様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１３　日現在

機関番号：　１２６０８
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７０１

研究課題名（和文）コピュラに基づく確率的な情報検索・情報推薦システムの実現と高精度化

研究課題名（英文）High Precision Information Retrieval and Recommendation based on Copulas

研究代表者
宮崎　純（Miyazaki, Jun）

東京工業大学・情報理工学院・教授

研究者番号：４０２９３３９４

交付決定額（研究期間全体）：（直接経費）　　13,800,000円

研究成果の概要（和文）：本研究では、コピュラを情報検索や情報推薦分野へ適用し、複数の指標間の複雑な因果関係を捉え、検索や推薦結果が説明可能かつ高精度な情報検索、情報推薦システムの構成方法を示した。
具体的には、複数のコピュラ関数を線形結合した混合コピュラモデルを応用し、良い混合コピュラを構成するために密度ベースクラスタリングを利用することで、検索ならびに推薦の高精度化が可能であることを示した。また、非線形や非単調なスコア関数でも、検索結果の上位k件を効率良く計算可能なアルゴリズムを開発した。情報推薦についても、各特徴パラメタを統計的な手法により前処理し、コピュラにより高精度の推薦が可能であることを示した。

研究成果の概要（英文）：In this research, we applied copulas which can consider complex dependencies among multiple features to the areas of information retrieval (IR) and recommender systems, and showed a method to design transparent and highly effective IR and recommender systems.
More specifically, we considered a mixture copula model which integrates multiple copulas with a linear combination for building effective IR and recommender systems. To estimate a good mixture copula which affects their effectiveness, we indicated that it is appropriate that a density-based clustering algorithm is applied in the copula estimation phase. In addition, we also developed an efficient top-k algorithm for quickly returning relevant results even if the scoring function is non-linear, such as copulas, and non-monotonic. Moreover, as for recommender systems, we showed that effective recommender systems can also be designed with mixture copulas, when preprocessing feature parameters with a statistical approach.

研究分野：　データ工学

キーワード：　情報検索　コピュラ　情報推薦　スコア統合

## １．研究開始当初の背景

近年の情報検索や情報推薦では、適合性の指標は、検索キーワードとテキスト間、すなわちトピックの適合度を表現するスコア値、また推薦アイテム等は存在の有無の二値でしか表現できず、それらの間の関連度や信頼度等は、様々な重み付け法により表現していた。このため、ユーザの情報要求や推薦要求に対して、トピックの適合度以外に、例えばコンテンツの新鮮度やテキストの複雑さ、文書構造などの指標を加味した全体の適合度は、それぞれのスコアの重み付き線形和などにより一意に決定され、それぞれの指標間の相互作用を適切に表現できず、検索結果の精度に影響を及ぼしていた。他にも機械学習手法を利用して適合度を計算する方法も提案されているが、検索結果や推薦結果に対する説明が困難という問題がある。

一方、データの状態やユーザの状況を考慮し、情報検索や推薦を行う試み、例えばデータの新しさに応じて出力スコア値を上げる、またユーザの状況に関しては特定の状態かどうかを 0/1 の二値とし、検索や推薦モデルの外でフィルタとして扱うのが一般的であり、システム全体の不透明さが問題となっていた。つまり、機械学習による方法と同様、検索結果や推薦結果導出過程がブラックボックス化されており、入力となる複数のスコアと、出力である全体のスコア値との因果関係が直観的に理解できない問題があった。

出力である検索や推薦結果の適合度のスコア値の一意性を緩めた Zaragoza らのクエリ尤度モデルでは、スコアは確率分布で表現し検索結果の尤もらしさを分布として直観的に表現できる。入力指標に関しても分布で表現する方が、より柔軟かつ自然であり、かつ良い検索・推薦結果が得られる可能性は非常に高い。データの新鮮さ以外にも、例えば商品の推薦において、ユーザにより商品が 5 段階で評価されている場合、平均値を推薦システムの入力とするのではなく、評価の度数分布を確率分布に変換して表現する方が評価の散らばりを推薦結果に直接反映することができる。また、ユーザの状況を表現する際にも人間の嗜好は曖昧であり、確率分布で表現する方が自然である。

## ２．研究の目的

センサデータベース分野では、ノイズ等をモデル化するため、既知の確率分布を入力データに付与する研究がある。一方、金融工学分野では複雑な金融商品のリスク解析のためにコピュラが注目されている。コピュラは同時分布関数とその周辺分布関数との関係を与える関数であり、債券等のリスクを複数の指標をそれぞれ周辺分布とし、これらの依存関係を同時分布として表現して解析する研究が多数ある。しかし、情報検索分野のクエリ尤度モデルでは、出力が単変量の分布のため複数の指標間の因果関係は扱えない。

本研究は、コピュラの優れた特徴を情報検索や情報推薦分野へ適用することを目的としている。情報検索分野で、複数の指標間の定性的な因果関係をコピュラで与え、各スコア値の線形結合等により得られたスコアに、同時分布の確率値をバイアスとして加味することで、総合的な適合度を計算する研究があるが、各指標間の因果関係を表現するコピュラ関数を既知として事前に与える点で本質的な問題がある。一般に、各指標の因果関係は未知であり、本研究ではそれぞれの指標の分布から、それらの因果関係を表現するコピュラ関数を推定することで、全体の適合度を同時分布としてとして直接算出することを目指す。

## ３．研究の方法

本研究では、情報検索・推薦において、コピュラを利用して入力から出力まで全てを確率分布で表現される単一モデルにより、より良い検索結果や推薦結果の導出を、因果関係を含めて提示可能とすることを目指すとともに、実用的な処理性能を得るための高速計算手法の開発、ならびに大規模なデータセットによる評価、さらにこれらの関連技術の研究を行った。具体的には、以下の研究を実施した。

1. コピュラを核とした、情報検索システムや情報推薦システムのモデル化
2. 複数のコピュラ関数の組み合わせによる複雑な確率分布のモデル化、ならびにそのコピュラ関数の推定方法
3. 非線形なコピュラ関数のための Top-k 検索アルゴリズムの開発と、効率的な検索結果の計算
4. プロトタイプシステムの実装ならびに大規模テストコレクションを利用した実験による提案手法の有効性評価
5. 情報検索や情報推薦のための効率的な統計量計算やコンテキスト情報の取り扱いとその応用

## ４．研究成果

本研究では、複雑な同時分布を持つデータを対象とするために、複数のコピュラを重み付き線形和で表現した混合コピュラを用いて、複数のモデルによる適合度を統合する手法を提案し、この手法を情報検索ならびに情報推薦に適用する研究を中心に行った。以下、(1)情報検索への適用、(2)情報推薦への適用についてそれぞれ述べる。

(1) 情報検索

既知のクエリとその適合文書から、適合文書の分布を混合コピュラにフィッティングさせ、適合文書の同時分布の推定を行うための手法が核となる。混合コピュラを用いることで、多峰的な同時分布を表現することが可能となり、局所的に強い相関を持つ領域を捉えることが可能となる。

ここでは、文書はk種類の検索モデルによって評価され、k 種類の適合度が算出されていることを仮定する。また、事前に適合していると判定されたN件の文書を学習データとして用いて、それら適合文書が生成される確率に関する多峰的な同時分布を混合コピュラによって構築することを考える。

混合コピュラは単峰コピュラの重み付き線形和として実現できる。混合コピュラを推定するアプローチとして、適合文書をクラスタリングし、各クラスタごとに対応する単峰コピュラを推定し、それらを線形和で結合するアプローチを採用した。具体的には、まず、学習データとして用いる N 件の適合文書を、距離ベースの階層型クラスタリング法を用いてP分割する。次に、クラスタごとにk種類の検索モデルについて、検索モデルによって算出された適合度の周辺分布をそれぞれ推定し、各々に推定した周辺分布を、単峰コピュラを用いて接合することで同時分布の推定が可能となる。最後に、クラスタごとに推定した同時分布を、各クラスタの適合文書数の比を重みとして線形和を構成することで混合コピュラを導く。混合コピュラを推定した後、その混合コピュラを用いて統合モデルを構成する。

ここで混合コピュラ導出時の改善方法として、適合文書をクラスタリングする際に、密度ベースクラスタリングアルゴリズムを適用することができる。これにより、適合文書群の密度の高い箇所のみを抽出し、精度を落とすノイズとなる外れ値を除外することで、検索精度の向上が期待できる。

評価実験では、TREC ClueWeb09 のカテゴリーB のうち、Wikipedia 文書を除いた文書を対象として評価実験を行った。本実験で対象とする文書数は約 4400 万件である。クエリは、TREC2011 アドホックタスクで用いられたものを利用した。実験では、BM25 とクエリ尤度モデルの二つの検索モデルによる適合度をコピュラで統合し、提案する混合コピュラを利用した手法の優位性を評価した。

その結果、ランキング上位 5 件を取得する際の検索精度、ならびに nDCG 指標において、いくつかの研究でその有効性が示されているベースラインの線形結合に対して、それぞれ 10%、15%以上精度が向上した。また、混合コピュラを用いて推定した同時分布と尤度の積で表されるモデルでは 20%の適合文書を取得するまでに、線形結合や本研究の先行研究となる Eickhoff らの単峰コピュラを用いた手法よりも 5%以上精度が向上し、有効であることを明らかにした。一方、密度ベースクラスタリングアルゴリズムを利用した改良手法では、距離ベースクラスタリングを利用した混合コピュラよりもさらに精度が上回っているケースが多く、密度ベースクラスタリングを混合コピュラ推定時に利用することが有効であることを明らかにした。特に取得する検索結果件数が多いほど精度が高くなり、これは密度ベースクラスタリングの利点である外れ値の除去が効果的に働いていると考えられる。しかしながら、密度ベースクラスタリングはクラスタの探索範囲を指定するパラメタの最適化が必要であることが問題であり、また、検索精度に大きく貢献するクラスタが存在することも判明している。このため、密度ベースクラスタリングにより、高精度検索が可能なよう自動的に優良なクラスタを選択するアルゴリズムの開発が今後の課題である。

(2) 情報推薦

情報推薦の代表的なアルゴリズムの一つであるコンテンツベースフィルタリングでは、ユーザが好むアイテムを事例として嗜好モデルを構築し、それに基づき推薦を行う。嗜好モデルは、SVM やニューラルネットワークといった機械学習手法により構築することが可能だが、これらの手法は推薦結果に対する理由付けが困難であるという問題がある。この問題に対して変数間の依存関係を捉えることが可能かつ学習結果の解釈が容易なコピュラを用いて構築した嗜好モデルに基づいてアイテムのランキング付けを行う推薦手法を提案した。

推薦対象のアイテムが持つ各特徴パラメタは、ある属性を満足するかどうかの二値の特徴パラメタや、商品の価格やユーザレビュー値などの連続値あるいは多値で表現されることが多い。二値属性の特徴パラメタは、推薦アイテムのスコアリングよりもむしろ、推薦対象アイテムのフィルタリングに利用されることが想定されるため、まず、簡単化のため、推薦対象アイテムの持つ特徴パラメタは、連続値で表現されているものと仮定する。ここで、コピュラを用いた情報検索の情報推薦への適用を考慮した場合、適合文書を適合アイテム、各適合度をアイテムの各特徴パラメタと読み替えることで、コピュラによる適合度統合手法を情報推薦のユーザプロファイリングに適用することができる。その際、情報推薦のコンテキストにおいては、全ての特徴パラメタを使用することが必ずしも適切とは限らない。文書の適合度はユーザに依存せず客観的に表現することが可能であり、高精度検索に貢献することが検証されているスコア関数や、スコア関数算出に用いられる統計量により定義される。それに対して、ユーザの嗜好はユーザごとに異なり、必ずしも全ての特徴を考慮しているわけではない。従って、特徴パラメタ全てを適合度と

して読み替えることは適切ではない。また、一般に機械学習では各特徴パラメタに重み付けを行うが、コピュラによる適合度統合式は対称であり、特定の特徴について重み付けされていない。そこで、ユーザの各特徴パラメタへの関心度を、KL 情報量を用いて定義し、関心度の値をもとに特徴パラメタの次元削減ならびに重要な特徴パラメタの検出を行った後、混合コピュラによる推薦のためのスコアリングを行う手法を提案した。具体的には、各特徴パラメタへのユーザの関心度の算出方法、関心度の分散から特徴パラメタの次元削減ならびにユーザが重要視している特徴パラメタの特定を行う手法の提案からなる。これらから、効果的な推薦スコアリング関数を、混合コピュラを利用して構成する。

さらに本推薦手法を一般化するために、離散値を取る特徴パラメタや、特異な嗜好ケース、例えば中庸な価格帯のアイテムを好むといったケースなどにも対応可能なよう、拡張を行った。離散値への対応については、カーネル密度推定におけるカーネル関数を tophat とし、カーネル関数のバンド幅を可能な限り小さくすることにより、コピュラに利用する密度関数を構成する手法を提案した。また、特異な嗜好ケースへの対応について、ユーザの関心度を高い精度で計測するために KL 情報量を改良し、KL 情報量の積分区間を変更した関心度関数を新たに定義した。この新しい関心度関数を利用して、関心の高い範囲のアイテムを抽出する許容範囲フィルタの構成方法を提案した。

東京都内のホテルのデータを利用し、被験者実験により評価を行った。その結果、提案手法のうち、次元削減を行った後に重要な特徴パラメタに重み付けを行う手法は、情報検索のための混合コピュラによるスコア統合手法を単に適用した手法に対して、推薦結果上位 20 件取得時の精度が有意水準 5%で統計的有意に向上することを示した。

さらに二値特徴パラメタや特異な嗜好を考慮する一般化のための拡張を行うことにより、過半数の評価指標で本拡張手法が最も良い結果を示した。加えて本拡張手法と拡張前の手法を比較した結果、統計的に有意に本手法が優れていることを確認した。許容範囲フィルタの採用については、それを採用しない結果に比べて、推薦精度が顕著に良いケースが多く、総合的に提案手法が優れていることを明らかにした。また、機械学習手法の一つであるランキング SVM よりも、推薦結果上位における推薦精度や、nDCG 指標について優れた性能が得られることを明らかにした。

しかしながら、許容範囲フィルタはコピュラの枠組みとは独立であり、ユーザの特異な嗜好をコピュラでモデリングすることが未達成である。これは即ち非単調な関数を、単調性を仮定するコピュラにどのように変換するかという問題に帰着し、今後の課題としたい。

コピュラによる情報検索ならびに情報推薦のモデル化に付随して、検索や推薦時のコンテキストを抽出して利用するための手法や、ユーザの検索意図の推定、位置情報やセンシング情報を検索や推薦のサービスに結び付けるための研究や、大規模な文書集合から効率よく統計量を計算するための高性能計算方式、混合コピュラによる非線形なスコア関数やその一般化である非単調な連続関数で定義されるスコア関数において、スコア値の上位 k 件を効率的に取得するためのアルゴリズムについて研究を行った。最後の上位 k 件を効率よく計算する新しい top-k アルゴリズムは、データではなくスコア関数の定義域である空間をインデキシングすることにより、動的に生成されるデータのインデキシングを必要とせず、また全件のデータをソートすることなく上位 k 件のデータを得ることを可能であることを明らかにした。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 14 件）


① Kentaro Noda, Yoshihiro Wada, Sachio Saiki, Masahide Nakamura, Kiyoshi Yasuda, Implementing Personalized Web News Delivery Service Using Tales of Familiar Framework, Proc. of IEEE International Conference on Pervasive Computing and Communications Workshop, pp.831-836, 2018, 査読有, http://www27.cs.kobe-u.ac.jp/achieve/data/pdf/1312.pdf

② Atsushi Keyaki, Jun Miyazaki, Kenji Hatano, Effective Mobile Search using Element-based Retrieval, IPSJ Transactions on Databases, Vol.10, No.3, pp.1-11, 2017, 査読有, DOI: 10.2197/ipsjjip.25.934

③ Toshiaki Wakatsuki, Atsushi Keyaki, Jun Miyazaki, A Case for Term Weighting using a Dictionary on GPUs, Proc. of DEXA 2017, pp.103-117, 2017, 査読有, DOI: 10.1007/978-3-319-64471-4_10

④ Shuhei Kishida, Seiji Ueda, Atsushi Keyaki, Jun Miyazaki, Skyline-based Recommendation Considering User Preferences, Proc. of APWeb-WAIM 2017, pp.133-141, 2017, 査読有, DOI: 10.1007/978-3-319-63564-4_11

⑤ Atsushi Keyaki, Jun Miyazaki, Part-of-speech Tagging for Web Search Queries Using a Large-scale Web Corpus, Proc. of SAC 2017, pp.931-937, 2017, 査読有, DOI: 10.1145/3019612.3019694

⑥ Hikaru Inomoto, Sachio Saiki, Masahide Nakamura, Shinsuke Matsumoto, Design and Evaluation of Mission-Oriented Sensing Platform with Military Analogy, Inter-

national Journal of Pervasive Computing and Communications, Vol.13, No.1, pp.1-17, 2017, 査読有, DOI: 10.1108/IJPCC-01-2017-0007

⑦ Yuka Teramoto, Kazuma Kusu, Takamitsu Shioi, and Kenji Hatano, Constructing a Judging Model of Closeness in Human Relations, Proc. of Culture and Computing 2017, pp.49-54, 2017, 査読有, DOI: 10.1109/Culture.and.Computing.2017.47

⑧ Yume Sasaki, Takuya Komatuda, Atsushi Keyaki, Jun Miyazaki, A New Readability Measure for Web Documents and its Evaluation on an Effective Web Search Engine, Proc. of iiWAS 2016, pp.357-364, 2016, 査読有, DOI: 10.1145/3011141.3011172

⑨ Takuya Komatuda, Atsushi Keyaki, Jun Miyazaki, A Score Fusion Method Using a Mixture Copula, Proc. of DEXA 2016, pp.216-232, 2016, 査読有, DOI: 10.1007/978-3-319-44406-2_16

⑩ Hikari Suganuma, Tkamitsu Shioi, Kenji Hatano, Constructing a Discriminant Model of Web Documents Suitable/Unsuitable for Search Result, Proc. of iiWAS 2016, pp.259-263, 2016, 査読有, DOI: 10.1145/3011141.3011204

⑪ Hiroaki Takatsuka, Seiki Tokunaga, Sachio Saiki, Shinsuke Matsumoto, Masahide Nakamura, KULOCS: Unified Locating Service for Efficient Development of Location-Based Applications, International Journal of Pervasive Computing and Communication, Vol.12, pp.154-172, 2016, 査読有, DOI: 10.1108/IJPCC-01-2016-0004

⑫ Long Niu, Sachio Saiki, Shinsuke Matsumoto, Masahide Nakamura,WIF4InL: Web-Based Integration Framework for Indoor Location, International Journal of Pervasive Computing and Communications, Vol.12, pp.49-65, 2016, 査読有, DOI: 10.1108/IJPCC-01-2016-0009

⑬ Kazuki Hagiwara, Kenji Hatano, Suggesting Sub-topics of an Issued Query Using Concept Structure, Proc. of ICADIWT 2016, pp.126-134, 2016, 査読有, DOI: 10.3233/978-1-61499-637-8-126

⑭ Hiroki Takatsuka, Sachio Saiki, Shinsuke Matsumoto, Masahide Nakamura, RuCAS: Rule-Based Framework for Managing Context-Aware Services with Distributed Web Services, International Journal of Software Innovation, Vol.3, pp.57-68, 2015, 査読有, DOI: 10.4018/IJSI.2015070105

〔学会発表〕(計 17 件)

① 朝日諒, 欅惇志, 宮崎純, 離散値特徴量や特異な嗜好ケースを考慮したコピュラを用いた推薦システム, DEIM 2018, 2018

② 左近健太, 欅惇志, 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, DEIM 2018, 2018

③ 柳本晟熙, 欅惇志, 宮崎純, GPU 上の MapReduce による大規模データ処理の最適な分割粒度の動的推定, DEIM 2018, 2018

④ 若月駿尭, 欅惇志, 宮崎純, GPU 上で辞書を利用した N-gram クエリ尤度の効率的な事前計算, DEIM 2018, 2018

⑤ 柳本晟熙, 欅惇志, 宮崎純, GPU 上の MapReduce による大規模データ処理の最適な分割粒度の動的推定, DEIM 2018, 2018

⑥ 田畑亮馬, 佐伯幸郎, 中村匡秀, 確率的位置情報算出アルゴリズムにおける実環境を考慮したシミュレーションによる特性評価, 電子情報通信学会サービスコンピューティング研究会, 2018

⑦ 岡本章平, 寺本優香, 楠和馬, 蒲原智也, 波多野賢治, アイテムの評価観点を用いたレビューの有用性判定, 電子情報通信学会総合大会, 2018

⑧ 鈴木崇弘, 欅惇志, 宮崎純, Copula を用いたユーザプロファイリング手法の提案, DEIM 2017, 2017

⑨ 佐々木夢, 欅惇志, 宮崎純, 検索対象データの事前インデックスを必要としない Top-k 検索アルゴリズムの提案と評価, DEIM 2017, 2017

⑩ 植田聖司, 欅惇志, 宮崎純, プレイリスト生成における遷移確率を用いたスコアリング手法の提案, DEIM 2017, 2017

⑪ 渡佑也, 欅惇志, 宮崎純, 中村匡秀, 多次元データに対する集約演算の効率化手法におけるデータ挿入スループットの向上, DEIM 2017, 2017

⑫ 寺本優香, 塩井隆円, 楠和馬, 波多野賢治, メール送受信者間の親疎関係判定モデルの構築, DEIM 2017, 2017

⑬ 岸田脩平, 欅惇志, 宮崎純, スカイライン演算を用いたユーザ思考を考慮した情報推薦のランキング手法の精度改善, DEIM 2016, 2016

⑭ 小松田卓也, 欅惇志, 宮崎純, 多峰性のあるコピュラを用いた文書の適合度の統合手法の提案及びその検証, DEIM 2016, 2016

⑮ 佐々木夢, 小松田卓也, 欅惇志, 宮崎純, 文書の可読性を考慮した Web 検索に関する一考察, DEIM 2016, 2016

⑯ 渡佑也, 欅惇志, 宮崎純, 中村匡秀, RDB と KVS を相互に利用した多次元データに対する集約演算の効率化, DEIM 2016, 2016

⑰ 菅沼ひかり, 塩井隆円, 波多野賢治, 情報要求を満たさない文書の判別モデ

ル構築と情報検索への活用，情報処理学会第 78 回全国大会，2016

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
宮崎　純（MIYAZAKI, Jun）
東京工業大学・情報理工学院・教授
研究者番号：40293394

(2)研究分担者
波多野　賢治（HATANO, Kenji）
同志社大学・文化情報学部・教授
研究者番号： 80314532

中村　匡秀（NAKAMURA, Masahide）
神戸大学・システム情報学研究科・准教授
研究者番号：30324859

欅　惇志（KEYAKI, Atsushi）
東京工業大学・情報理工学院・助教
研究者番号：00733958