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

左近　健太† 　欅　 惇志††,††† 　宮崎　 純††

† 東京工業大学情報工学科　〒 152–8552 東京都目黒区大岡山 2-12-1
†† 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
††† 国立研究開発法人科学技術振興機構，ACT-I　〒 332-0012 埼玉県川口市本町 4-1-8
E-mail: † sakon@lsc.cs.titech.ac.jp, †† keyaki@lsc.cs.titech.ac.jp, †††† miyazaki@cs.titech.ac.jp

あらまし　本稿では，複数の検索モデルの適合度を統合する際，密度ベースのクラスタリングの導入によって，多峰性コピュラに基づく情報検索の精度改善手法を提案する．従来の適合度の統合には，線形結合，ランキング学習などが使われていた．近年，コピュラを用いた統合手法が提案され，複雑な依存関係を捉えられるようになったことで，精度が向上した．本研究においては，クラスタリングの手法を距離ベースから密度ベースに変更することで，外れ値の影響を減らし，多峰性コピュラの統合手法を改良することを提案する．実験の結果，P@15，P@20，$nDCG$@5，$nDCG$@20，iP@0.1，iP@0.2，$MAiP$ で既存の手法を上回った．また，P@15，P@20，$nDCG$@20，$MAiP$ において複数の既存手法に対して統計的有意に精度が向上した．

キーワード　コピュラ，情報検索，クラスタリング，DBSCAN

1. はじめに

情報検索システムとは，大規模な情報データベースから適切な検索アルゴリズムにしたがって利用者 (ユーザ) の入力した検索要求 (クエリ) に適合するデータを返答するシステムである．近年のインターネットと Web の普及により，ユーザの求める情報をより高速かつ正確に取り出す必要性が高まっており，情報検索システムのさらなる改善が望まれている．

本研究で扱う情報検索システムは，大量の文書データベースからユーザのクエリである検索語に対する適合度 (relevance) を計算し，適合度の順に表示するものとする．ここで言う適合度とはクエリに適合している度合いを表す．適合度の計算では理論的に絶対的な手法が確立されておらず，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 [1,18]．その評価手法を研究するワークショップとして TREC(注1)がある [29]．

ただ，単一の検索モデルにより算出された適合度では，多様なユーザのクエリに適切に対応することが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る適合度を統合する手法が提案されている [3]．

例えば，複数の適合度を線形結合して統合した適合度を算出するものがある [28]．しかし線形結合は適合度間の依存関係がうまくとらえれらないという問題がある．また，機械学習による統合手法も提案されている [2] が，これには検索結果の理由づけが困難であるという課題がある．

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 [20] を利用した統合が考えられ成果が出ている [6,7]．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，各適合度間の複雑な依存関係をとらえることができると同時に，検索結果に対する解釈が可能となる．

また，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依存関係も捉えられる [16]．

ただし，既存手法では距離ベースのクラスタリングを行っていたので，適合度の分布に外れ値がある場合，精度が低下する可能性があった．本研究では，この問題を解決するため，密度ベースのクラスタリングを行うことによる混合コピュラを用いた手法を提案する．これにより，クラスタリングにおいて外れ値の影響を排除することができ，より精度の高い適合度の統合を行うことが期待できる．また，複数のクラスタリング結果を合成して，不適合文書をできるだけ排除するようにクラスタ領域の再構築を行うことで，さらなる精度向上を目指す．

本稿の以降の構成を述べる．2. 節でコピュラ，3. 節でクラスタリングの概要を述べ，4. 節で関連研究を述べる．5. 節で密度ベースのクラスタリングを用いた多峰性コピュラによる適合度の統合手法を提案し，6. 節で提案手法の評価実験について述べる．最後に 7. 節でまとめと今後の課題を述べる．

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，複数の周辺分布を一つの同時分布に写像する関数のことである．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 [27]．(スクラーの定理)

(注1)：http://trec.nist.gov/

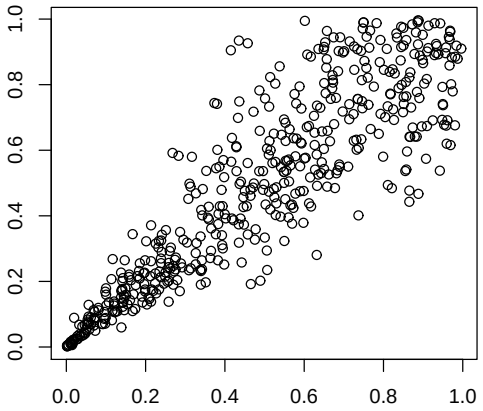

図 1　クレイトンコピュラに従った分布の例

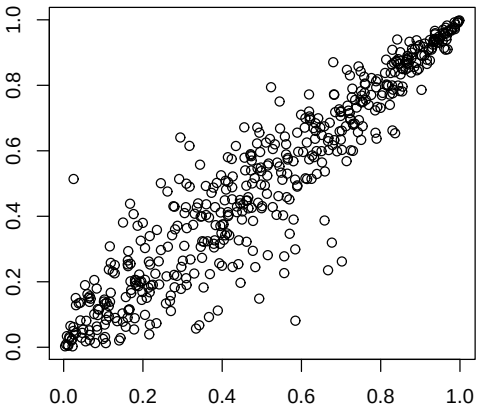

図 2　グンベルコピュラに従った分布の例

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

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

2.2.1　アルキメデス型コピュラ

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．以下に示す積コピュラ，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．

- 独立コピュラ (積コピュラ)

$$C_{indep}(U) = \prod_{i=1}^{n} u_i$$

で与えられるコピュラである．u_i に依存関係がない場合独立コピュラとなる．

- 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

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

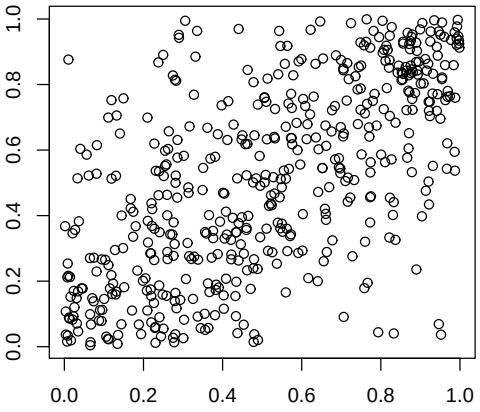

図 3　フランクコピュラに従った分布の例

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

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

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．

2.2.2　楕円型コピュラ

ガウス分布や t 分布のような標準的な分布から導かれたコピュラを楕円コピュラと呼ぶ．

$$C_{Gaussian}(U) = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n))$$

を満たす $C_{Gaussian}$ がガウシアンコピュラ (正規コピュラ) である．ここで Φ_Σ は相関行列 Σ を持つ n 変数標準正規分布の累積密度関数を，Φ^{-1} は 1 変数標準正規分布の累積密度関数の逆関数を表す．同様に t 分布から作られるコピュラを t コピュラと呼ぶ．

2.2.3　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 [5]．経験コピュラは以下の式により与えられる [31]

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

N は学習データの数，t_i^k は N 個の n 変量データ $(x_1^1,\ldots,x_n^1)$, $\ldots$, $(x_1^N,\ldots,x_n^N)$ が観測されたとき，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．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. クラスタリングの概要

3.1 クラスタリング

クラスタリングとはラベルなしデータ $D=\{x_1,\ldots,x_m\}$ を k 個の “性質の近い” 集団であるクラスタ $\{C_j|j=1,\ldots,k\}$ へとグループ分けすることである [14]．ただしクラスタ同士には重なりはなく，基本的にすべてのデータはいずれかのクラスタに属する．クラスタリングには “性質の近さ” をどのように定義するかによってさまざまな手法が提案されているが，主なものには距離ベースである分割的手法と階層的手法，密度ベースの手法がある [30,33]．

3.2 分割的手法

“性質の近さ” をデータの間の距離として計算し，一定の基準に基づいて最適な k 個のクラスタに分割する手法である．もっとも有名な手法として k 平均法 (k-means) がある [17]．ただし，k 平均法は，あらかじめクラスタ数を与える必要がある．

3.3 階層的手法

階層的手法はトップダウンの分割型とボトムアップの凝集型に分かれる．分割型は一つのクラスタを徐々に分解していくものである．一方，凝集型はまず各データを一つのクラスタとみなし，距離の近いものから順に一つずつ結合していくものである [4]．

クラスタ間の距離の定義により代表的なものとして以下のような手法がある [32]．なお，x_1 と x_2 の距離を $dist(x_1,x_2)$，クラスタ C_1 と C_2 の距離を $dist(C_1,C_2)$ とする．

- 最短距離法：各クラスタに含まれる点の距離のうち，最短のものをクラスタ間の距離とする．

$$dist(C_1,C_2) = \min_{x_1\in C_1, x_2\in C_2} dist(x_1,x_2)$$

- 最長距離法：各クラスタに含まれる点の距離のうち，最長のものをクラスタ間の距離とする．

$$dist(C_1,C_2) = \max_{x_1\in C_1, x_2\in C_2} dist(x_1,x_2)$$

- 群平均法：各クラスタに含まれるすべての点間の距離の平均をクラスタ間の距離とする．

$$dist(C_1,C_2) = \frac{1}{|C_1||C_2|}\sum_{x_1\in C_1}\sum_{x_2\in C_2} dist(x_1,x_2)$$

3.4 密度ベースの手法

この手法ではデータの密度が高い部分を一つのクラスタとみなす．DBSCAN [8] は代表的な密度ベースのクラスタリング手法であり，$(\varepsilon, MinPts)$ をパラメータとしてクラスタリングを行う．ある点 x から距離 ε 以内の点が $MinPts$ 個以上あれば，それらは一つのクラスタを形成する．これをすべての点に対して行い，より大きなクラスタを形成する．どのクラスタにも属さない点を外れ値とする．図 4 は DBSCAN を用いたクラスタリングの概略を示している．$x1$，$x2$，$x3$，$x4$ はクラスタの中心を表しており，最終的なクラスタは図 4 右のようになる．

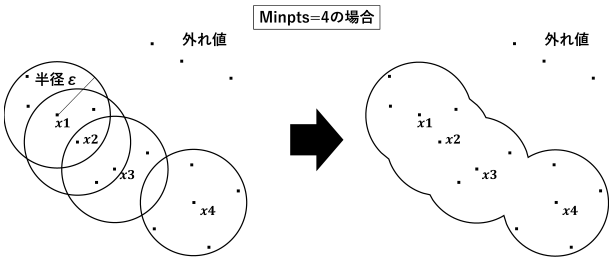

図 4　DBSCAN の概略図

DBSCAN は事前にクラスタ数を与える必要がなく，任意の形状のクラスタを見つけることができる．また外れ値 (ノイズ) を排除することができるので，外れ値を含むデータのクラスタリングに強みがある．

4. 関 連 研 究

4.1 情報検索モデル

文書の適合度を算出するモデルとして，従来は重みづけ出現頻度ベースの TF-IDF 法が用いられていた [15,25]．近年，古典的確率モデルである BM25 [24] や確率的言語モデルであるクエリ尤度モデル [23] が提案され，TF-IDF 法より高精度の結果が得られている．

4.2 適合度の統合

適合度の統合について，初期には単に適合度を加算する方法がとられた [9]．また，二つの順位付けから全体の順位付けを求める選好投票を利用した多数決方式もとられた [19]．

その後，Vogt らにより複数の適合度を線形結合して統合した適合度を算出する手法が提案され，精度の向上が得られた [28]．線形結合は複数の適合度が独立である場合は一定の効果があるが，実際にはそれらの間に依存関係がある場合が多く，この依存関係が線形結合ではとらえられない結果有効性が低くなる．Gerani らは非線形変換を行った後に線形結合を行うことにより，一定の精度改善を行った [10]．

また，他の方法として機械学習による統合手法も提案されている [2,11–13,21,22]．この手法では精度の向上が得られるが，どのようにしてその結果が出たのかを人間が適切に理由づけすることができないという課題がある．

4.3 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 [6,7]．適合度の統合にコピュラを用いることにより，各適合度間の複雑な依存関係をとらえることができると同時に，結果の理由づけが可能となる．

図 5 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，単峰

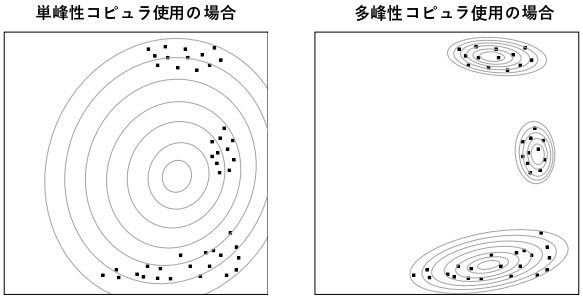

図 5　単峰性コピュラと多峰性コピュラの違い

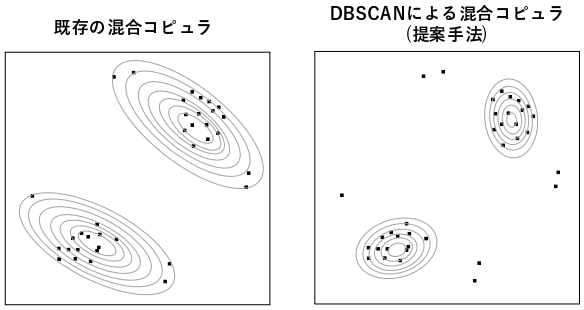

図 6　既存の混合コピュラと DBSCAN による混合コピュラの比較

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 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 [16]．また，Sasaki 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 [26]．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．

5. 提 案 手 法

5.1 提案手法の概要

混合コピュラを用いた適合度の統合において，Komatsuda らは階層的手法の群平均法によるクラスタリングを用いていた．しかしこの手法では，外れ値もクラスタに含めてしまうため，データに外れ値が多いと精度が悪くなってしまう．そこで本研究では密度ベースのクラスタリング手法である DBSCAN を用いた手法を提案する．これにより外れ値の影響が少なくなり，精度の向上が見込まれる (図 6)．

ただし DBSCAN では，不適合文書の分布を考慮していないため，不適合文書を多く含むクラスタが生成され，精度が低下する可能性がある．そこで，第二の提案手法として，異なる複数の ε ごとに DBSCAN で求めたクラスタ領域すべてを合成して，不適合文書をできるだけ含まないように再構築を行う．以降，これを拡張 DBSCAN と呼ぶ．これにより，不適合文書をできるだけ排除しつつ適合文書を多く抽出することを目指す．

提案手法の概要は以下のとおりである．クラスタリング手法は密度ベースの DBSCAN やその拡張を利用するという点を除いて，Komatsuda らの手法に準拠する．

（1） DBSCAN，または拡張 DBSCAN による適合文書のクラスタリングを行う．

（2） 各クラスタごとに，周辺分布とコピュラのパラメータの推定を行う．

（3） 推定した周辺分布とコピュラから混合コピュラを構築する．

（4） 混合コピュラを用いて，統合モデルを構築する．

5.2 拡張 DBSCAN

拡張 DBSCAN では，クラスタに含まれる不適合文書をできるだけ排除することによって，DBSCAN よりも精度の向上を目指す．

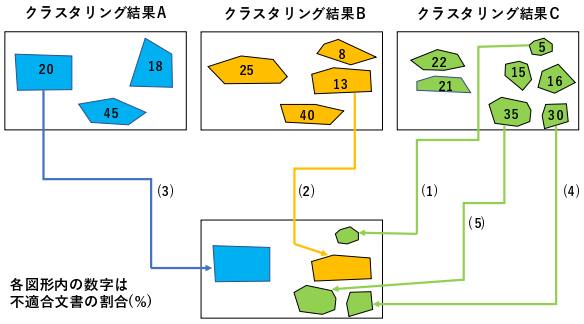

図 7　拡張 DBSCAN の再構築イメージ図

クラスタに属する点の集合を $\{(x_1, y_1), (x_2, y_2), \cdots, (x_k, y_k)\}$ とすると，拡張 DBSCAN のクラスタ領域は，それらの点をすべて含む凸包とする．凸包を考えることで余分な領域を含めないようにする．

ここで不適合文書の割合を，クラスタ領域の凸包内部に含まれる全文書数に対する不適合文書数の割合と定義する．拡張 DBSCAN では，不適合文書の割合が少ない順にクラスタ領域を追加していく greedy アルゴリズムを使用した．図 7 に，クラスタ領域の再構築のイメージを示す．図中のクラスタリング結果 A，B，C は同一データに対して ε の値を 3 通りに変えてそれぞれクラスタリングを行った三つの結果を示している．まず不適合文書の割合が最も小さいクラスタリング結果 C の (1) が採用され，次に残ったクラスタ領域の中で既に追加されたクラスタ領域と重ならず，不適合文書の割合が最も小さいクラスタリング結果 B の (2) が採用される．このようにして順次 (3)，(4)，(5) が追加される．

5.3 周辺分布とコピュラの推定

これ以降の処理は Komatsuda らの手法と同様である．まずクラスタ毎に周辺分布を推定した後，コピュラを推定する．周辺分布は任意に選択可能であり，適合度の分布を考慮して適切に選択する．例えばガウス分布の累積分布関数や経験分布の累積分布関数が選択できる．(コピュラを用いるためその入力は累積分布関数とする必要がある)．

経験分布の場合の累積分布関数は以下となる．

$$\hat{F}(x) = \frac{1}{N} \sum \mathbf{1}\{X_i \leqq x_i\}$$

N は経験分布関数の推定に用いた学習データ，$X_i (i = 1, \ldots, N)$ は適合文書の適合度を表す．

コピュラは 2. 節で述べたようにいくつかの種類があるが，周辺分布の依存関係を適切に表せるものを選択する．アルキメデス型コピュラのパラメータ θ の推定には最尤推定を用いる．

5.4 混合コピュラによる統合モデルの作成

二つのモデルを使用し，実際のデータで検証を行う．

一つ目の統合モデルは，以下のように混合コピュラの累積分布関数 C_{mix} を利用するものである．k はコピュラの個数，$C_i(1 \leqq i \leqq k)$ はコピュラ，重み p_i は，全適合文書のうちクラスタ i に属する適合文書の割合とする．

$$C_{mix}(U) = \sum_{i=1}^{k} p_i C_i(U)$$

二つ目の統合モデルは，以下のように U の尤度と混合コピュラの累積分布関数 C_{mix} の積を用いるものである．ただし U の尤度はその成分同士が独立な確率変数であると仮定している．

$$C_{mix-prod} = C_{mix}(U) \prod_{i=1}^{k} u_i$$

いずれも混合コピュラによりクラスタに分解した周辺分布間の複雑な依存関係をとらえることができる．

6. 実　　験

二種類の検索モデルのスコアを統合する際，提案手法が既存の手法や Komatsuda らの手法に対して優位性を持つことを検証する．

6.1 検索モデルと評価指標

本実験で対象とするデータセットは ClueWeb09-b(注2) であり，約 4,400 万件の Web 文書から構成される．統合する 2 種類の検索モデルは，別途算出された BM25 [24] とクエリ尤度モデル [23] を用いる．

精度評価には Precision，nDCG(normalized Discounted Cumulative Gain)，補間精度 (Interpolated Precision)，平均補間精度 (Mean Average Interpolated Precision)，ERR(Expected Reciprocal Rank) を用いた．

表 1　Precision・Recall

	検索された	検索されない
適合する	A	B
適合しない	C	D

- Precision・Recall (表 1 を参照のこと．)

Precision とは検索された文書のうち適合する文書の割合を示す．

$$P = \frac{A}{A+C}$$

Recall とは適合する文書のうち検索された文書の割合を示す．

$$R = \frac{A}{A+B}$$

$P@k$ は上位 k 件のうちの適合文書の割合を表す．

(注2)：boston.lti.cs.cmu.edu/Data/clueweb09.

- nDCG

$nDCG$ は，適合文書の順位が上位であるほど高くなるように設計された評価尺度であり，以下の式で定義される．ここで $iDCG$ は $DCG@k$ が取りうる最大の値であり，rel_i は上位 i 件目の文書の適合性を 0 か 1 の 2 値で表した変数である．

$$nDCG@k = \frac{DCG@k}{iDCG@k}$$

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

- 補間精度

補間精度 (Interpolated Precision) とは，Recall が i 以上における Precision の最大値を表す．

$$iP@i = \max_r \{P@r | R@r \geqq i\}$$

また，平均補間精度 (MAiP) は，iP@0.0，iP@0.1，...，iP@1.0 の平均である．

$$MAiP = \frac{\sum_{i \in \{0.0, 0.1 \ldots 1.0\}} iP@i}{11}$$

- ERR

ERR とは，逆数順位 $\frac{1}{r}$ を重みとした期待値であり，

$$ERR = \sum_{r=1}^{n} \frac{1}{r} R_r \prod_{i=1}^{r-1} (1 - R_i)$$

で与えられる．ただし R_i はユーザが i 番目の文書を選ぶ確率であり，

$$R_i = \frac{2^g - 1}{2^{g_{max}}}$$

となる．g は適合度の集合であり，$g \in \{g_{min}, \cdots, g_{max}\}$ である．

6.2 比 較 手 法

線形結合，独立コピュラおよび Komatsuda らによるクラスタリングに群平均法を用いた混合コピュラの手法を比較対象とする．以下で確率変数ベクトル X の成分 x_i は，適合度を正規化した値を，確率変数ベクトル U の成分 u_i は x_i を累積分布関数 $F_i()$ で写像した値を表している．

まず線形結合についてであるが，その式は以下のように表される．

$$LIN(X) = \sum_{i=1}^{k} \lambda_i x_i$$

線形結合の重み λ_i は 0.05 から 0.95 まで 0.05 ずつ動かし，学習データに対する最適値 0.95 を採用した．

次に独立コピュラについてである．独立コピュラの詳細は 2. 節を参照されたい．

6.3 提案手法におけるコピュラの決定

提案手法のモデルはコピュラの種類，周辺分布の種類，ε の値から構成され，その最適なモデルは表 2 に示す通りであり，決定方法は以下に記述する．また，表 3 はクラスタの再構築を行った拡張 DBSCAN の場合のものであるが，この場合はクラスタごとに ε の値は異なるため省略する．コピュラの種類

表 2　最適なモデルの決定

手法	周辺分布	コピュラ	ε
混合コピュラ (C_{mix})	ガウス分布	経験コピュラ	0.008
混合コピュラ × 尤度 ($C_{mix-prod}$)	経験分布	経験コピュラ	0.01

表 3　最適なモデルの決定 (拡張 DBSCAN の場合)

手法	周辺分布	コピュラ
混合コピュラ (C_{mix})	経験分布	経験コピュラ
混合コピュラ × 尤度 ($C_{mix-prod}$)	経験分布	経験コピュラ

には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，ガウシアンコピュラ，経験コピュラの 5 種類が存在する．周辺分布の種類には経験分布，ガウス分布の 2 種類が存在する．DBSCAN では，ε の値を 0.07(多峰的な同時分布となる上限値) から 0.005 まで変化させて評価を行い，精度が最大となる ε の値を特定した．拡張 DBSCAN については，ε の範囲が 0.005 から 0.07 のときの DBSCAN のクラスタ結果を用いて，クラスタの再構築を試みた．

最適なモデルを決めるにあたっては $P@k(k = 5, 10, 15, 20)$，$nDCG@k(k = 5, 10, 15, 20)$，$iP@i(i = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5)$，$MAiP$，$ERR$ の計 16 個の指標において各々の精度の最大値を求め，その最大値を最も多く含むコピュラの種類，周辺分布の種類，ε の組み合わせを採用した．なお，$C_{mix-prod}$ については ε の値を変化させても精度が変わらなかったので，$\varepsilon = 0.01$ として評価を行った．

6.4 実験結果

既存手法，Komatsuda らの手法 (既存 C_{mix}，既存 $C_{mix-prod}$) と DBSCAN による手法 (提案 C_{mix}，提案 $C_{mix-prod}$)，拡張 DBSCAN による手法 (提案 (拡張)C_{mix}，提案 (拡張)$C_{mix-prod}$) を用いた際の Precision，nDCG(normalized Discounted Cumulative Gain)，補間精度 (Interpolated Precision)，平均補間精度 (Mean Average Interpolated Precision)，ERR を表 4 に示す．各指標に対して最も数値の高いものを太字で示している．

表 4 内のシンボル $*$，†，§ はそれぞれ線形結合，独立コピュラ，既存 $C_{mix-prod}$ に対して統計的有意に精度が向上したことを示す．有意水準 1%で統計的に有意なものは $**$ のように同じシンボルの連続で表現した．また，有意水準 5%で統計的に有意なものは $*$ のように単一シンボルで表現した．

以下 DBSCAN による提案 C_{mix} について考察する．Precision に関しては上位 15 件 (P@15)，上位 20 件 (P@20) の文書では最も値が高くなっている．また P@15 では線形結合，既存 $C_{mix-prod}$ に対して統計的有意に精度が向上した．nDCG に関しては上位 5 件 (nDCG@5)，上位 15 件 (nDCG@15)，上位 20 件 (nDCG@20) の文書では値が高くなっており，特に nDCG@20 では線形結合，既存 $C_{mix-prod}$ に対して有意水準 1%で統計的有意に精度が向上している．P@15，nDCG@15，P@20，nDCG@20 が高いことから，検索の上位 15，20 件については DBSCAN による提案 C_{mix} の精度は高くなると考えられる．iP に関しては iP@0.1，iP@0.2 で他の手法よりも値が高くなっており，適合文書を 10%から 20%取得する際に有効であるといえる．MAiP に関しては他の手法よりも値が高くなっており，線形結合，既存 $C_{mix-prod}$ に対して統計的有意に精度が向上している．ERR に関しては線形結合，既存 Cmix，既存 Cmix-prod よりも高い値をとっている．線形結合，既存 $C_{mix-prod}$ に対しては統計的有意に精度が向上している．

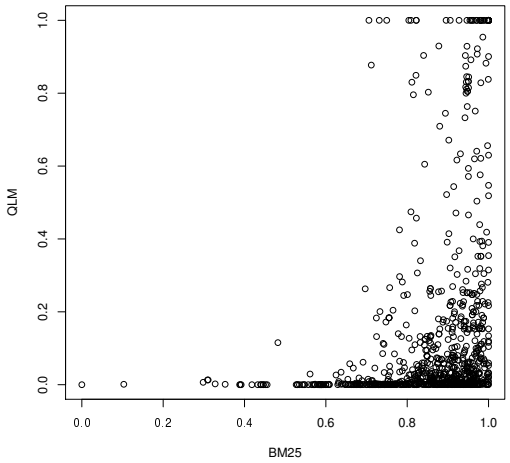

図 8　クラスタリング前の状態

図 8 はクラスタリング前の適合文書のプロット図である．また，図 9 はクラスタリング後の適合文書のプロット図であり，クラスタリングの様子をプロットされた図形と色で表している．両者を比較すると，DBSCAN により多くのクラスタに分割され，外れ値も除去されている．これが混合コピュラの精度向上に貢献したと考えられる．

次に，拡張 DBSCAN について考察する．拡張 DBSCAN による手法は，P@5，nDCG@10，iP@0.0 において提案 C_{mix} を上回っていて，上位数件の文書に関しては拡張 DBSCAN による効果があるといえる．また，ERR については提案 C_{mix} よりも値が高くなっている．

総合的に見ると，拡張 DBSCAN は提案 C_{mix} をしのぐほどではなかった．この理由としては，拡張 DBSCAN では複数の小さなクラスタ領域に分かれる傾向があるので，適合文書をとりこぼしているからだと考えられる．

図 10 は拡張 DBSCAN による適合文書のプロット図である．図 9 と比較すると，拡張 DBSCAN では，BM25 のスコアが高いときのクラスタ領域が抽出できているが，クエリ尤度モデルのスコアが低いときのクラスタ領域は抽出できないものもある．これは図 11 からもわかるように，クエリ尤度モデルのスコアが低いところに不適合文書が多く存在するので，その部分が排除されている．これによって，もとの DBSCAN との差が出ていると考えられる．

7. まとめ

本研究では，密度ベースクラスタリングである DBSCAN を用いた混合コピュラにより，適合度を統合する手法を提案し，その有効性を検証した．評価実験により，既存の距離ベースクラスタリングを利用した混合コピュラに基づく手法よりも有効

表 4　DBSCAN・拡張 DBSCAN の実験結果

手法	BM25	クエリ尤度モデル	線形結合	独立コピュラ	既存 C_{mix}	既存 $C_{mix-prod}$	提案 C_{mix}	提案 $C_{mix-prod}$	提案 (拡張) C_{mix}	提案 (拡張) $C_{mix-prod}$
P@5	0.248	0.124	**0.264**	0.256	0.252	0.2	0.252§§	0.252§§	0.256§§	**0.264**§§
P@10	0.226	0.116	0.236	0.25	**0.256**	0.228	0.242	0.248	0.244	0.24
P@15	0.2147	0.1173	0.224	0.2413	0.2387	0.2187	**0.244***§§	0.2387§	0.224	0.224
P@20	0.215	0.119	0.222	0.219	0.227	0.211	**0.234**†§	0.227†	0.224	0.219
nDCG@5	0.1668	0.0773	0.1774	0.1829	0.1774	0.1346	**0.1851**§§	0.1801§§	0.1816§§	0.1839§§
nDCG@10	0.1564	0.0802	0.1675	**0.1853**	0.1842	0.1548	0.1791§§	0.1836§§	0.1851§§	0.1803§§
nDCG@15	0.154	0.0831	0.1632	**0.1804**	0.1778	0.1534	0.1801*§§	0.1793*§§	0.1751§§	0.1744§§
nDCG@20	0.1596	0.0857	0.1683	0.1755	0.1778	0.1559	**0.182****§§	0.1787§§	0.1799§§	0.1761§§
iP@0.0	0.3846	0.1957	0.3956	0.4084	**0.4193**	0.3637	0.4065§§	0.4075§§	0.4148§§	0.4138§§
iP@0.1	0.2751	0.1168	0.2793	0.2877	0.2836	0.2611	**0.2932**§§	0.2883§	0.2733	0.2739
iP@0.2	0.1996	0.0677	0.2008	0.2096	0.1944	0.1791	**0.21**§	0.206§	0.1919	0.203§
iP@0.3	0.0841	0.0405	0.0844	**0.0973**	0.0893	0.0888	0.0955*	0.094	0.0867	0.0902
iP@0.4	0.0348	0.0183	0.035	0.0384	**0.0402**	0.0387	0.0397*	0.0395	0.0358	0.0374
iP@0.5	0.0139	0.0047	0.0141	0.012	0.0149	**0.0181**	0.0159	0.0138	0.013	0.013
MAiP	0.0912	0.0405	0.0927	0.0963	0.0952	0.0871	**0.0973***§§	0.0958§§	0.0927	0.0943§§
ERR	0.08205	0.04403	0.08644	0.09483	0.09183	0.0772	0.09335*§§	0.09495*§§	**0.09507***§§	0.09379*§§

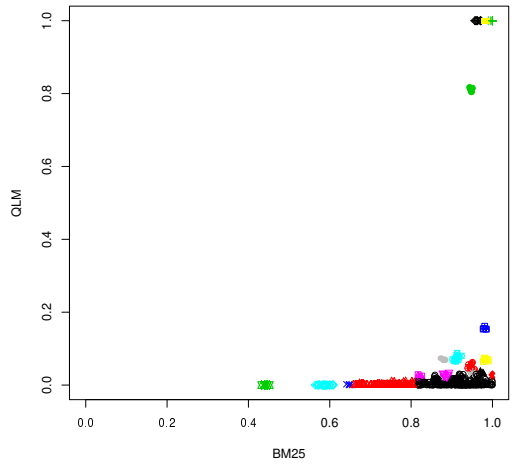

図 9　DBSCAN($\varepsilon = 0.008$) によるクラスタリング結果

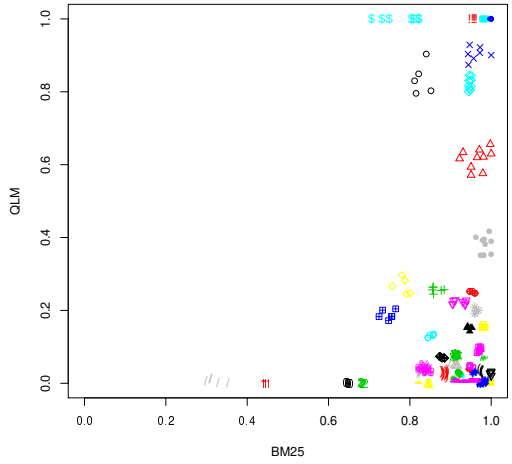

図 10　拡張 DBSCAN のクラスタリング結果

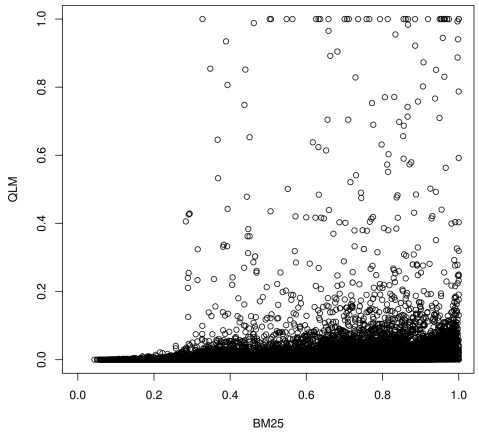

図 11　不適合文書のプロット図

であることを示すことができた．これは DBSCAN による外れ値の除去が効果的に働いていると考えられる．

また，クラスタ領域を再構築した拡張 DBSCAN では，一部の指標において DBSCAN を上回ったが，全体としては精度が向上したとはいえない．これは不適合文書を排除することによって，不適合文書の割合が少ないクラスタ領域が抽出できたが，逆に不適合文書の割合が多いものの精度向上に貢献するクラスタ領域が排除されたため，それほど精度向上が見られなかったと考えられる．

今後の課題として (1)DBSCAN のパラメータ ($\varepsilon, MinPts$) の最適値をより効率的に求める方法を検討する．また，(2)DBSCAN 以外のクラスタリング方法も検討予定である．その他，(3) ユーザの要求をより正確に捉えるため，三つ以上の検索モデルの統合を行う予定である．

謝　　辞

本研究の一部は，JSPS 科研費 (JP15H02701，JP16H02908，JP15K20990，JP17K12684)，JST ACT-I の助成を受けたものである．ここに記して謝意を表す．また，本研究の実験をご支援をいただいた，株式会社日立製作所の小松田卓也氏，佐々木夢氏に心より感謝の意を表す．

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] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96. ACM, 2005.

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

[4] William H. E. Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, Vol. 1, No. 1, pp. 7–24, 1984.

[5] 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.

[6] 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.

[7] Carsten Eickhoff, Pavel Serdyukov, and Arjen P De Vries. A combined topical/non-topical approach to identifying web sites for children. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 505–514. ACM, 2011.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231. AAAI, 1996.

[9] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In Donna K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pp. 243–252. NIST Special Publication 500-215, 1994.

[10] Shima Gerani, ChengXiang Zhai, and Fabio Crestani. Score transformation in linear combination for multi-criteria relevance ranking. In *Advances in Information Retrieval*, pp. 256–267. Springer, 2012.

[11] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 55–64. ACM, 2016.

[12] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. A position-aware deep model for relevance matching in information retrieval. *CoRR*, Vol. abs/1704.03940, , 2017.

[13] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. Position-Aware Representations for Relevance Matching in Neural Information Retrieval. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 799–800. International World Wide Web Conferences Steering Committee, 2017.

[14] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, Vol. 31, No. 3, pp. 264–323, 1999.

[15] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21, 1972.

[16] 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.

[17] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, Vol. 28, No. 2, pp. 129–137, 1982.

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

[19] Mark Montague and Javed A Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 538–548. ACM, 2002.

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

[21] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. A study of matchpyramid models on ad-hoc retrieval. *CoRR*, Vol. abs/1606.04648, , 2016.

[22] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. Deeprank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 257–266. ACM, 2017.

[23] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275–281. ACM, 1998.

[24] 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.

[25] Gerard M. Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, Vol. 24, No. 5, pp. 513–523, 1988.

[26] 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.

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

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

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

[30] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 2012.

[31] 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.

[32] 敏弘神蔦. データマイニング分野のクラスタリング手法 (1) : クラスタリングを使ってみよう! 人工知能学会誌, Vol. 18, No. 1, pp. 59–65, 2003.

[33] Zhi-Hua Zhou 著, 宮岡悦良, 下川朝有訳. アンサンブル法による機械学習 – 基礎とアルゴリズム –. 近代科学社, 2017.