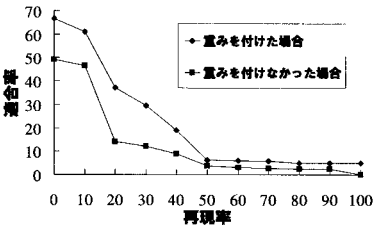


図 3: 重みを用いた場合と用いなかった場合の比較

## 4 おわりに

本稿では，PDF など複数のメディアで構成された電子文書を検索する際に用いられる，テキストと画像の評価値に対して重み付けを行う手法について提案した．以前の研究ではテキストの評価値と画像の評価値は異なる評価基準で評価値が求められており，各メディアの評価値の分布は異なるため，各メディアの評価値が文書の評価値に正しく反映されていなかったが，本稿で提案した手法により各メディアの評価値への重みを考慮することによって，文書の評価値へ正しく反映されるようになった．最後に，実際に PDF 文書を用いて実験を行い，重みの有効性を実証することができた．

今後は，利用者が問合せに重みを付ける方法について考慮する予定である．

**謝辞** 本研究の一部は，文部科学省科学研究費基盤研究（課題番号: 11480088，12680417，12780309），ならびに科学技術振興事業団戦略的基礎研究推進事業による．ここに記して謝意を表します．

## 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. TOD 11, Oct. 2001. (to appear).

[2] Gerard Salton, Edward A. Fox, and Harry Wu. Extended Boolean Information Retrieval. *Communications of the ACM*, Vol. 26, No. 11, pp. 1022 – 1036, 1983.