博 士（情報科学） 谷 口 剛

学 位 論 文 題 名

# A STUDY ON CORRELATION MINING BASED ON CONTRAST SETS

（コントラストセットに基づく相関マイニングに関する研究）

## 学位論文内容の要旨

近年、大規模なデータベースから思いがけない知識を発見することを目的とするデータマイニングの研究領域において、頻出アイテム集合マイニングに関する研究が成熟期を迎え、多くの成果が報告されている。従来のデータマイニングの研究では、与えられたデータベースにおいて、頻出アイテム集合のような出現確率が高いアイテム集合に注目することが多かった。しかしながら、そのような特徴的なアイテム集合はユーザが既に知っているような当たり前な情報であることも多い。このような場合、まだ特徴として顕在化していないようなアイテム集合にも注目したいユーザも存在し得る。本論文では、例え出現確率が高くないという意味で特徴的ではないアイテム集合の組であっても、ある条件付けにより劇的に相関が変化するアイテム集合の組は潜在的に重要であるという立場に立ち、その手法の開発を行なう。

第一章では序論として、本研究の背景, 動機および概要について詳しく述べている。

第二章では、相関マイニングとコントラストセットマイニングの二つの研究を取り上げている。相関マイニングとは、与えられたデータベースにおいて相関しているアイテム集合の組の探索に関する研究である。また、コントラストセットマイニングとは、与えられた二つのデータベースにおけるアイテム集合の出現確率の差を基準に、一方のデータベースを特徴付けるようなアイテム集合を探索する研究である。本研究では、二つのデータベースにおける相関の度合いを比較するため、これらの研究と高く関連している。これらの研究との違いとしては、相関マイニングとコントラストセットマイニングは共に特徴的なアイテム集合 (の組) に注目することに対して、本研究では相関変化に注目し、高い相関を示さないような特徴的ではないアイテム集合の組でさえ注目する。また、相関マイニング、コントラストセットマイニングは共に評価尺度がアイテム集合の包含関係に関して非単調に変化し、探索対象を枝刈りすることが難しい。本研究でも、これらの研究と同様の難しさを持つ。

第三章では、潜在的に重要なアイテム集合の組に注目するための一つのアプローチを与えている。具体的には、与えられたグローバルデータベースとある条件付けによって得られるローカルデータベースにおけるアイテム集合間の相関の度合いを比較し、その違いが大きいアイテム集合の組を探索する問題を提案している。このようなアイテム集合の組のことを **DC pair** と呼ぶ。ローカルデータベースにおいて高い相関を示すわけではない **DC pair** を、潜在的に重要な関係として特に注目する。**DC pair** を探索する問題は、相関マイニングとコントラストセットマイニングの双方の

問題の難しさを併せ持つ。この難しい問題に対し、DC pair の組み合わせの候補をトップダウンに識別し、その候補を分割することにより DC pair を発見する方法を与えている。詳細に述べると、相関変化が顕著な DC pair を発見しやすくするために新たにパラメータ与え、求める DC pair を限定する。この限定された問題に対し、そのパラメータに依存した探索空間の下限を理論的に示す。実験において、DC pair の組み合わせの候補を識別するために、この探索空間の理論的下限によって、一部の探索対象を安全に枝刈りすることができたが、結局ほとんどのアイテム集合を探索しなければならないことがわかった。

第四章では、DC pair の組み合わせ候補のトップダウン探索法の問題点を議論し、DC pair の要素のボトムアップ探索法を与えている。トップダウン探索では、DC pair に分割できないような組み合わせの候補が多く識別されていた。そこで、ボトムアップに DC pair の要素の候補を識別し、その組み合わせを調べることにより DC pair を発見する方法を提案している。ここでは、要素の候補の探索空間の部分空間における相関変化の上限と下限を理論的に明らかにする。さらに、この性質を上手に利用するために、部分空間を効率よく識別する技術を利用する。実験において、トップダウン探索に比べ、ボトムアップ探索の方が、DC pair を効率的に探索できることを示している。

第五章では、潜在的に重要な DC pair のみを探索する方法を与えている。前章におけるアルゴリズムでは、ローカルデータベースにおいて高い相関を示すような DC pair が多く発見された。このような DC pair は、相関マイニングの研究でも発見できる関係である。そこで、ローカルデータベースにおける相関に関する制約を導入し、ローカルデータベースにおいて高い相関しか示さないような DC pair の要素を明らかにできる場合を示す。このことにより、要素の候補をさらに制限し、組み合わせを調べるための計算負担を軽減することができる。

第六章では、本研究の枠組みを時系列データに適用している。前章までは DC pair を探索する一般的なアルゴリズムを与え、効率的な探索の可能性を示したが、グローバルデータベースに対し、ローカルデータベースにおけるトランザクション件数は比較的に少数な場合を想定していた。本章においては、時間の変化に対し件数がそれほど変化せず、特定のタイムスタンプを持つローカルデータベースのサイズが小さくない場合にも有効な DC pair 検出法について考察し、そのための新たな枝刈規則も導入している。実験では、国勢調査のデータに対し検証を行い、実際に潜在的に重要な DC pair が発見され、時系列データに対して、さらなる応用を目指す可能性があることを示している。

第七章では、DC pair マイニングの枠組みをニュース記事の解析に適用している。近年のインターネット環境の普及にともない、様々な立場から発信された情報を利用することができるようになってきている。そこで、同じ事象について、複数の情報源が、どのように異なった情報を伝えているかを分析する。具体的には、4 つの国の新聞記事をグローバルデータベース、それぞれの国の新聞記事をローカルデータベースとする。また、ユーザが興味を示す国が複数ある場合を想定し、二つの国名を表すキーワードに対して、それぞれの国において顕著な相関変化が生じるキーワードを発見した。それぞれの国に対する記事において、興味深い論調の違いがあることがわかった。

第八章では、ローカルデータベースを得るための条件探索について論じている。本研究では、ユーザがローカルデータベースのための条件を与えることを仮定している。しかしながら、適切な条件付けをユーザ自身が認識していない状況もあり得る。したがって、グローバルデータベースの適切な条件付けの探索は重要な課題である。そのような観点から、条件探索の一つの定式化とし

て、コントラストセットに基づき与えられた時系列トランザクションデータを分割する問題を提案している。また、バイオインフォマティクスの実データに対し、本研究の応用の可能性について述べている。

第九章では、本論文の総括を与え、残された研究課題について述べている。

# 学位論文審査の要旨

主　査　教　授　原　口　　誠
副　査　教　授　田　中　　譲
副　査　教　授　有　村　博　紀

学位論文題名

# A STUDY ON CORRELATION MINING BASED ON CONTRAST SETS

## （コントラストセットに基づく相関マイニングに関する研究）

大量データの収集・蓄積とその高速な計算機処理が可能になったことから、データベースからの知識発見・データマイニングの研究が1990年代以降、活発に行われてきた。この間、多くのデータマイニングのアルゴリズムが提案されてきたが、基本的には頻出性に代表されるように、比較的に検出しやすい領域に対し特に成功を収めてきた。一方、そうした頻出パターンやルールは、多数の個別的なデータに対して成立することから、一般的であり意外性を伴わない場合も多い。何をもってデータマイニングによって検出すべきターゲットルールとするかは、問題設定にも依存するが、本研究においては、特にアイテム集合間の相関検出問題に対し、相関の度合いがそれほど高くはないアイテム集合対で、特定の局所的な部分データベースにおいて全体と比較して顕著な相関度の上昇を持つものの検出を試みている。これは、全体のデータベースにおいて、低頻度で低相関のアイテム集合対であっても、特定の時間や場所等によってデータの範囲を特定した局所的なデータベースにおいて相関が顕著に上昇する場合は、全体と対比した部分において潜在的な現象が生じている可能性を認めるものであり、探索上より困難な領域において潜在的に興味深いパターンやルールを発見するためのオリジナルな発想に基づいていると述べている。こうした性質を持つアイテム集合対を、所与のデータベースとその局所的な部分データベースに対して検出するために、探索範囲の理論的下限と上限を明らかにし、さらに、無駄なアイテム集合を枝刈するための探索ルールを実装した数種類のアルゴリズムを開発し、それらの有効性と比較を実験的に行っている。さらには、タイムスタンプ付のデータに対し、特定の時間を指定したときに顕著なパターンを発見できることを実験的に明らかにしている。

第一章では序論として、本研究の動機と概要を述べ、本論文の位置付けを行っている。

第二章では、本論文の関連研究である相関マイニングとコントラストセットマイニングについて詳しく論じ、本研究との比較を行っている。相関マイニングとは、与えられた1つのデータベースにおいて特徴的な相関を持つアイテム集合対を目標にしており、また、コントラストセットマイニングにおいては、所与の二つのデータベースに対し、一方のデータベースで特徴的なアイテム集合を検出する問題であるとしている。いずれも特徴的な相関検出を目的とするが、一方、本研究にお

いては、全体と部分という2つのデータベースにおいて、部分において必ずしも特徴的とは限らないものを探索のターゲットにしており、よって、異なる問題であると述べている。

第三章では、全体と部分における相関比により、DCペアの概念を導入し、部分において潜在的でかつ顕著な相関発見問題をDCペア検出問題として定め、さらに、DCペアからなる探索空間の構造を解析し、トップダウンに動作する理論的なアルゴリズムを提案している。DCペアの空間においては、相関比がある一定以上のものだけが興味の対象であり、この条件から、全体および部分のアイテム集合束における頻度に関する制約に基づく探索下限と上限を与えている。さらに、この理論的成果をアルゴリズムの枝刈り規則として利用することが可能であり、トップダウンアルゴリズムの形で示している。

第四章では、DCペアの組合せ候補をトップダウンに探索する前章のアルゴリズムの問題点を明らかにし、先んず、部品となるアイテム集合を検出しそれを組み合わせるボトムアップ法を与えている。具体的には、ペアの要素となりえる候補アイテム集合の相関変化の上限および下限に関する性質に基づいて探索すべき部分空間を効率よく識別する技術を利用したボトムアップ法を与え、実験的にその効果を示している。

第五章では、前章のボトムアップ法をさらに洗練化した手法を与えている。局所データベースにおいて高い相関を示す部品アイテム集合対は、特徴的なアイテム集合対の検出を試みる相関マイニングを用いて対応でき、また本研究では局所データベースにおいて特徴的な相関は興味の対象外であることに注目し、そうしたアイテム集合対の枚挙を抑制するための新たな制約を導入し、集合対の組合せに要する計算負荷をさらに軽減させるアルゴリズムに発展させ、実装実験においてその有効性を実験的に示している。

第六章では、本研究の枠組みをタイムスタンプを持つデータベースに適用するために、さらなるアルゴリズムの改良を行っている。すなわち、時間の変化に対し件数がそれほど変化せず、特定のタイムスタンプを持つローカルデータベースのサイズが小さくない場合においても有効なDCペア検出法について考察し、そのための新たな枝刈規則も導入している。実験では、国勢調査のデータに対し検証を行い、実際に潜在的に興味深いDCペアが特定の年代で検出できることを示している。

第七章では、DCペアマイニングの枠組みを、複数の情報源からのニュース記事を解析し、情報源に依存して相関が顕著に異なるキーワードの対を検出する問題に応用している。具体的には、4つの国の新聞記事をグローバルデータベース、それぞれの国の新聞記事をローカルデータベースとし、それぞれの国において顕著な相関変化が生じるキーワードの発見に成功している。国毎の論調の違いを認識する手法としての可能性について述べている。

第八章では、局所データベースを得るための条件を探索する問題の1つの定式化として、コントラストセットに基づき与えられた時系列トランザクションデータを分割する問題を提案し、バイオインフォマティクスの実データに対し、本研究の応用の可能性について論じている。

第九章では、本論文の総括を与え、残された研究課題について述べている。

これを要するに、著者はデータマイニングにおける潜在的に有意味な相関を持つアイテム集合対の発見に関する新知見を得たものであり、大規模データベースからの知識発見に対して情報科学上貢献するところ大なるものがある。よって著者は、北海道大学博士(情報科学)の学位を授与される資格あるものと認める。