検索結果を統合するための関数選択手法
An Automatic Selection Method of Combination Functions

鈴木 優♡ 波多野 賢治◇ 吉川 正俊♠
植村 俊亮◇ 川越 恭二♡

Yu SUZUKI Kenji HATANO
Masatoshi YOSHIKAWA
Shunsuke UEMURA Kyoji KAWAGOE

本論文では，複数の検索システムから出力された検索結果を統合するための最適な統合関数を問合せごとに推定する方法として，シャノンの情報量の概念を援用した尺度を用いて，スコアの分布から統合関数の適合度を測定する方法を提案する．本提案では，利用者にとって必要な検索対象の数が検索対象全体の数と比較してきわめて少なく，検索結果に含まれる高いスコアの数が少ないとき，その検索結果は十分に正解集合を絞り込んでいると考えることができるため，利用者の検索目的に適した統合関数であると仮定した．評価実験を行うことによって，実際に適した統合関数を選択することができることを示した．

In this paper, we propose an automatic selection method of combination functions to integrate multiple results of retrieval subsystems. In our method, we automatically select using the distributions of relevance scores. When a retrieval system calculates a small number of high relevance scores, the system can identify relevant retrieval targets. Therefore, we suppose that the sureness of the combination functions should depends on the number of high relevance scores. Then, we propose a calculating method of the sureness of combination function using Shannon's information measure. From our experimental results, we find out that our proposed method can select better combination functions.

1. はじめに

本論文では，複数の検索システムによる検索結果からなる統合検索結果を検索する際に，それぞれの検索結果に含まれるスコア，つまり問合せと検索対象との類似度を統合するための最適な統合関数を，検索システムによって推定する手法の提案を行う．

現在様々な研究者によって，検索システムの精度を向上させるための手法として，複数の異なる検索システムを統合する方法が考えられている．これらの検索システムは，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 [4] 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．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ため，この手間を軽減するためには，検索システムによって検索結果だけから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の分布に注目することによって，統合関数が最適である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の分布を測定し，統合関数の自動選択に用いた．

♡ 正会員 立命館大学情報理工学部情報コミュニケーション学科
{yusuzuki,kawagoe}@is.ritsumei.ac.jp
◇ 正会員 奈良先端科学技術大学院大学情報科学研究科
{hatano,uemura}@is.naist.jp
♠ 理事 名古屋大学情報連携基盤センター
yosikawa@itc.nagoya-u.ac.jp

2. 最適な統合関数の自動選択手法

本章ではまず複数の検索サブシステムを統合した検索システムの全体像を示し，統合関数を自動的に選択するための詳細を示す．

2.1 検索システムの全体図

本論文では，検索サブシステムを組み合わせた検索システムを考える．図 1 に，本研究で想定している検索システムを画像検索に用いた場合の全体図を示す．この検索システムは，大きく分けて三つの部分で構成されており，それぞれ次のような機能を持つ．

1. **検索サブシステムによる検索**
   利用者が検索システムに問合せを入力した時，その問合せを検索サブシステムに再び入力する．次に，検索サブシステムが，それぞれの検索手法を用いて検索結果を出力する．ここで検索結果は，一つの検索対象に対して，複数の検索サブシステムが計算したスコアを付与している．
2. **スコアの統合**
   スコアの正規化手法を用いて，各々の検索サブシステムが出力したスコアを正規化する．さらに，統合関数を用いてそれぞれのスコアを統合し統合スコアを計算する．最後に検索対象を統合スコアの昇順に並べる．つまり，統合関数の数と同じ数の統合検索結果が得られる．
3. **統合関数の選択**
   統合検索結果群におけるスコアの分布から，最も検索精度が良い統合関数を推定し，その統合関数による統合検索結果を利用者に対して出力する．

ここで，検索システムの精度を上昇させる方法として特に 3. の部分に注目する．以前の著者らによる研究 [6] では，Lee [2] の研究において用いられている 29 種類の関数を統合関数として用いた場合の検索システムの精度の変化を調査した結果，問合せによって最適な統合関数が異なることが分かった．そこで，本研究では問合せごとに最適な統合関数を決定する方法について述べる．

2.2 基本的な考え方

2.2.1 適合検索対象数が少ない場合

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 [3] や NTCIR [5] 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用意されている．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，高いスコアが付与された検索対象の数が少なければ，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る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

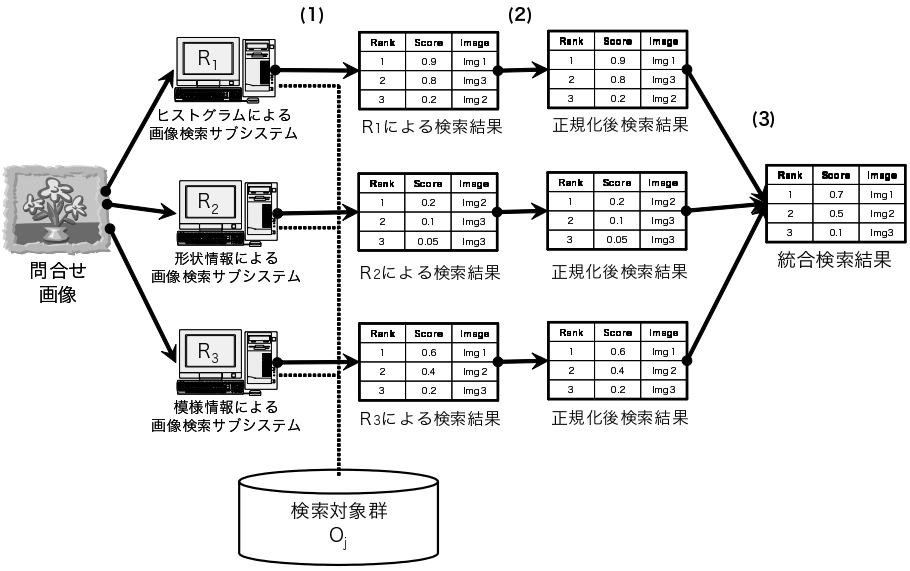


図 1: 複数の検索サブシステムを組み合わせた画像検索システム
Fig. 1 An image retrieval system with multiple image retrieval subsysyems.

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と考えられる．

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．

2.2.2 適合検索対象数が多い場合

前節では，正解検索対象の数が全検索対象の数と比べて比較的少ない場合の議論を行ったが，実際には正解検索対象集合の数が非常に多い場合も存在する．例えば，データベースに関連する論文群だけを検索対象とした場合，利用者が検索キーワードとして“データベース”と入力したときには，多くの検索対象を正解とするべきであると考えられる．そこで，このような場合には，どのようなスコア分布が良い分布であるかを考える必要がある．

ここで，前節と同様，統合関数 F_1，F_2 をそれぞれ用いた検索システム $R(F_1)$，$R(F_2)$ のスコア分布が図 2 であったとする．ところが以前の条件と異なり，正解検索対象数が多いという条件から，検索システムは検索対象群から少数の正解検索対象へ絞り込む必要がない．つまり，検索システムは多くの検索対象に高いスコアを付与した場合（図 2 (b)）であっても，それらの検索対象は全て正解である可能性がある．同様に，少数の検索対象に高いスコアを付与した場合（図 2 (a)）には，低いスコアが付与された検索対象が正解である可能性がある．つまり，利用者の検索意図に合致した統合関数をこれら二つの統合関数から選択することは難しい．

一方，利用者はスコアが高い順に高々数百件だけを閲覧するだけである点に注目する．この場合，どのような方法で数百件の検索結果を得たとしても，その中に正解検索集合が含まれる可能性は高い．つまり，本節における前提条件の場合には，検索結果の上位部分だけに着目した場合には，どのような統合関数を用いたとしても再現率は低く，適合率は高いと考えられるため，統合関数の選択は検索精度に影響が無いと考えられる．

以上の議論から，正解検索対象の数の大小にかかわらず，本提案手法を利用することによって検索精度が向上すると考えられることを示した．次節では，これらの手法を定式化し，実際に統合関数を選択するためのアルゴリズムを示す．

2.3 統合関数の自動選択手法

2.2 節で述べた，統合関数の自動選択手法に関する議論を基に，統合関数の自動選択手法を示す．提案手法は次に示す三つの手順で表すことができる．

1. **スコアの分布を求める**
   まず，統合後のスコアを $[0,1]$ の範囲となるように正規化を行う．次に，スコアの範囲を等しく分割し，それぞれのスコアの範囲に含まれる検索対象の数を求める．
2. **各々の検索対象に対して情報量を求める**
   検索対象集合全体の数に対する各々のスコアの範囲における検索対象の数の割合を，各々の検索対象がそのスコアを付与される確率と考えられる．そこで，そのスコアの値に対する情報量をシャノンの情報量 [4] を用いて定める．
3. **利用者の検索意図への統合関数の適合度を計算する**
   それぞれの検索対象に付与された情報量を統合して，ある統合関数が利用者の検索意図にどの程度適合しているかを計算する．

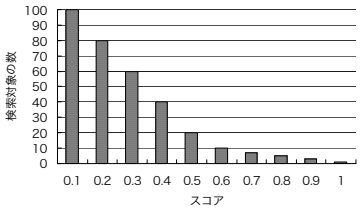
以下，それぞれの手順について説明する．

2.3.1 スコアの分布

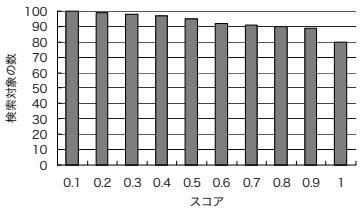
まず，統合関数 F_i $(i = 1, 2, \cdots, M)$ を用いた検索システム $R(F_i)$ を用いて，検索対象 O_j $(j = 1, 2, \cdots, N)$ に対してスコア $S(F_i, O_j)$ を計算する．次に，これらのスコア群から，スコアの分布 $G(F_i, k)$ $(k = 0.1, 0.2, \cdots, 1)$ を計算するために，次式を用いる．

$$G(F_i, k) = |(\mathbf{O}|0 \leq S(F_i, O_j) \leq k)| \quad (1)$$

ここで，$\mathbf{O}$ は検索対象集合であり，スコアが 0 以上 k 以下のものである．さらに，$|\mathbf{O}|$ は $\mathbf{O}$ の要素数，つまり条件を満たした検



(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
Fig. 2: Relevance score distributions of retrieval systems $R(F_1)$ and $R(F_2)$.

索対象集合の数である．つまり，$G(F_i,k)$ はスコアが 0 以上 k 以下である検索対象数を表している．

2.3.2 検索対象における情報量の計算

次に，スコアの分布を基に，それぞれの検索対象に対して情報量を計算する．統合関数 F_i を用いた場合の，検索対象 O_j の情報量 $I(F_i,O_j)$ を次式で示す．

$$I(F_i,O_j) = -log\frac{G(F_i,k)}{N} \tag{2}$$

つまり，検索対象 O_j が区分されている $G(F_i,k)$ の値を用いて情報量を計算する．

2.3.3 統合関数への適合度の計算

最後に，全ての検索対象に対して計算された情報量を統合して，統合関数への適合度を計算する．統合関数 F_i への適合度 $T(F_i)$ は，次式を用いて計算する．

$$T(F_i) = \sum_{j=1}^{N} I(F_i,O_j) \tag{3}$$

以上に示した適合度を，あらかじめ準備した全ての統合関数に対して計算する．そして，適合度が最も高い統合関数は利用者の検索意図に最も適合すると考えられるため，利用者にはその統合関数を用いて計算された検索結果を提示する．

次に，提案手法を用いることによって検索精度が上昇することを，実際の検索システムを用いて示す．

3. 予備実験

本論文で提案した提案手法が有効であることを確かめるために，小規模な予備実験を行った．本実験では，提案手法によって選択された統合関数を用いた場合とランダムに統合関数を用いた場合の比較を行い，提案手法を用いることによって検索精度が向上することを示すことを目的としている．

3.1 実験方法

本実験では，複数の検索システムを比較するために，複数の画像検索サブシステムを組み合わせた検索システムを用いる．実験の手順を以下に示す．

1. 問合せを検索システムに入力する．
2. 検索システムはあらかじめ用意した複数の統合関数を用い，検索結果を出力する．
3. 提案手法を用いて，利用者の検索意図に合致した統合関数を求める．
4. すべての検索結果から，再現率と適合率を計算する．
5. 提案手法によって選択された統合関数による再現率，適合率を他の統合関数の場合と比較する．

再現率，適合率を求める場合には，テストコレクションと呼ばれる問合せとそれに対応する正解検索対象の組を用意する必要がある．ところが，我々は画像検索システムに適用することが可能なテストコレクションを発見することができなかった．そのため，本実験では第一著者を含む数人の研究協力者の協力を得て画像検索エンジンのためのテストコレクションを作成した．検索対象となる画像の数は約 30000 枚であり，問合せと正解検索対象の組の数は 15 個である．

本実験で用いた検索サブシステムは三種類であり，それぞれ画像の色ヒストグラム，画像に写っている対象物の模様情報と形状情報の三つの特徴量を用いた検索システムである．これは，現在提案されている主な画像検索エンジンは，これら三つの種類の特徴量のうちのどれかが用いられていることが多いためである．

3.2 実験に用いた統合関数

本実験で用いた統合関数は CombSUM，CombMNZ，CombANZ の三つであり，それぞれ次の式 (4)，(5)，(6) で定義されている [1].

$$IS(O_j) = \sum_{i=1}^{N} S(R_i,O_j) \tag{4}$$

$$IS(O_j) = \sum_{i=1}^{N} \left(S(R_i,O_j)\cdot K(R_i)\right) \tag{5}$$

$$IS(O_j) = \sum_{i=1}^{N} \left(\frac{S(R_i,O_j)}{K(R_i)}\right) \tag{6}$$

3.3 実験結果

図 3 に，各々の問合せを用いた場合の平均適合率を示す．*MAX* は三つの統合関数のうち最適なものを選択した場合，*MIN* は最も悪いものを選択した場合，*AVG* は三つの統合関数用いた場合のそれぞれの平均適合率を平均したもの，*Proposed* は提案手法によって選択された統合関数を用いた場合の平均適合率を表している．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．また，全ての問合せにおいて最も平均適合率が低い統合関数は選択されなかった．さらに，三つの統合関数全てが一つ以上の問合せにおいて最も平均適合率が悪いことが分かった．つまり，一般的な手法である，ある一つの統合関数を用いた場合には最も平均適合率が低い統合関数を選択する場合があることから，提案手法を用いることによって検索精度が向上するといえる．

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場合に適合率が上昇す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．

4. 関連研究

本研究の目的である，複数の検索システムを統合することによる検索システムを実現する方法として，提案手法のようにスコア

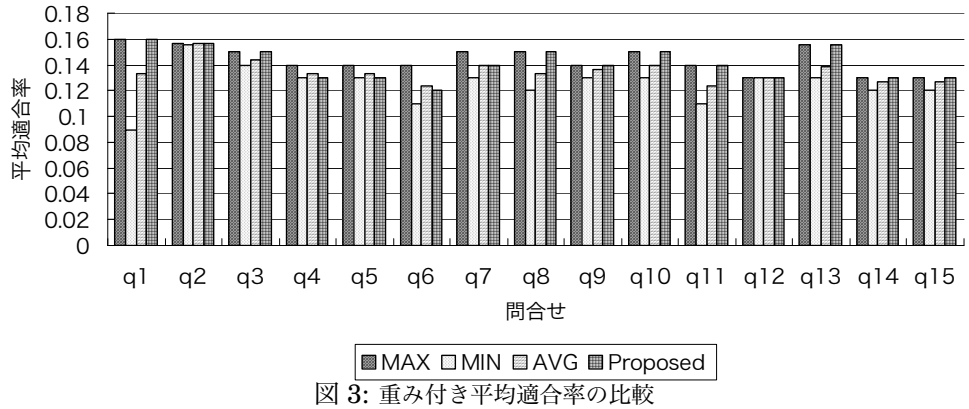


図 3: 重み付き平均適合率の比較
Fig. 3: Comparisons of weighted averaging precisions.

の分布から最適な統合関数を推定する方法の他に，関連フィードバックを用いる方法や機械学習による方法など様々な方法が考えられる．そこで，それぞれの方法と我々の手法との差異について述べる．

関連フィードバックによる方法とは，一度何らかの検索結果を利用者に提示した後，利用者によって必要であると考えられる検索対象を利用者自身が選択することによって，統合関数を選択する方法である．利用者自身の検索対象の選択を統合関数の選択に直接反映させることができるため，利用者の検索意図が反映されやすく，検索精度は大幅に向上すると考えられる．ところが，この方法では検索システムが最初に提示する検索結果に利用者の検索意図を反映させることができない．さらに，最初に提示する検索結果の上位に正解検索対象が含まれていない場合には，利用者が必要であると考えている検索対象を選択することができず，検索精度が低下する場合があると考えられる．つまり，提案手法を用いた後の検索結果に対して関連フィードバックを用いることによって，さらに検索精度を向上することも考えられる．

一方，ある利用者にとって特定の統合関数を用いることが良いと推定される場合，利用者による複数の問合せ，正解検索対象の入力を用いて，利用者ごとに特定の統合関数を推定することが機械学習などの方法を使って可能であると考えられる．ところが，我々は利用者の問合せごとに最適な統合関数は異なると考えており，利用者が連続して複数の検索を行った場合であっても，それらの検索意図は相互に異なると考えている．そのため，利用者の検索意図に対する傾向だけを用いて統合関数を定めた場合に，検索意図の差異が統合関数の選択に反映されず，利用者の検索意図に合致した統合関数を選択することができないと考えた．

以上の議論から，提案手法では，利用者によるフィードバックを行わない方法であり，しかも利用者の検索意図の長期的傾向に基づかない手法を用いた．

5. おわりに

本論文では，検索システムの精度を向上させるために，統合関数を問合せごとに自動的に選択する方法を提案した．ここで，検索対象に付与されたスコアから統合関数ごとにスコア分布を求め，統合関数に対して利用者の検索目的への適合度を計算した．さらに，スコア分布から適合度を求めるためにシャノンの情報量を用いた．最後に，予備実験を行うことによって，提案手法を用いて選択した統合関数は確かに利用者の検索意図に適合していると考えられることが分かった．

本論文では予備実験として，Fox らが提案した三つの統合関数から一つの統合関数を選択する実験を行った．ところが，情報検索の分野では今までに 29 種類の統合関数が提案されていることが Lee [2] によって示されている．そこで，これらの統合関数を用いた場合であっても提案手法が有効であることを示す必要があると考えられる．

また，予備実験では三つの検索サブシステムを全て用いたが，これらのうちいくつかは，検索精度の向上に寄与していない場合もあると考えられる．つまり，これらの検索サブシステムのうち二つだけを組み合わせた場合や一つだけを用いた場合には，三つ全て用いた場合よりも検索精度が向上する場合があると考えられる．このような場合に対応するために，提案手法を拡張する方法を提案しなければならないと考えられる．

[文献]

[1] E. A. Fox and J. A. Shaw. Combination of Multiple Searches. In *The Second Text REtrieval Conference (TREC2)*, pp. 243 – 252, 1993.

[2] J. H. Lee. Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval. Technical Report TR95-1501, Cornell University, 1995.

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

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

[5] 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/.

[6] 鈴木, 波多野, 吉川, 植村. 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, 42(SIG 10 (TOD 11)):11 – 21, 2001.

鈴木 優　Yu SUZUKI
立命館大学情報理工学部情報コミュニケーション学科講師．マルチメディア電子文書検索に関する研究に従事．ACM，情報処理学会，日本データベース学会，各会員．

波多野 賢治　Kenji HATANO
奈良先端科学技術大学院大学情報科学研究科助手．情報検索に関する研究に従事．ACM，IEEE Computer，情報処理学会，電子情報通信学会，日本データベース学会，各会員．

吉川 正俊　Masatoshi YOSHIKAWA
名古屋大学情報連携基盤センター教授．データベースシステムの研究に従事．ACM，IEEE Computer，情報処理学会，電子情報通信学会，各会員，日本データベース学会理事．

植村 俊亮　Shunsuke UEMURA
奈良先端科学技術大学院大学情報科学研究科教授．データベースシステムの研究に従事. 情報処理学会フェロー. 電子情報通信学会フェロー, IEEE Fellow, 日本データベース学会会員.

川越 恭二　Kyoji KAWAGOE
立命館大学情報理工学部情報コミュニケーション学科教授．情報システム，ネットワークサービス，データベースに関する研究に従事. 情報処理学会，IEEE，日本データベース学会各会員.