メタ検索における検索結果の統合方法の提案

大野　成義†‡　太田　学‡　片山　薫‡　石川　博‡

†職業能力開発総合大学校情報工学科　〒229-1196 神奈川県相模原市橋本台 4-1-1

‡東京都立大学大学院工学研究科　〒192-0397 東京都八王子市南大沢 1-1

E-mail: †ohno@cs.uitec.ac.jp, ‡{ohta, katayama, ishikawa}@eei.metro-u.ac.jp

あらまし　メタ検索において利用される検索エンジン数に対応する次元をもつ空間を考える．各検索エンジンの検索結果リストは本当の検索結果から各次元軸上に射影した結果であるとする．メタ検索における検索結果の統合では，一般に、その軸の直交性をもとに統合を行っている．これは，各検索エンジンでそれぞれ異なった方針に基づきランキングを行っているため独立の検索結果であると仮定しているからである．しかし，各検索結果間にはなんらかの相関が存在し，直交でなく斜交していると考えた方が自然である．本研究では，各検索エンジンの検索結果について，その間の相関から斜交の角度を決定し，統合する方法を提案する．この方法は事前トレーニングによる重みづけと異なり，検索毎に相関を決める動的な統合方法である．最後に，提案方法と直交を仮定した統合方法の比較検討を行う．

キーワード　情報検索，Web とインターネット，メタ検索，相関，斜交

Rank Combination Method of Metasearch

Shigeyoshi OHNO†‡　Manabu OHTA‡　Kaoru KATAYAMA‡　and　Hiroshi ISHIKAWA‡

†Department of Information and Computer Science ,Polytechinic University,
4-1-1 Hashimotodai, Sagamihara, Kanagawa, 229-1196

‡Graduate School of Engineering, Tokyo Metropolitan University, 1-1 Minami-Osawa, Hachioji Tokyo, 192-0397

E-mail: †ohno@cs.uitec.ac.jp, ‡{ohta, katayama, ishikawa}@eei.metro-u.ac.jp

Abstract　We suppose a space with its dimension being the number of search engines used in Metasearch and suppose the results of each engine to be features of some “universal result” mapped on these axes. The assumption of orthogonality often adopted for these axes will not be validated unless the results of the each engine are independent of each other. However, some correlations among these results are often existed. In this sense, it is more natural to think that the axes are oblique. In this work we propose the method of determining the crossing angles among these axes, which gives better results. This is a dynamic fusion method in the sense that correlations between the each respective retrieval result are taken into account and hence the angles are dynamically calculated, unlike the method by training. Finally, we compare our method with other ones in which an orthogonality is assumed among the axes.

Keyword　Information Retrieval, Web and Internet, Metasearch, Correlation, Oblique

1. はじめに

複数の検索エンジンの結果を利用するメタ検索エンジン[1][2]には，各検索エンジンの検索結果をどのように統合するのかというところに問題の一つがある．これは各検索エンジンでそれぞれ異なった方針に基づき検索され，ランキングされていることに起因する．

そこで，まず，検索エンジンの数と同じ次元数の多次元空間を考える．検索対象のページはこの多次元空間上に存在しており，各検索エンジンの検索結果は、それらを各次元軸に射影したものであると考える．そこで，各検索エンジンから得られた情報から多次元空間上の各ページの位置を求めることでより正しいランキングが得られると期待できる．従来の統合方法[3]では，これらの各軸は互いに直交しているとの仮定のもとで行われている．しかし，各検索結果を表す軸は斜交していると考えた方が自然である．

そこで，本研究では，斜交した軸を用いて行う統合方法を提案する．軸のなす角度は各検索結果間の相関からそれぞれ決定する．

なお，提案方法を定量的に評価するために適合する正解が存在する TREC(Text REtrieval Conference) [4] 参加グループの検索結果を利用した．提案方法は他の統合方法より高い適合率を得た．

2. 関連研究

2.1. スコア情報を用いた統合方法

TREC のデータを利用する取り組みとしては他に Aslam[5] らの研究など多くのもの[6-10]がある．TREC の参加グループの検索結果にはランキングの他にスコアが付加されており，ランキングよりも詳細データであると考えられていた．このため，これらの研究ではスコアを用いて統合を行っていることが多い．しかし，現実のメタ検索で利用する各検索エンジンに必ずしもスコアが付加されているとは限らず，スコアは補助的に扱いランキングを基準とする必要がある．また，TREC はコンテストであり評価はランキング付けで決まるため信頼できるスコアを算出しているとは限らない．TREC の 2000～2002 年のデータにおいてスコアを用いた統合方法とランキングを用いた統合方法を比較した場合，特に 2001 年以降では必ずしもスコアを用いた方法が良いとは限らない．[11]

そこで本研究では、最新の TREC 結果を利用するがスコアは使わずランキングを基準に統合を行う．

2.2. トレーニングを必要とする結合方法

TREC のデータを利用したこれまでの取り組みは，トレーニングの有無で分類することもできる．ここでいうトレーニングとは，事前にテスト検索を行うことである．テスト検索を行うことで，メタ検索に利用される個々の検索エンジンに重みづけを行い，より高い平均適合率を得られるように工夫している．[6] [7] [8]

メタ検索で使う検索エンジンが固定しており，検索エンジンの性能が変化しない場合は良いが，実際にはそうとは限らない．検索エンジンの性能はより精度を上げようと努力されており，検索エンジンの性能がかわるごとにテスト検索を行う必要がある．

そこで，本研究では事前のテスト検索を必要としない統合方法を検討する．

2.3. 直交性にもとづく結合方法

検索結果を統合する場合，一般に，各検索結果に対して直交性が期待されていると考えられる．

メタ検索で使用する検索エンジンの数と同じ次元を持つ空間を考える．各検索エンジンから得られた検索結果のリストを空間の軸上に並べる．検索エンジンを対等に扱い，ランキングの上位なものほど原点から離れるように並べるため以下のように規格化する．

$$r_i(a) ? (num_i ?1? R_i(a))/num_i \qquad (1)$$

ここで，添え字の i は検索エンジンを表し，$r_i(a)$ はページ a の規格化されたランキング値，$R_i(a)$ はページ a のランキングの値，num_i は検索結果リストに含まれるページ総数とする．

各検索エンジンを $1,2,3,,,n$ とするとページ a は n 次元空間上の座標 $(r_1(a), r_2(a), r_3(a),,, r_n(a))$ をもつ点であり，メタ検索における最も単純な統合方法 (Borda-fuse)

$$r(a) ? ? \sum_{i?1}^{n} r_i(a) \qquad (2)$$

は n 次元空間上の $n-1$ 次元超面で描かれた等高線（超面）によってランキングの値を決められていると解釈できる．

例として表２のような２つの検索結果を統合する．

表１　２つの検索結果の一覧

ランク	検索エンジン１	検索エンジン２
1	ページ A	ページ A
2	ページ B	ページ C
3	ページ C	ページ B
4	ページ D	ページ D

ページ A,B,C,D の統合したランクキング値は式（１）と式（２）より

$$r(A) ? \frac{4?1?1}{4} ? \frac{4?1?1}{4} ? 2, \quad r(B) ? \frac{5}{4}, \quad r(C) ? \frac{5}{4}, \quad r(D) ? \frac{1}{2}$$

となり，２次元空間（平面）で表現すると図１のようになる．統合したランキング値が等しいページ（ページ B とページ C）は図１の点線で表現した直線（値 5/4 の等高線）上にのる．統合したランキング値が小さいほど原点に近い直線上にのる．式（１）で規格化したランキング値を使っているため，統合したランキング値が小さいということは本来のランキングで下位に位置することになる．

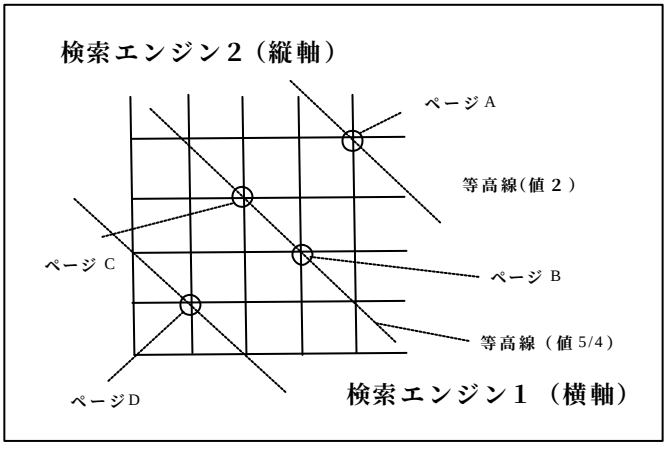

図１　２次元空間での表現

Fox[12]らが提案し，Lee[9]が実験を行った統合方法

$$r(a) = n_a^{\gamma} \sum_{i=1}^{n} r_i(a) \qquad (3)$$

（ここで n_a はページ a を含む検索結果を出した検索エンジンの個数を表し，γ は1とする．0 の場合は Borda-fuse と一致する．）は，図1で軸上に位置する（検索エンジン1の軸上にある場合は検索エンジン2の検索結果にランキングされていないことを意味する．）ページは同じ等高線上にあっても統合した順位を低くしていることになる．

また，対数を用いて上位のものに重みを付けて統合する方法[11]

$$r(a) = \sum_{i=1}^{n} (1 - Log_{num} R_i(a)) \qquad (4)$$

は，図1で各軸の目盛りを原点から離れるほど大きくすることと等価である．

事前トレーニングを行い検索エンジン i の重み α_i を計算してから統合する Vogt[6]らの方法

$$r(a) = \sum_{i=1}^{n} \alpha_i r_i(a) \qquad (5)$$

も，事前トレーニングによって軸ごとに目盛りの大きさを変えているだけである．

以上のように全て直交性を前提に統合を行っている．これは各検索結果間には依存関係が存在しないと期待しているからである．

本研究では各検索結果にはなんらかの依存関係が存在するとした方が自然であると考え，斜交軸を導入する．

3. 斜交性にもとづく結合方法の提案

従来の統合方法は直交性を前提に各検索エンジン間のなす角度という概念は考慮していない．しかし、各検索エンジン間にはなんらかの依存関係があると仮定した方が自然であり、各検索エンジンの検索結果間の相関関係を統合に利用する．

本研究で提案する統合方法は得られた各検索エンジンの検索結果の相関を計算し斜交角度を求めるステップと決定した斜交軸を用いて検索結果を統合するステップの2段階で行う．

3.1. 相関係数による斜交角度の決定

一つの検索要求に対して得られた各検索エンジンの検索結果の相関を計算し、多次元空間の軸のなす角度とする．相関を調べる一般的な方法としてはピアソンの相関係数があり，検索エンジン i と j の検索結果の相関係数 r_p は式（1）で規格化されたランキング値 $r_i(a)$ をもちいて以下の式で求められる．

$$r_p = \frac{\frac{1}{n-1} \sum_a (r_i(a) - \overline{r_i})(r_j(a) - \overline{r_j})}{\sqrt{\frac{1}{n-1} \sum_a (r_i(a) - \overline{r_i})^2} \sqrt{\frac{1}{n-1} \sum_a (r_j(a) - \overline{r_j})^2}}$$

$$\overline{r_i} = \frac{1}{n} \sum_{a=1}^{n} r_i(a)$$

ここで n は検索エンジン i と j の検索結果に含まれるページの総数であり num_i や num_j に等しいとは限らない．検索エンジン j の検索結果に含まれるが検索エンジン i の検索結果に含まれないページがあれば、そのページ数だけ n は num_i より大きい数となる．また，ページ a が検索エンジン i の検索結果に含まれない場合は $r_i(a) = 0$ とする．求めた r_p は検索エンジン i と j のなす角度の余弦として斜交軸を構成する．

つまり検索エンジン i と j の検索結果間の相関係数が α でなったとする．検索エンジン i と j の検索結果を表す軸をそれぞれ単位ベクトル $\vec{w}_i$ と $\vec{w}_j$ で表すと，そのベクトルの内積 $\vec{w}_i \cdot \vec{w}_j$ が α となる．

特殊な例として検索エンジン i と j の検索結果が全く同じ場合 r_p は1となる．同様に完全に逆順の検索結果である場合は-1となり問題はない．しかし，両方の検索結果に含まれるページが全くなかった場合 r_p は0とはならず，負の値になり弱い逆相関があるかのような値になってしまう．両方の検索結果に含まれるページが全くない場合，その検索結果間には相関がない，つまり無相関になるべきである．そこでそのような場合，相関係数が0となり，無相関になるような相関係数の定義を考える．

ピアソンの相関係数は自然数の相関を調べる場合スピアマンの順位相関係数 r_s と同値である．

$$r_s = 1 - \frac{6 \sum_a d_a^2}{n^3 - n}, \qquad d_a = r_i(a) - r_j(a)$$

スピアマンの順位相関係数は $r_i(a)$ が0の場合（ページ a が検索エンジン i の検索結果に含まれていない場合），定義されていない．そこで，両方の検索結果に含まれるページが全くない場合は0となるように，以下のような修正した相関係数 r_m を定義する．

$d_a = r_i(a) - r_j(a)$ $r_i(a) - r_j(a) \neq 0$ の場合

$d_a = \sqrt{\frac{(n^2 - 1)}{12}}$ $r_i(a) - r_j(a) = 0$ の場合

$r_m = 1 - \frac{6 \sum_a d_a^2}{n^3 - n}$

上記の d_a は, 同一ページが相関を調べようとしている２つの検索エンジンでの順位の差を意味している. 順位の差が全て 0 のとき強相関となり，n， $n-2$，$n-4$,,, 2, 0 の値をとるとき逆相関となる. そこで、一方の検索エンジンの検索結果に含まれるが，もう一方の検索エンジンの検索結果に含まれないページに関しては，0 と n の間の定数値とする．全てがこのような場合，r_m が 0 となるように定数値を決めた.

3.2. 直線上への射影による統合計算

相関係数を求めることで，統合する検索エンジンの検索結果を表す軸上の単位ベクトル $\vec{w}_i$ が統合する検索エンジンの数だけ得られる．各検索エンジンの検索結果は多次元空間上の各ページの射影であると考え，各検索結果から各ページの位置を求める．これは検索結果を表す軸上のランキングの値からこの軸に垂直な超平面を考え，全ての超平面の交わる点が求める位置となる．求める多次元空間上のページ a の位置を表すベクトルを $\vec{v}_a$ とすると，$(\vec{v}_a - r_i(a)\vec{w}_i) - r_i(a)\vec{w}_i = 0$ を満たす $\vec{v}_a$ を求めればよい．ここで、-はベクトルの内積を表し，$r_i(a)$ は式（１）の規格化されたランキングの値である.

図２で多次元空間上の各ページの位置を黒丸で表し，各検索エンジンでのランキングを白丸で表す．簡単にするため統合される検索エンジンは２つとし，図２では多次元空間を２次元平面で表現する.

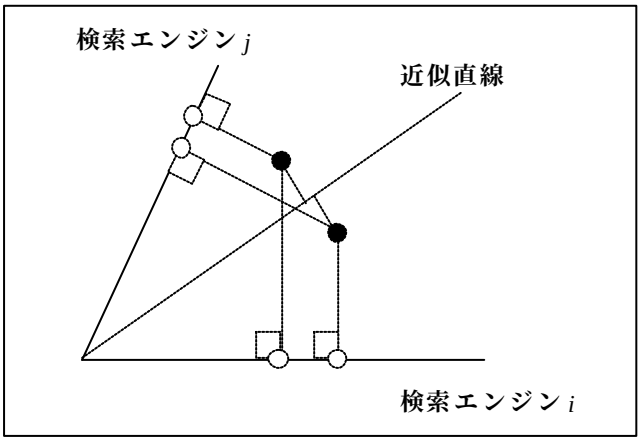

図２　各ページを示す点と各検索結果を表す軸

各ページは空間上に分布しており，一つの直線上に並んできるとは限らない．統合結果であるランキングのリストを得るために空間上に分布している点の集合を一つの直線で近似する必要がある．多次元空間上の各ページの重心を求め，原点とその重心を通る直線をその近似直線とする．従って，この近似した直線はベクトル $\vec{V} = \frac{\sum_a \vec{v}_a}{|\sum_a \vec{v}_a|}$ で表現できる． ここで $\sum_a$ は統合される検索エンジンの検索結果に含まれる全てのページに関して和をとることを意味する．全ての検索結果に含まれている必要はない.

各ページからこの近似直線に下ろした垂線の足が最も本当の検索結果に近い検索結果であると期待できる．従って，$|\vec{v}_a(\vec{V} - \frac{\vec{v}_a}{|\vec{v}_a|})|$ がページ a の統合したランキングの値 $r(a)$ となる.

4. 実験結果と考察

4.1. ランキングが 1000 位までである場合の統合結果

提案方法の妥当性を評価するために TREC 2002 と TREC 2001 と TREC-9 の Web Track のデータを利用する．まず，それぞれ上位５グループの検索結果に対して統合を行った．表２はその５グループのグループ名であり，表３がその適合率の平均である．表 3 で 5 docs，10 docs，15 docs，20 docs とあるのはそれぞれ統合結果上位 5,10,15,20 位までででの適合率である.

表２　利用した上位５グループのグループ名

TREC-9	TREC 2001	TREC 2002
iit00m	iit01m	pltr02wt2
jscbt9wll1	ok10wtnd1	uog04cta2dph
jscbt9wcl1	csiro0mwa1	Mercah
jscbt9wll2	ok10wtnd0	icttd1
ric9dpn	flabxtd	uog03ctadqh

表３　上位５グループの適合率

	TREC-9	TREC 2001	TREC 2002
平均適合率	0.2856	0.2763	0.1697
5docs	0.4352	0.5536	0.2506
10docs	0.38	0.4916	0.218
15docs	0.3427	0.4451	0.1948
20docs	0.3437	0.4032	0.17632
R 適合率	0.3196	0.3044	0.1991

５グループの検索結果をそれぞれ統合した結果が表４，表５，表６である．TREC の Web Track は毎年 50 のクエリに対する検索コンテストである. そのため 50 のクエリに対する５グループの検索結果をクエリごとに相関を求め，各検索エンジン間のなす角度を決

定して統合を行った.

表４　TREC-9 のデータを使った統合結果（各検索結果の 1000 位までを統合）

	単純統合	対数統合	斜交統合	修正斜交統合
平均適合率	0.3243	0.331	0.3475	0.3505
5 docs	0.488	0.488	0.504	0.504
10 docs	0.416	0.414	0.428	0.428
15 docs	0.3893	0.384	0.3947	0.3947
20 docs	0.363	0.361	0.372	0.372
R 適合率	0.3433	0.3599	0.3633	0.3651

表５　TREC 2001 のデータを使った統合結果（各検索結果の 1000 位までを統合）

	単純統合	対数統合	斜交統合	修正斜交統合
平均適合率	0.3728	0.3771	0.3879	0.3909
5 docs	0.672	0.652	0.678	0.684
10 docs	0.574	0.604	0.608	0.61
15 docs	0.5187	0.524	0.528	0.5427
20 docs	0.468	0.487	0.489	0.501
R 適合率	0.3934	0.3872	0.3964	0.4014

表６　TREC 2002 のデータを使った統合結果（各検索結果の 1000 位までを統合）

	単純統合	対数統合	斜交統合	修正斜交統合
平均適合率	0.195	0.2002	0.2046	0.206
5 docs	0.298	0.3061	0.302	0.3161
10 docs	0.2286	0.2327	0.2306	0.2326
15 docs	0.1973	0.2042	0.2068	0.2068
20 docs	0.199	0.198	0.198	0.198
R 適合率	0.2217	0.2302	0.2308	0.2301

TREC の各検索結果は最大 1000 位までがリストされており,統合結果も最大 1000 位までとした.従って,平均適合率だけが良くなっても実際のメタ検索で有効とは限らない．検索エンジンを使うユーザーは一般にはリストの上位５から２０位くらいまでしか見ないことが多いと考えられ,メタ検索でも同様だと思われる.従って，上位５位から２０位くらいまでての適合率が最も重要であると考えられる.

表で単純統合は単純な統合方法(Borda-fase)（以下，単純統合と記述）を，対数統合は対数を用いて各検索結果の上位に重みをつけた統合方法（以下，対数統合と記述）を，斜交統合はピアソンの相関係数を用いて斜交軸のなす角度を決めて統合を行った提案方法（以下，斜交統合と記述）を，修正斜交統合は修正した相関係数を用いて角度を決め統合を行った提案方法（以下，修正斜交統合と記述）をそれぞれ意味する.

比較しやすいように単純統合との比をグラフ化したものが図３，図４，図５である．グラフでは平均適合率と上位 5 位までと 10 位までての適合率のみ表示した．縦軸は単純統合に比べてどの程度適合率が変化したか変化率をパーセントで表している.

斜交統合は単純な統合方法よりも適合率が良くなっており，一部例外を除いて対数統合よりも良い．修正斜交統合は斜交統合とほとんど同じ傾向であるが,更に適合率が向上している.

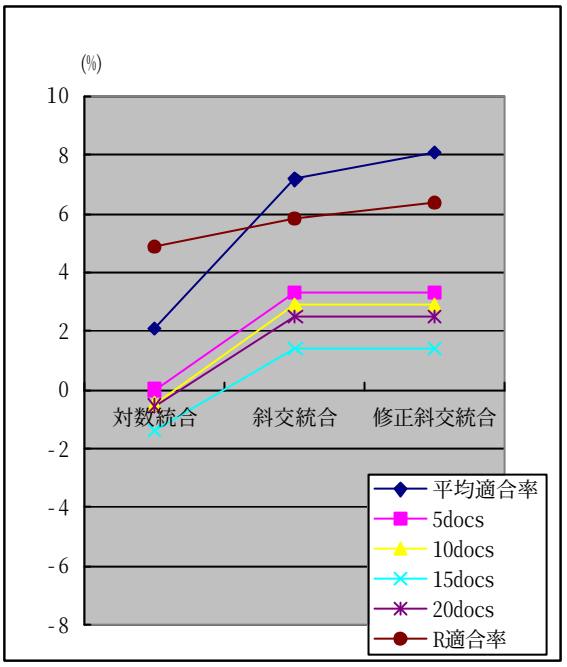

図３　TREC-9 のデータを使った統合結果（各検索結果の 1000 位までを統合）

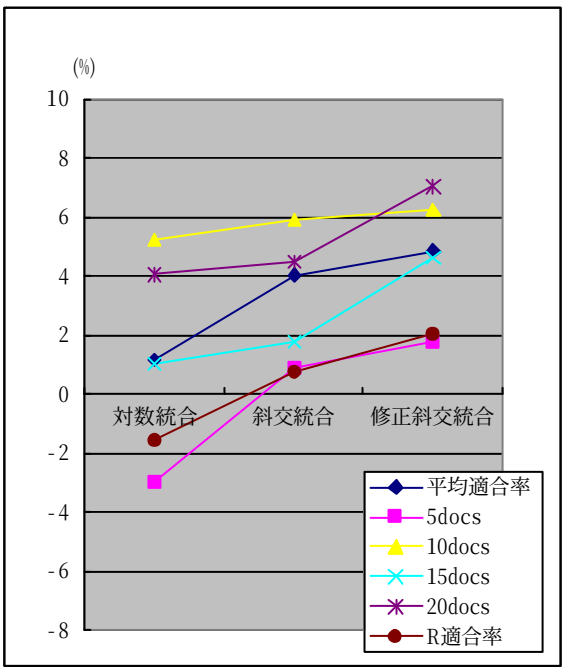

図４　TREC 2001 のデータを使った統合結果（各検索結果の 1000 位までを統合）

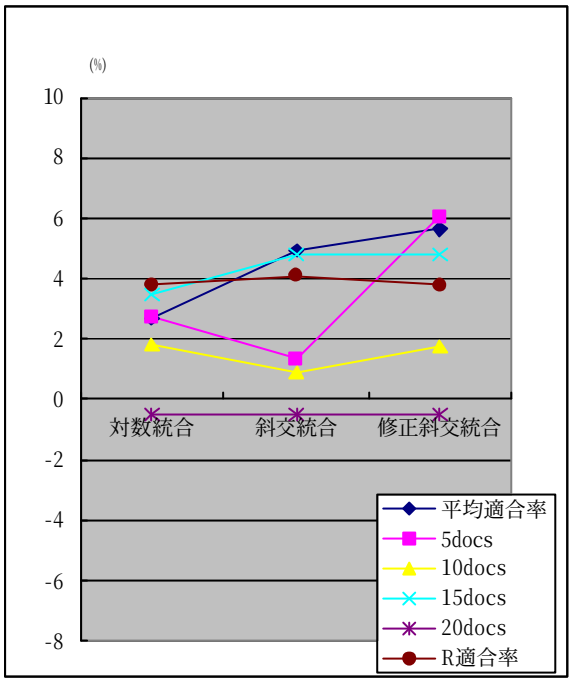

図５　TREC 2002 のデータを使った統合結果（各検索結果の 1000 位までを統合）

4.2. ランキングを 100 位までに制限した場合の統合結果

TREC のコンテストは一つのクエリに対して最大 1000 位まで検索結果を求めることができる．しかし，実際のメタ検索エンジンで利用する各検索エンジンはそれぞれ 1000 位までの検索結果を返すとは限らない．むしろ,最大 50 位までや 100 位までに限定して利用することが考えられる．そのような場合でも提案方法が有効であるか確認するために，統合する各グループの検索結果を 100 位までに制限し，統合した結果も 100 位までに制限した結果を表７，表８，表９に示す．

統合方法による違いを比較しやすいように，1000 位までのときと同じように図６，図７，図８でグラフ化する．100 位までに制限した場合は 1000 位までと異なり斜交統合は対数統合より必ずしも良いとはいえない．むしろ，単純統合よりも統合効果が低い場合さえある（図７）．一方，修正した相関係数を用いた斜交統合は一部例外はあるものの他の統合方法より適合率が向上していることが確認できる．

表７　TREC-9 のデータを使った統合結果（各検索結果の 100 位までを統合）

	単純統合	対数統合	斜交統合	修正斜交統合
平均適合率	0.2714	0.2823	0.2805	0.2899
5 docs	0.4640	0.4760	0.48	0.476
10 docs	0.4060	0.4000	0.404	0.408
15 docs	0.3673	0.3720	0.368	0.3747
20 docs	0.3390	0.3450	0.346	0.3472
R 適合率	0.3298	0.3395	0.3398	0.341

表８　TREC 2001 のデータを使った統合結果（各検索結果の 100 位までを統合）

	単純統合	対数統合	斜交統合	修正斜交統合
平均適合率	0.3003	0.2984	0.2865	0.3052
5 docs	0.65	0.652	0.64	0.66
10 docs	0.57	0.574	0.566	0.582
15 docs	0.469	0.5033	0.504	0.5147
20 docs	0.4	0.4127	0.4061	0.412
R 適合率	0.3429	0.3447	0.3264	0.3440

表９　TREC 2002 のデータを使った統合結果（各検索結果の 100 位までを統合）

	単純統合	対数統合	斜交統合	修正斜交統合
平均適合率	0.1733	0.1752	0.1751	0.1761
5 docs	0.3002	0.302	0.308	0.308
10 docs	0.2247	0.2265	0.2247	0.2286
15 docs	0.2068	0.2109	0.2082	0.2095
20 docs	0.199	0.2031	0.199	0.199
R 適合率	0.2188	0.2242	0.2238	0.2279

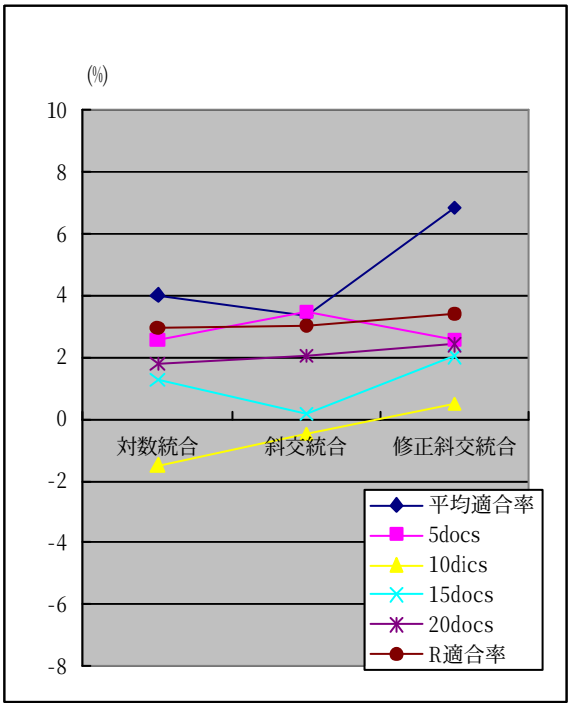

図６　TREC-9 のデータを使った統合結果（各検索結果の 100 位までを統合）

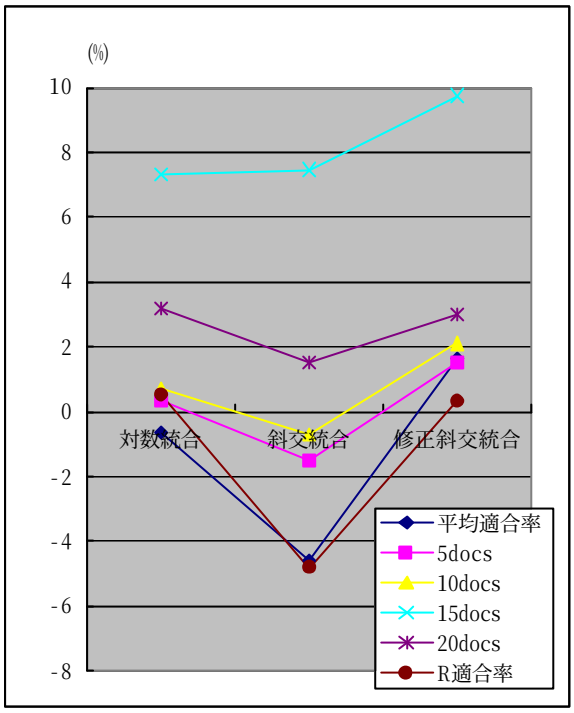

図７　TREC 2001 のデータを使った統合結果（各検索結果の 100 位までを統合）

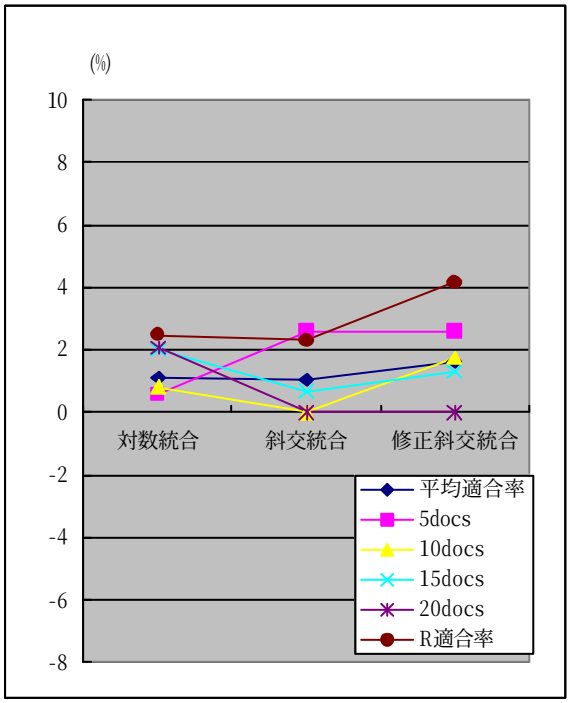

図８　TREC 2002 のデータを使った統合結果（各検索結果の 100 位までを統合）

4.3. ピアソンの相関係数と修正した相関係数の比較

ピアソンの相関係数 r_p と修正した相関係数 r_m を比較した結果が表１０である．TREC-9，TREC 2001，TREC 2002 において 1000 位までで統合したものと，同じく 100 位までで統合したものの 6 つに分けて，全ての角度，全てのクエリに対して $(r_m ? r_{p)/} r_m$ を求めて平均した．単位はパーセントである．

表１０　ピアソンの相関係数と修正した相関係数との差

TREC-9	1000 位までで統合	12.4±0.2
TREC 2001	1000 位までで統合	11.4±0.8
TREC 2002	1000 位までで統合	13.3±0.3
TREC-9	100 位までで統合	10.5±0.6
TREC 2001	100 位までで統合	11.3±0.2
TREC 2002	100 位までで統合	10.5±0.2

ピアソンの相関係数 r_p と修正した相関係数 r_m との差は全て正の値で，明らかに前者の方が小さい．相関を調べる２つの検索結果で片方にしか含まれないページのランキングの値を $r_i(a) ? 0$ としたが，これが相関係数を小さくするように働いたためであると考えられる．r_m は負の値をとることはまれであったが，r_p が負の値をとることは r_m に比べてかなりあった．

4.4. 考察

一般的な相関係数を用いた斜交統合は，他の統合方法に比べて優位な場合もあるが，図５や図７で明らかなように必ずしも良い方法とはいえない．特にランキングを 100 位までに制限した場合は単純統合にさえ劣る場合があった．一方で修正した相関係数を用いた斜交統合は，ランキングを 100 位までに制限した場合も、1000 位までにした場合も他の統合方法より優位であることが確かめられた．

表１０でピアソンの相関係数と修正した相関係数の違いを調べたが，その違いがどのように統合に寄与するか実験では明らかにできなかった．しかし，一般的な相関係数が修正した相関係数より小さい値をとってしまうことが，ピアソンの相関係数を使う斜交結合を有効性の高いものにするのを妨げていると考えられる．

実験的に，修正した相関係数の方が斜交統合では有効であることが確かめられた．斜交統合では，一方の検索結果にだけ含まれるページが多い程，検索結果間

の相関は無相関に近くなるように相関関係を決めるべきであるといえる．しかし，今回の修正した相関係数の定義が最も良いかどうかは検討の余地がある．

提案方法はクエリごとに各検索結果間の相関を求め，各検索エンジン間のなす角度を決定して統合をおこなう．これは事前にテスト検索を行い，各検索エンジンの性能を評価し，検索エンジンごとの重みをつけて統合する方法とは二つの意味で異なる．

一つは検索エンジンごとの重みづけでは自由度が統合する検索エンジンの数であるのに対して提案方法は検索エンジンの数を n とすると $\frac{n(n-1)}{2}$ だけの自由度をもっている．４つ以上の検索エンジンを統合する場合は自由度だけをみてもより精密な統合を行える可能性がある．

二つ目の違いはそもそもテスト検索を行う必要がない点である．インターネット上に存在する検索エンジンは日々データを更新しており，検索性能も改良が加えられていると思われる．日を変えて検索を行えば同じ結果が返ってくるとは限らない．従って，そのような検索エンジンを利用するメタ検索でテスト検索を行っておいても，すぐまたテスト検索を行わなければ今現在の検索エンジンの状況を反映したものにならない．また，新たな検索エンジンがインターネット上に登場しても，テスト検索が終わるまでは利用できないことになる．一方で，提案方法は新たな検索エンジンが登場しても，テスト検索することなくすぐ利用できる．

提案方法はスコアでなくランキングのみで統合を行い，クエリの内容も考慮していない．つまり，どのような検索を行っても利用できるより汎用性の高い統合方法であるといえる．さらにいえばメタ検索におけるランキングの統合だけでなく，より広範囲な分野でのランキングの統合に利用できる統合方法[13]であることを意味している．今回の実験で，修正した相関係数を用いた斜交統合は TREC-9，TREC 2001，TREC 2002 の３つで，さらにランキングを最大の 1000 にとった場合も 100 に限定した場合も有効であることが確認できた．これは他の分野のランキングの統合でも有効である可能性を意味している．良い例とはいえないかもしれないが，例えば，テニスやサッカーの世界ランキングのように，各大会でのランキングからそれらを統合して総合ランキングを決める場合などにも利用できる可能性がある．

5. おわりに

メタ検索における各検索エンジンの検索結果の相関から斜交軸を求めて統合を行う方法を提案した．また，その提案方法を評価するために TREC のデータを使って実験を行った．単純な統合方法と比較して最大で 8%以上適合率が向上した．

同じページを全く含まない検索結果間の相関を無相関になるように修正した相関係数を定義したが，その正当性に関しては検討の余地がある．

文　献

[1] http://www.ceek.jp/

[2] http://bach.istc.kobe-u.ac.jp/metcha/

[3] W.Meng, C.Yu and K.Liu, Building Efficent and Effective Metasearch Engines, ACM Computing Surveys, Vol.34, No.1, pp48-89, Mar.2002

[4] TREC http://trec.nist.gov/

[5] J.A.Aslam and M.Montague, Models for Metasearch. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.276-284. 2002.

[6] C.C.Vogt and G.W.Conttrell, Fusion via a Linear combination of scores, Information Retrieval,1(3), pp151-173, Oct.1999.

[7] L.Si and J.Callan, Using Sampled Data and Regression to Merge Search Engine Results, Procceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002.

[8] B.Yuwono and D.Lee. Server Ranking for Distributed Text Retrieval System on Internet. Procceedings of the International Conference on Database Systems for Adv. Applicatons, pp.41-49, 1997

[9] J.H.Lee, Analyses of multiple evidence combination, Procceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp267-275, 1997.

[10] J.Savoy, A.L.Calve, and D.Vrajitoru, Report on the TREC-5 experiment:Data fusion and collection fusion, The Fifth Text Retrieval Conference (TREC-5), Gaithersburg, MD, 1997.

[11] 大野成義，太田学，片山薫，石川博，“メタ検索における検索結果の統合方式の検討，”情報処理学会第 65 回全国大会講演論文集，分冊 3，3E-2，pp.23-24, Mar.2003.

[12] E.A.Fox and J.A.Shaw, Combination of multiple searches, The Second Text Retrieval Conference, pp.243-249, Gaithersburg, MD, USA, Mar.1994.

[13] R.Fagin, R.Kumar, D.Sivakumar, Comparing top k lists, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 2003