コピュラを用いたスコア統合方式の改良による情報検索の高精度化

左近　健太†　　宮崎　　純†

† 東京工業大学情報理工学院　〒 152–8552 東京都目黒区大岡山 2-12-1
E-mail: †{sakon@lsc., miyazaki@}cs.titech.ac.jp

あらまし　本研究では，ユーザの多様な情報要求に対応するために，複数の検索モデルで求めた各文書の適合度を確率モデルであるコピュラを用いて統合し，情報検索の高精度化を実現する統合方式を提案する．三つ以上の検索モデルのスコア統合においても，検索モデルのスコア分布に関する統計的性質を調べることによって，統合方式としてコピュラや線形結合の組み合わせのうち，高精度化につながる方法を自動的に判定できるようにした．

キーワード　コピュラ，情報検索，スコア統合

1　はじめに

情報検索システムとは，各種のデータベースや Web などの大規模な情報データベースから，ユーザの検索要求であるクエリにできるだけ合致したデータを返答するシステムである．情報検索システムでは適合度を調べ，そのランキングの高いデータから順番に結果を出力している．適合度とは，ユーザのクエリにデータが適合する度合いである．適合度の計算では理論的に確定した手法がなく，これまでさまざまな適合度を計算する検索モデルが提案されてきた [1], [8]．その評価手法を研究するワークショップとして TREC[1] がある [14]．

しかしながら，ユーザの情報検索要求は多種多様であり，単一の検索モデルにより算出された適合度では，複雑な検索要求に適切に対応することは難しいという問題がある．この問題に対応するため，複数の検索モデルにより算出された適合度を統合する手法が提案されてきている [2]．

例えば，複数の適合度の重み付き和である線形結合を用いたものがあり，検索モデル間の適合文書の重複が多く，不適合文書の重複が少ない場合などに有効性が示されている [13]．ただ，線形結合はモデル間の非線形な関係をうまく捉えることが難しく，適用範囲には限界がある．

そこで，金融工学においてリスク管理に用いられていたコピュラ [9] を利用した統合が考えられ成果が出ている [5]．コピュラとは，多変量同時分布を各変量の分布 (周辺分布) と周辺分布間の関数としてとらえたものであり，その関数形により各適合度間の非線形な依存関係をとらえることができる．

さらに，適合度の分布をクラスタリングしてからそれぞれのクラスタ分けされた小領域にコピュラを当てはめ，それらのコピュラを線形結合することによって，より精度の高い適合度の統合を行うという多峰性コピュラ (混合コピュラ) の手法が提案されている．これにより，適合度間の非線形な部分的依存関係も捉えられる [6]．

さらに，いろいろな例を詳細に検討してみたところ，統合する検索モデルの種類によっては，コピュラだけではなく，コピュラ以外の統合方法もうまく活用する必要があることがわかった．今回はコピュラ以外の統合方法には線形結合を用いている．実験の結果，複数の検索モデルの分布に関する分散などの統計的指標を用いることで，コピュラと線形結合のどちらを用いたほうが精度が高いかを推定できることが分かった．これらの複数の指標の大小関係に基づいてコピュラや線形結合を用いることで，より多くの検索モデルの組み合わせに対応できるようになり，さまざまな検索モデルおよび三つ以上の検索モデルの統合に関してもコピュラや線形結合の組み合わせのうち，高精度化につながる方法を自動的に判定できるようにした．

本稿の以降の構成を述べる．2 節でコピュラ，3 節で関連研究を述べる．4 節で統計的指標に関する予備実験について述べ，5 節でコピュラおよび線形結合を用いた三つ以上の検索モデルの統合方法の提案について述べ，6 節で提案手法の評価実験について述べる．最後に 7 節でまとめを述べる．

1：http://trec.nist.gov/

2　コピュラの概要

2.1　コピュラの定義と性質

コピュラとは，多変数の分布関数とその周辺分布関数の関係を表す関数のことである．n 個の連続な 1 次元周辺分布関数 $F_1, \ldots, F_n$ を持つ n 次元同時分布関数を F とすると，以下の関係を満たす C が一意的に存在する [12]．(スクラーの定理)

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n))$$

この C をコピュラと呼ぶ．コピュラは n 次元単位立方体 $[0,1]^n$ から単位区間 $[0,1]$ への関数であり，以下の性質を持つ．

- コピュラ $C(u_1, u_2, \ldots, u_n)$ は単調増加 (n-increasing) である．
- u_i 以外の要素をすべて 1 にしたときコピュラの値が u_i となる．すなわち，$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$
- 少なくとも 1 つの要素 u_i が 0 である場合，コピュラの値は 0 となる．すなわち，$C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_n) = 0$

2.2　代表的なコピュラ

- アルキメデス型コピュラ

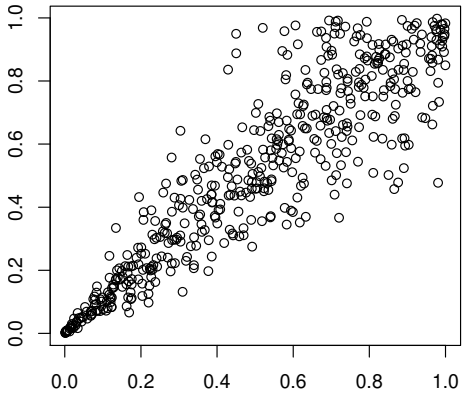

図 1　クレイトンコピュラに従った分布の例

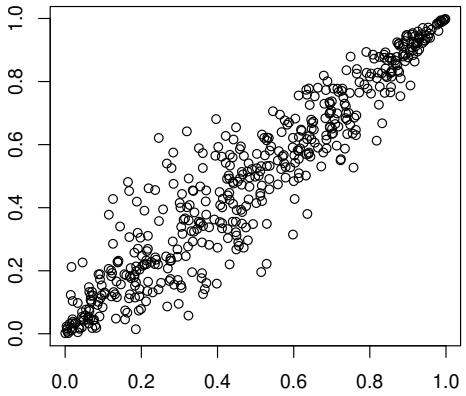

図 2　グンベルコピュラに従った分布の例

区間 $[0,1]$ 上で定義され，正の実数値をとる単調減少凸関数 φ が $\varphi(1)=0$ を満たすとする．このとき，

$$C_\varphi(U)=\varphi^{-1}(\varphi(u_1)+\varphi(u_2)+\ldots+\varphi(u_n))$$

を n 次元アルキメデス型コピュラと呼ぶ．φ は C_φ の生成素 (ジェネレーター) といい，通常一つのパラメータ θ を含む．以下に示すクレイトンコピュラ，グンベルコピュラ，フランクコピュラはアルキメデス型コピュラである．

- クレイトンコピュラ

$$C_{Clayton}(U)=(1+\sum_{i=1}^{n}(u_i^{-\theta}-1))^{-\frac{1}{\theta}}$$

であらわされるコピュラである．ジェネレータは $\varphi(t)=(t^{-\theta}-1)/\theta$ である．図 1 に 2 次元クレイトンコピュラに従った分布の例を示す．

- グンベルコピュラ

$$C_{Gumbel}(U)=\exp(-(\sum_{i=1}^{n}(-\log u_i)^\theta)^{\frac{1}{\theta}})$$

であらわされるコピュラである．ジェネレータは $\varphi(t)=(-\log t)^\theta$ である．図 2 に 2 次元グンベルコピュラに従った分布の例を示す．

- フランクコピュラ

$$C_{Frank}(U)=-\frac{1}{\theta}\log\left(1+\frac{\prod_{i=1}^{n}(\exp(-\theta u_i)-1)}{(\exp(-\theta)-1)^{n-1}}\right)$$

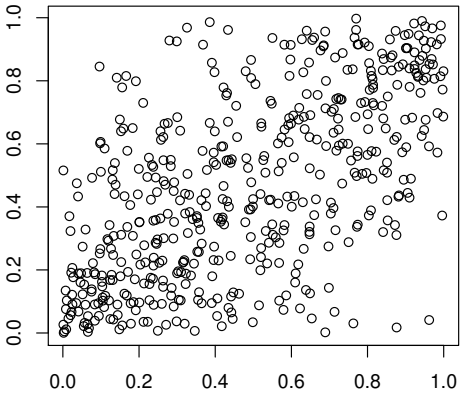

図 3　フランクコピュラに従った分布の例

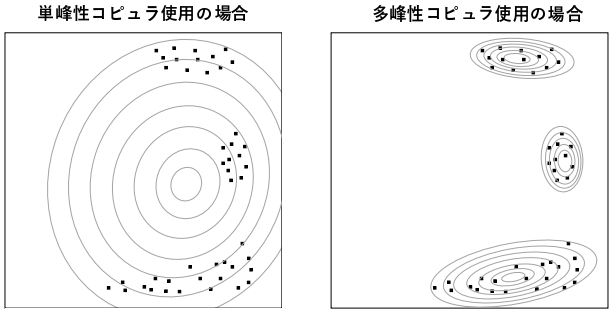

図 4　単峰性コピュラと多峰性コピュラの違い

であらわされるコピュラである．ジェネレータは $\varphi(t)=-\log((\exp(-\theta t)-1)/(\exp(-\theta)-1))$ である．図 3 に 2 次元フランクコピュラに従った分布の例を示す．

- 経験コピュラ

経験的な同時分布から，周辺分布をその経験分布で与えて導いたコピュラを経験コピュラと呼ぶ [3]．経験コピュラは以下の式により与えられる [16]

$$\hat{C}(U)=\frac{1}{N}\sum_{k=1}^{N}\prod_{i=1}^{n}\mathbf{1}\{t_i^k\leqq u_i\}$$

N は学習データの数，N 個の n 変量データ $(x_1^1,\ldots,x_n^1),\ldots,(x_1^N,\ldots,x_n^N)$ が観測され，第 i 変量の値を小さい順に並べ替え，$x_i^k(k=1,\ldots,N)$ が r_i^k 番目になったとしたとき，$t_i^k=r_i^k/N$ となる．このコピュラにより，学習データの分布を正確に再現した同時分布を推定することができる．

3　関 連 研 究

3.1　コピュラを用いた適合度の統合

Eickhoff らは，金融工学においてリスク管理に用いられていたコピュラに注目し，これを適合度の統合に利用した研究を行い，その有効性を示した [5]．適合度の統合にコピュラを用いることにより，各適合度間の非線形な依存関係をとらえることができるようになった．

しかしながら，図 4 のように，複数の検索モデルによる適合度の分布が局所的に相関の高いいくつかのクラスタに分かれている場合，Eickhoff の用いた単峰性コピュラではクラスタごとの依存関係を的確に捉えることは難しい．そこで，Komatsuda

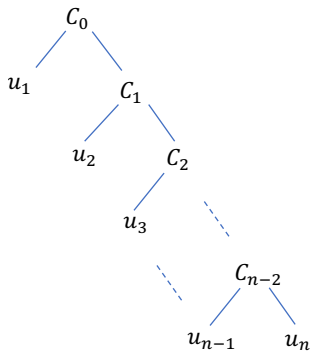

図 5　nested copula における統合

らは適合度の分布を群平均法でクラスタリングしてからそれぞれのクラスタにコピュラを当てはめ，コピュラを線形結合することによって，精度の高い適合度の統合を行うという多峰性コピュラ (混合コピュラ) の手法を提案した [6]．また，佐々木らの研究においてもコピュラにより適合度と文書の可読性の統合が行われ，精度が向上した [11]．ただし，適合度の分布に外れ値がある場合，精度低下の可能性があった．そのため，クラスタリング手法を距離ベースから密度ベースに変えることで，それを改善する研究が行われた [17]．

3.2　nested copula を用いた適合度の統合

Eickhoff らは，三つ以上の検索モデルのスコア統合において，nested copula の手法を提唱した [4]．nested copula とは，パラメータの異なるコピュラを用いて検索モデルのスコアを順に統合していく方法であり，複数の検索モデルのスコアの関係をより適切に捉えることができるため，一つのコピュラで複数の検索モデルのスコアを同時に統合するよりも精度を向上させることができた．nested copula における統合は図 5 の通りである．$u_1, u_2, \cdots, u_n$ は検索モデルを，$C_0, C_1, \cdots, C_{n-2}$ はコピュラを表しており，図 5 の全体は nested copula による統合の一例を表している．

nested copula はパラメータの異なるコピュラを用いることができるが，その各パラメータが仮に検索モデルに対してあらかじめ固定的に決まっているとすると，統合結果は統合順序に依存しない．しかしコピュラパラメータを各二つのスコア統合時に最尤推定などによりその場で求めるとすると，統合順序によって統合結果が変わってしまうので，より精度を上げるには統合順序を考慮する必要がある．Eickhoff はこの問題を提起しているが, [4] では統合順序をランダムに行うとして，明確な回答を与えていない．多峰性コピュラの場合，ネストした時に統合順序に結果が依存しないという数学的裏付けがないため，この統合順序が問題となる．本論文ではこれに一つの回答を与えるものである．

4　予 備 実 験

本節では，検索モデルのスコア統合後の精度を実際に統合を行わずに推定することによって，統合方法としてコピュラがよいのか線形結合がよいのかを選択できることを示すことを目的とする．この研究における精度には $nDCG@5$ を用いるものとする．二つの検索モデルのスコア統合において，コピュラの方が精度が高くなる場合と，線形結合の方が精度が高くなる場合がある．ただし，どちらの方が精度が高くなるかを判断するのに実際に統合して決めるとすると手間がかかってしまう．

まず，二つの検索モデルのスコアの統合において，検索モデルのスコア分布に関する統計的性質によって，コピュラで統合したときの精度と，線形結合で統合したときの精度を推定でき，それによって統合方法を選択できると仮定し，その統計的性質とは何なのかを予備実験によって明らかにする．

4.1　統計的指標

本研究では統合手法を選択する統計的指標として，相関係数，ケンドールのタウ，および分散共分散行列の固有値の合計を検討し，このうち適切な指標を予備実験によって明らかにする．

4.1.1　相 関 係 数

相関係数とは，二つの変数間の線形関係を測定する指標であり，分布が線形関係に近いほど高い値になる．相関係数は -1 から 1 までの値を取る．相関係数を用いるにあたっては，データが正規分布であることが前提となっており，パラメトリックな方法であるといえる．X，Y を検索モデルのスコアの集合として X，Y のスコア分布を考えたとき，相関係数 cor の式は以下のように表される．

$$cor = \frac{Cov_{XY}}{S_X \cdot S_Y}$$

Cov_{XY} は X，Y の共分散を，S_X は X の標準偏差を，S_Y は Y の標準偏差を表している．

4.1.2　ケンドールのタウ

ケンドールのタウ (ケンドールの順位相関係数) は，順位間の相関を表す指標であり，二つの変数の値の順序関係が似通っているものほど値が高くなる．ケンドールのタウは -1 から 1 までの値を取る．ケンドールのタウはノンパラメトリックな方法である．ケンドールのタウ τ の式は以下のように表される．

$$\tau = \frac{K - L}{{}_nC_2}$$

K は n 要素から 2 要素を選んだときに順位関係が一致する組の数である．L は n 要素から 2 要素を選んだときに順位関係が一致しない組の数である．データの大小関係のみを考慮し，データの値そのものは使用しない．

4.1.3　分散共分散行列の固有値

分散共分散行列の固有値 [15] は，分布の広がりを示す指標であり，固有値の合計が大きいほど散らばりの大きい分布となる．分散共分散行列とは，分散と共分散から成る行列のことである．X，Y を検索モデルのスコアの集合として X，Y のスコア分布を考えたとき，分散共分散行列 Σ は次のような形式で表される．

$$\Sigma = \begin{pmatrix} V_X & Cov_{XY} \\ Cov_{YX} & V_Y \end{pmatrix}$$

V_X，V_Y はそれぞれ X，Y の分散を，Cov_{XY}，$Cov_{YX}(= Cov_{XY})$ は X，Y の共分散を表している．分散共分散行列

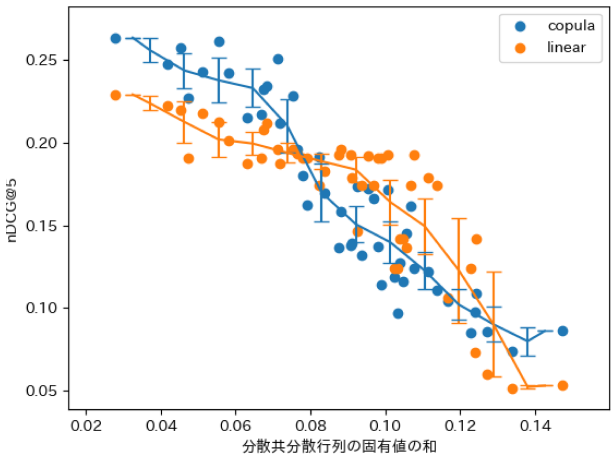

図 6　分散共分散行列の固有値の和と精度 ($nDCG@5$) との関係

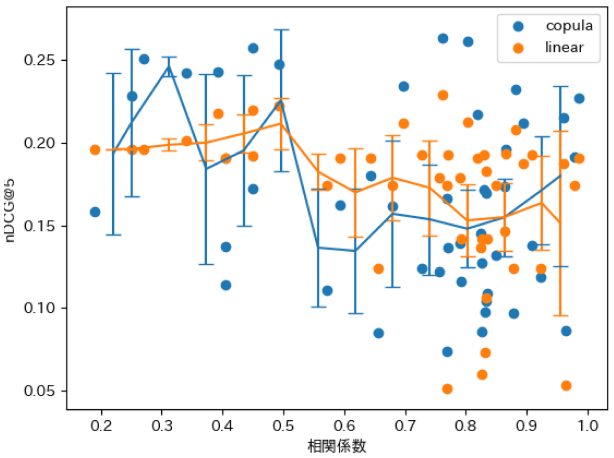

図 7　相関関係と精度 ($nDCG@5$) との関係

を固有値分解して固有値を求めた後，それらの固有値の合計を取ることによって固有値の合計を求めることができる．なお，固有値の合計は，元の変数の分散の合計に等しくなる．

4.2　統計的指標に関する予備実験

まず，二つの検索モデルのスコアの組み合わせをさまざまに変化させた場合，どういう条件の場合にコピュラの方が精度が高くなるのか，それとも線形結合の方が精度が高くなるのかを示すための予備実験を行った．

検索モデルの組み合わせによっては，コピュラの方が線形結合よりも精度が良い場合もあり，線形結合のほうがコピュラよりも精度が良い場合もある．

そこで，統合する検索モデルのスコア分布（散布図）に関する統計的性質を調べることで，コピュラで統合したときの精度と，線形結合で統合したときの精度を推定でき，それによって統合方法を選択できると仮定し，その統計的性質を予備実験によって明らかにする．ここで用いたコピュラは多峰性コピュラであり，事前に行っておくべきクラスタリング手法は密度ベースクラスタリングの Density Peak [7] としている．

検索モデルのスコア分布に関する統計的性質としては，分散や相関関係が考えられる．分散を表す指標として，固有値の合計を用いた．また，相関関係の指標として，相関係数およびケンドールのタウを用いた．固有値の合計と精度との関係，相関関係と精度との関係，ケンドールのタウと精度との関係をそれぞれ図 6，図 7，図 8 に示す．

検索モデル二つが与えられたときのそれぞれの指標 (固有値の合計，相関係数，ケンドールのタウ) の値とコピュラの精度との関係を図中の青い点で表現し，検索モデル二つが与えられたときのそれぞれの指標の値と線形結合の精度との関係を図中のオレンジの点で表現している．検索モデル二つの組み合わせの数だけ青い点 (または，オレンジの点) が存在している．それぞれの図の横軸は指標の値を，縦軸はコピュラあるいは線形結合の精度を表している．ここでの精度は $nDCG@5$ を用いている．

また，それぞれの図に対して横軸に関するスライディングウィンドウを考えて，それぞれの範囲で代表的な点を考えてそ

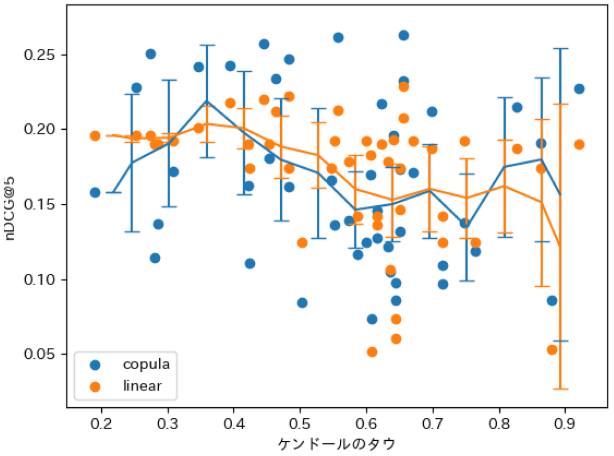

図 8　ケンドールのタウと精度 ($nDCG@5$) との関係

れらの代表的な点を結ぶことによって，それぞれの指標とコピュラの精度の変化，及び線形結合の精度の変化を示した．ここではスライディングウィンドウの窓の幅を横軸の座標全体の $\frac{2}{13}$，移動距離を横軸の座標全体の $\frac{1}{13}$ としている．代表的な点の横軸の座標，縦軸の座標に関しては，各ウィンドウの中心を代表的な点の横軸の座標とし，各ウィンドウに含まれている点に対して，それらの縦軸の座標の平均を考えることで代表的な点の縦軸の座標とした．

さらに，コピュラあるいは線形結合のそれぞれの代表的な点を中心として，縦軸方向に標準誤差のエラーバーを引いた．それぞれのエラーバーについては，各ウィンドウに含まれている点を対象として 95%信頼区間を表示している．

図 6，図 7，図 8 それぞれのエラーバーに注目すると，図 7，図 8 では，どのウィンドウをとってもコピュラと線形結合の精度には差があると判定できない．そのため，相関係数やケンドールのタウは，コピュラ，線形結合の精度の大小関係には影響があるとは言えない．一方，図 6 では，固有値の合計が小さければ，コピュラの代表的な点に関するエラーバーが線形結合の代表的な点に関するエラーバーよりも上回っている．一方，固有値の合計が大きい範囲では，線形結合の代表的な点に関するエラーバーがコピュラの代表的な点に関するエラーバーより

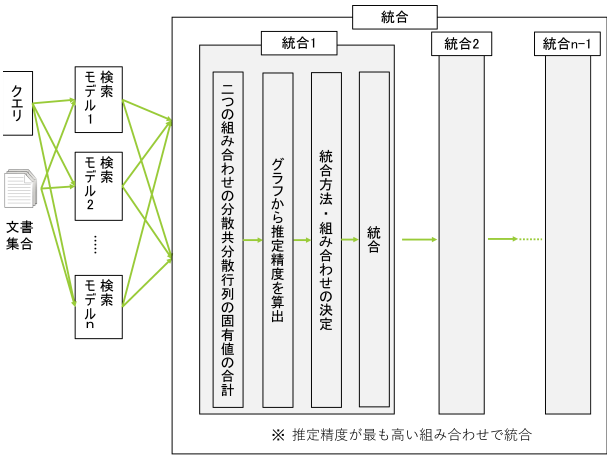

図 9　提案手法の統合方法

も上回っている箇所もある．そのため，固有値の合計がコピュラ，線形結合の精度の大小関係に対して影響していると考えられる．したがって，固有値の合計を用いれば，コピュラで統合したときの精度，線形結合で統合したときの精度を推定でき，統合方法を選択できる．

実際に検索モデルが与えられた場合，検索モデルのスコア二つの分散共分散行列の固有値の合計を求めた後，図 6 の曲線から推定精度を求めることになる．

5　提 案 手 法

三つ以上の検索モデルのスコア統合は，組み合わせ最適化問題となり，最適解を求めるためには，すべての統合の組み合わせを実際に行って精度を比較する必要がある．ただし，モデル数が多くなると，組み合わせの数が爆発的に多くなるので，すべての組み合わせを試行するのは現実的ではない．そこで今回は，二つずつの検索モデルの統合に還元し，各統合段階で最も推定精度の高い組み合わせとその統合方法を選び，順次統合していく Greedy アルゴリズムを検討した．これを Greedy-integrate と名付ける．

精度の推定に関しては，事前に多数のモデル，例えばパラメータ k と b の範囲をさまざまに変えた BM25 や TF-IDF，QLM(クエリ尤度モデル) などを用意してそれらの組み合わせに対する分散共分散行列の固有値の和，およびコピュラと線形結合のそれぞれで実際に統合を行ったときの精度を求め，グラフにプロットする．そのプロットをもとに，分散共分散行列の固有値の和から精度を推定できるようにする．

提案手法は図 9 の通りである．複数の検索モデルが与えられたとき，そのうち各二つずつのモデルの組み合わせにおいて，分散共分散行列の固有値の和を考えることで，図 6 からそれぞれの統合方法 (線形結合およびコピュラ) における推定精度を算出する．そのうち最も推定精度が高くなる組み合わせと統合手法 (線形結合かコピュラか) を求め，各地点で最も推定精度が高くなるようにモデル二つのスコアを統合する．この統合を繰り返していき，残りのモデルが一つになった地点で統合を終了する．

nested copula と比較すると，本研究では統合方法としてコピュラ，線形結合の両方を用いており，また，多峰性コピュラの順序を考慮した統合を行えるので，高精度化が実現できると考えられる．

6　実　　験

本節では，検索モデルのスコアが三つ，あるいは四つ与えられた場合に，Greedy-integrate が有効であることを示す．検索モデルの全ての統合方法の中で実際に最も精度が高い統合方法の精度と Greedy-integrate との精度とを比較し，Greedy-integrate の精度が，実際に最も精度が高い統合方法の精度と比べてどれだけ差があるかを調べる．Greedy-integrate の精度が，実際に最も精度が高い統合方法の精度に近ければ近いほど，Greedy-integrate が統合方法の選択として有効であるといえる．

ただし，実験の正当性を保証するため，実際に用いる検索モデルは，図 6 で使用していないモデルを用いる．

また，各統合段階で，ある二つの検索モデルの分散共分散行列の固有値の合計が図 6 の範囲外になってしまい，推定精度を全て求められない可能性がある．その場合は統合方法として，基本的に線形結合を採用する．図の範囲外における精度の推定については今後の課題としたい．

6.1　実験で使用する検索モデル

実験で使用する検索モデルは，索引語の重み付け方法として BM25 [10] を用いたブーリアンモデルで，これは文書の適合度を算出する代表的なモデルのうちの一つである．単語の重み $w(t,d)$ は以下のように表される．

$$w(t,d) = \frac{(k+1)tf(t,d)}{k(1-b+b\frac{dl(d)}{avdl})+tf(t,d)} \cdot idf(t)$$

ただし

$$idf(t) = log\left(\frac{N-df(t)+0.5}{df(t)+0.5}\right)$$

ここで，$tf(t,d)$ は文書 d 中の単語 t の出現頻度，$dl(d)$ は文書 d の語数，$avdl$ は文書集合全体の平均語数，N は文書数，$df(t)$ は文書中に単語 t を含む文書数，k，b はパラメータである．

6.2　検索モデルのスコアが三つの場合の実験結果

このパターンで用いた三つの検索モデルは，それぞれ BM25，QLM，TF-IDF で，これらの三つの検索モデルを (1)，(2)，(3) とする．(1) については，BM25 のパラメータ b，k の値をそれぞれ $(b,k) = (0.75, 1.0)$ としている．

最初に検索モデルのスコア二つを選んで統合する場合，それぞれの固有値の合計，コピュラあるいは線形結合で統合するときの推定精度 (図 6 から求めた) は表 1 の通りである．

表 1 から Greedy-integrate では，まずはじめに (1) と (2) をコピュラで統合すると判定される．(1) と (2) をコピュラで統合したモデルを (4) とすると，次に (4) と (3) を統合する場合の固有値の合計，コピュラあるいは線形結合で統合するときの推定精度は表 2 の通りである．この場合は，固有値の合計が図

表 1　最初に検索モデルのスコア二つを統合する場合の固有値の合計と推定精度

検索モデルの組み合わせ		(1), (2)	(1), (3)	(2), (3)
固有値の合計		0.0766	0.1114	0.1487
推定精度	コピュラ	**0.1980**	0.1207	-
	線形結合	0.1922	0.1470	-
実際の精度	コピュラ	**0.2125**	0.1372	0.1084
	線形結合	0.1942	0.1942	0.1438

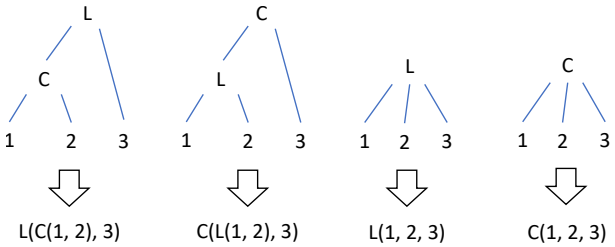

図 10　統合方法について

6 の範囲外となってしまったので，推定精度は求められない (表中では - としている)．そのため，(4) と (3) の統合方法として線形結合を採用している．

表 2　次に検索モデルのスコア二つを統合する場合の固有値の合計と推定精度

検索モデルの組み合わせ		(4), (3)
固有値の合計		0.2046
推定精度	コピュラ	-
	線形結合	-
実際の精度	コピュラ	0.0952
	線形結合	**0.2125**

また，Greedy-integrate とその他の統合方法との実際の精度の比較を表 3 に示す．それぞれの評価指標で最も高い精度が出た箇所は太字にしている．

ここで，図 10 は統合の順序と記号を表している． $L(1,2,3)$ とは検索モデル (1) と (2) と (3) を線形結合で同時に統合したことを表しており，$C(1,2,3)$ とは検索モデル (1) と (2) と (3) をコピュラで同時に統合したことを表している．また，$L(C(1,2),3)$ とは検索モデル (1) と (2) をコピュラで統合後，(3) を線形結合で統合したことを，$C(L(1,2),3)$ とは検索モデル (1) と (2) を線形結合で統合後，(3) をコピュラで統合したことを表している．他の統合方法も同様である．

表 3　検索モデル三つにおけるさまざまなスコア統合方法の精度比較

手法	$L(1,2,3)$	$C(1,2,3)$	$C(C(2,3),1)$	***Greedy*** $L(C(1,2),3)$	⋯	C(L(1,2),3)	C(C(1,2),3)
nDCG@5	0.1942	0.1328	**0.2225**	0.2125	⋯	0.1372	0.0952
nDCG@10	0.1836	0.1494	**0.2067**	0.1868	⋯	0.135	0.1171
nDCG@15	0.1764	0.1544	**0.1963**	0.179	⋯	0.128	0.1258
nDCG@20	0.1809	0.1562	**0.1985**	0.1797	⋯	0.1342	0.1242
iP@0.0	0.4355	0.3391	**0.4945**	0.4459	⋯	0.3272	0.2773
iP@0.1	0.3069	0.2629	**0.3275**	0.291	⋯	0.2183	0.2347
iP@0.2	0.2357	0.1915	**0.239**	0.2273	⋯	0.1565	0.1675
iP@0.3	**0.1241**	0.1222	0.1155	0.1186	⋯	0.0775	0.1079
P@5	0.276	0.188	**0.3**	0.288	⋯	0.188	0.132
P@10	0.258	0.21	**0.272**	0.246	⋯	0.184	0.166
P@15	0.244	0.2133	**0.2547**	0.2293	⋯	0.164	0.1853
P@20	0.24	0.208	**0.247**	0.224	⋯	0.167	0.178
ERR	0.08547	0.08023	**0.10325**	0.09653	⋯	0.07306	0.0526

今回の場合，Greedy-integrate は $L(C(1,2),3)$ に相当する．また，これらの統合方法のうち精度 ($nDCG@5$) が最大になるのは $C(C(2,3),1)$ になる．Greedy-integrate は最適解とは同じにならないものの，nDCG@5 の差が 0.01 と十分小さい差であることから，最適解に近い結果となる．

6.3　検索モデルのスコアが四つの場合の実験結果

このパターンで用いた四つの検索モデルは，それぞれ b，k の値を変えた 3 種類の BM25 と QLM で，3 種類の BM25 を (1)，(2)，(3) とし，QLM を (4) とする．(1) は，BM25 のパラメータ b，k の値をそれぞれ $(b,k)=(0.75,1.0)$ としており，(2) は，BM25 のパラメータ b，k の値をそれぞれ $(b,k)=(0.50,3.0)$ としており，(3) は，BM25 のパラメータ b，k の値をそれぞれ $(b,k)=(0.25,6.0)$ としている．

最初に検索モデルのスコア二つを選んで統合する場合，それぞれの固有値の合計，コピュラあるいは線形結合で統合するときの精度は表 4 の通りである．なお，推定精度は図 6 から求めている．

表 4　1 回目に検索モデルのスコア二つを統合する場合の固有値の合計と精度

検索モデルの組み合わせ		(1), (2)	(1), (3)	(1), (4)	(2), (3)	(2), (4)	(3), (4)
固有値の合計		0.0477	0.0570	0.0766	0.0654	0.0850	0.0943
推定精度	コピュラ	**0.2426**	0.2368	0.1980	0.2309	0.1656	0.1481
	線形結合	0.2109	0.2014	0.1922	0.1989	0.1876	0.1790
実際の精度	コピュラ	**0.2485**	0.2442	0.2125	0.2436	0.2367	0.2215
	線形結合	0.2117	0.2137	0.1942	0.2095	0.2095	0.1826

表 4 から Greedy-integrate では，まずはじめに (1) と (2) のスコアをコピュラで統合すると判定される．(1) と (2) のスコアをコピュラで統合したモデルを (5) とすると，次に (5) と (3) と (4) のスコアを統合する場合の固有値の合計，コピュラあるいは線形結合で統合するときの精度は表 5 の通りである．この場合は，一部の検索モデルの組み合わせにおいて固有値の合計が図 6 の範囲外となってしまったので，推定精度が求められていない箇所がある (表中では - としている)．

表 5　2 回目に検索モデルのスコア二つを統合する場合の固有値の合計と精度

検索モデルの組み合わせ		(3), (4)	(3), (5)	(4), (5)
固有値の合計		0.0943	0.2478	0.2674
推定精度	コピュラ	0.1481	-	-
	線形結合	**0.1790**	-	-
実際の精度	コピュラ	0.2215	0.2231	0.2390
	線形結合	0.1826	**0.2485**	**0.2485**

表 5 から Greedy-integrate では，次に (3) と (4) のスコアを線形結合で統合すると判定される．それを統合したモデルを (6) とすると，検索モデル三つの場合と同様にして表 6 から (5) と (6) のスコア統合方法として線形結合を採用する．

また，Greedy-integrate とその他の統合方法との実際の精度の比較を表 7 に示す．今回の場合，Greedy-integrate は $L(C(1,2),L(3,4))$ に相当する．また，これらの統合方法の

表 6　3 回目に検索モデルのスコア二つを統合する場合の固有値の合計と精度

検索モデルの組み合わせ		(5), (6)
固有値の合計		0.2483
推定精度	コピュラ	-
	線形結合	-
実際の精度	コピュラ	0.2406
	線形結合	**0.2485**

うち精度 ($nDCG$@5) が最大になるのは $C(L(C(1,3),4),2)$ になる．Greedy-integrate は最適解とは同じにならないものの，nDCG@5 の差が 0.0064 と十分小さい差であることから，最適解に近い結果となる．

表 7　検索モデル四つにおけるさまざまなスコア統合方法の精度比較

手法	C(L(C(1,3),4),2)	C(C(L(1,4),3),2)	···	***Greedy*** $L(C(1,2),L(3,4))$	···	L(L(L(3,4),2),1)	L(L(L(3,4),1),2)
nDCG@5	0.2549	0.2549	···	0.2485	···	0.2117	0.2113
nDCG@10	0.2202	0.2202	···	0.2211	···	0.2012	0.1979
nDCG@15	0.2112	0.2112	···	0.2101	···	0.196	0.1955
nDCG@20	0.2108	0.2108	···	0.2072	···	0.1921	0.1942
iP@0.0	0.5192	0.5192	···	0.4759	···	0.4707	0.4816
iP@0.1	0.3185	0.3185	···	0.3287	···	0.3197	0.312
iP@0.2	0.2313	0.2313	···	0.2009	···	0.235	0.2296
iP@0.3	0.1179	0.1179	···	0.0935	···	0.1296	0.1299
P@5	0.344	0.344	···	0.328	···	0.284	0.276
P@10	0.282	0.282	···	0.284	···	0.272	0.258
P@15	0.2667	0.2667	···	0.2653	···	0.264	0.2547
P@20	0.256	0.256	···	0.249	···	0.243	0.243
ERR	0.10987	0.10987	···	0.11186	···	0.09649	0.10022

手法	L(1,2,3,4)	C(1,2,3,4)
nDCG@5	0.2137	0.2332
nDCG@10	0.2068	0.2127
nDCG@15	0.1956	0.1991
nDCG@20	0.1981	0.197
iP@0.0	0.4773	0.4541
iP@0.1	0.3167	0.3037
iP@0.2	0.2419	0.2048
iP@0.3	0.1368	0.1173
P@5	0.288	0.308
P@10	0.276	0.268
P@15	0.2587	0.2453
P@20	0.251	0.237
ERR	0.09869	0.09993

7 ま と め

本研究では，ユーザの多様な情報要求に対応するために，複数の検索モデルで求めた各文書の適合度を確率モデルであるコピュラを用いて，情報検索の高精度化を実現する統合方式を提案し，その統合方式の評価を行った．そして分散共分散行列の固有値の合計を考えることでコピュラで統合したときの精度，線形結合で統合したときの精度を推定でき，統合方法を選択できるとわかった．この統合方法を Greedy-integrate と名付けた．これにより，検索モデルのスコアが三つあるいは四つの場合でも統合方法を自動的に判断して高精度化につなげることができた．一般に組み合わせ最適化問題において，Greedy アルゴリズムは近似解を与えることができる．今回の実験では，Greedy-integrate により求めた解が最適解に近い結果になった．

今後の課題について述べる．課題の一つ目としては，検索モデル二つのパターン数を増やして図 6 の範囲外となるケースを避け，推定精度を求められるようにすることである．課題の二つ目としては，検索モデルのスコアが五つ以上の場合でも，Greedy-integrate が有効であることを実験で検証していくことである．課題の三つ目としては，Greedy-integrate が必ずしも統合後の検索精度が最も良くなるとは限らないと考えられるので，Greedy-integrate 以外の方法を検討してみることである．これらによってさらなる情報検索精度の向上に努めていきたい．

謝　　辞

本研究の一部は，JSPS 科研費 (18H03242, 18H03342, 19H01138A) の助成を受けたものである．

文　　献

[1] Pia Borlund. The concept of relevance in IR. *Journal of the American Society for information Science and Technology*, Vol. 54, No. 10, pp. 913–925, 2003.

[2] Ronan Cummins. Measuring the ability of score distributions to model relevance. In *Information Retrieval Technology*, pp. 25–36. Springer, 2011.

[3] P. Deheuvels. La fonction de dépendance empirique et ses propriétés – Un test non paramétrique d'indépendance. *Académie Royale de Belgique – Bulletin de la Classe des Sciences*, Vol. 65, No. 5, pp. 274–292, 1979.

[4] Carsten Eickhoff and Arjen P. de Vries. Modelling complex relevance spaces with copulas. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pp. 1831–1834, 2014.

[5] Carsten Eickhoff, Arjen P. de Vries, and Kevyn Collins-Thompson. Copulas for information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 663–672. ACM, 2013.

[6] Takuya Komatsuda, Atsushi Keyaki, and Jun Miyazaki. A Score Fusion Method Using a Mixture Copula. In Sven Hartmann and Hui Ma, editors, *Database and Expert Systems Applications*, pp. 216–232. Springer, 2016.

[7] Rashid Mehmood, Guangzhi Zhang, Rongfang Bie, Hassan Dawood, and Haseeb Ahmad. Clustering by fast search and find of density peaks via heat diffusion. *Neurocomput.*, Vol. 208, No. C, pp. 210–217, October 2016.

[8] Stefano Mizzaro. Relevance: The whole history. *Journal of the American society for information science*, Vol. 48, No. 9, pp. 810–832, 1997.

[9] Roger B. Nelsen. *An Introduction to Copulas*. Springer, 1999.

[10] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, pp. 109–126. NIST Special Publication 500-225, 1995.

[11] Yume Sasaki, Takuya Komatsuda, Atsushi Keyaki, and Jun Miyazaki. A new readability measure for web documents and its evaluation on an effective web search engine. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, iiWAS '16, pp. 355–362. ACM, 2016.

[12] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de Paris*, pp. 229–231, 1959.

[13] Christopher C. Vogt and Garrison W. Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, Vol. 1, No. 3, pp. 151–173, 1999.

[14] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiment And Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. MIT Press, 2005.

[15] 矢野桂司. 地理行列への直接因子分析法の適用に関する一考察. 地理学評論 Ser. A, Vol. 58, No. 8, pp. 516–535, 1985.
[16] 戸坂凡展, 吉羽要直. コピュラの金融実務での具体的な活用方法の解説. 金融研究, Vol. 24, 別冊 2, pp. 115–162, 2005.
[17] 左近健太, 欅淳志, 宮崎純. 密度ベースクラスタリングによる多峰性コピュラを用いた情報検索の高精度化. 第 10 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2018）論文集 D5-3, 2018.