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

令和　３ 年　 ６ 月 ２１ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)（一般）
研究期間：2017～2019
課題番号：１７Ｈ０１７８８

研究課題名（和文）弱閉集合の代数的構造の解明と知識発見への応用

研究課題名（英文）Properties of Weakly Closed Itemsets and their Application to Knowledge Discovery

研究代表者
山本　章博（Yamamoto, Akihiro）

京都大学・情報学研究科・教授

研究者番号：３０２３０５３５

交付決定額（研究期間全体）：（直接経費）　　11,400,000円

研究成果の概要（和文）：本研究では，２つの離散値属性間の２項関係における閉集合にノイズを認めるために，弱閉集合を，集合論を用いて定式化した上で，弱閉集合を列挙するためのアルゴリズムを構築した．閉集合にグラフ理論を用いた解釈を与えることができること範にして弱閉集合の定義を与え，閉集合の高速列挙アルゴリズムを範として列挙アルゴリズムを設計した．さらに，旅行者の経路を集めたデータに対して，弱閉集合の定義と列挙アルゴリズムを経路データ向けに修正した上で適用することにより，旅行者がよく辿る経路を弱閉集合として列挙することに成功した．また閉集合の持つ不動点意味論を，弱閉集合には一般には与えることができないことも明らかにした．

研究成果の学術的意義や社会的意義
２つの離散値属性間の２項関係における閉集合は，知識発見における意味を持つだけでなく，数学的な性質を数多く持ち，しかも高速な列挙方法が開発されるなど，知識発見において重要な概念である．しかし，ノイズを全く認めないことが実用上の障害となることもあった．そこで，閉集合に対してノイズを許容する方法が必要であるが，離散値属性を扱う際にノイズを数量的に定義することは適切とは限らない．そこで，弱閉集合をグラフ理論を範にして集合論を用いて定式化した上で，弱閉集合を列挙するためのアルゴリズムを構築した．実応用として，旅行者の経路を集めた実データから旅行者がよく辿る経路を弱閉集合として列挙することに成功した．

研究成果の概要（英文）：In this research, in order to admit noise in closed sets in a binary relation between two discrete-valued attributes, we formulated weakly closed sets using set theory and constructed an algorithm for enumerating weakly closed sets. We defined weakly closed sets based on the fact that closed sets can be interpreted using graphs. We designed an algorithm for enumerating weakly closed sets with modifying the well-known fast enumeration algorithm for closed sets. Furthermore, by modifying the definition of weakly closed sets and the enumeration algorithm to the trajectory data collected from travelers, we succeeded in enumerating the routes frequently followed by them as weakly closed sets. We also showed that the fixpoint semantics of closed sets cannot be given to weakly closed sets in general.

研究分野：知能情報学

キーワード：　知識発見　２項関係　閉集合　弱閉集合

１．研究開始当初の背景

World Wide Web などの広い意味でのデータベース内に蓄積されている大量かつ多様なデータから有用な知識を効率的に導出する知識発見手法は益々重要になっている．その中で，自然言語データにおける本文とキーワードという関係，**Web** ページ間のリンク構造の場合は参照元と参照先の関係ように，２つの離散値属性間の２項関係は基本的かつ容易に構築できるデータである．一つの２項関係(例えば図 **1(1)**の●印)において有用な部分関係を抽出することは，知識発見における重要なテーマの一つである．**Wille** らは特に有用な部分関係として，図 **1(2)**に示すような●で埋め尽くされる矩形で表される閉集合に着目し，閉集合について束理論を用いた順序代数的構造を明らかにした**[1]**．２項関係の閉集合は，買い物データからのバスケット分析などに端を発するアイテム集合データベースからの知識発見に応用され，有用な知識としての頻出閉集合という概念を導出した**[2]**．離散値属性を２部グラフの節点とみなすとき，閉集合は完全部分グラフになることから，完全部分グラフの効率的生成研究が主に我が国を中心に展開された**[3]**.

閉集合は大きな功績を残したものの，**Web** ページ間のリンク構造を用いたコミュニティ抽出などグラフの特性が直接対象となるような分野では数学的に定義が強すぎる．さらに，実データはノイズも考慮しなければならない．そこで近年のコミュニティ抽出などにおいては有用な知識として密な部分関係が注目されている．密な部分関係とは図１**(3)**に示すような，十分な濃度を持つ部分関係であり，**[4]**は行列を用いて密な部分関係を抽出している．図 **1(3)**をみるとわかるように，×印１か所を矩形内に入れるだけで閉集合よりずっと単純で理解しやすい矩形が得られる．

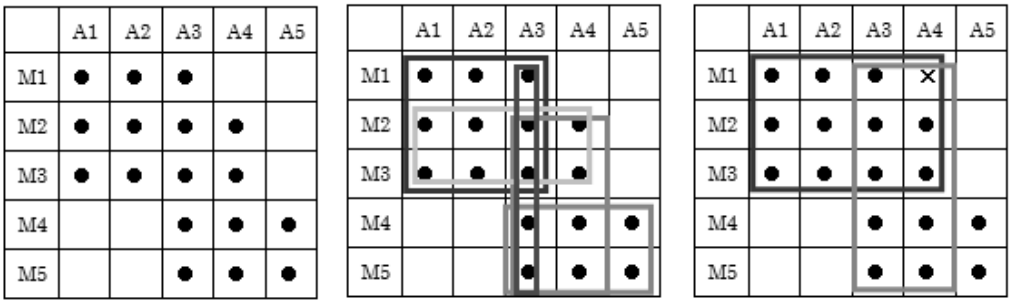

図１　２項関係データと閉集合，弱閉集合

２．研究の目的

もともと閉集合を構成した **Wille** は，数学における **Galois** 対応を範として，閉集合に対する順序代数的かつ俯瞰的な理論を構成し，それに対して **Uno** らが開発したアルゴリズムを適用するという方向で研究が進んできた．一方，　密な部分関係はデータマイニングからの実応用から生まれたため，そのような順序代数的かつ俯瞰的な理論は我々の知る限り提案されていない．一方，現在数多くなされている密な部分関係は行列や統計的最適化理論に基づく局所的な解析に基づいている．　そこで本研究の目的を，密な部分関係を弱閉集合とよぶとき，閉集合全体の順序代数的性質と対比可能な形で弱閉集合全体のなす性質を提示した上で，弱閉集合を求める効率的なアルゴリズムを構成することとする．

２．研究の方法

弱閉集合の代数的性質を求めるという目標に対して，以下を研究する．

1. 弱閉集合の集合論的定式化：密度の概念を適切に定義することで，閉集合を自然に拡張した密な部分関係を集合論的に定式化する．密度の概念は単に数値としての密度を導入するのではなく，集合論的に定式化する．弱閉集合として密な部分関係を定義するためには，何らかの形で密度の概念が必要である．しかし，対象データは離散値属性であり，密度は実数を基礎とした概念であるため，直接的な密度の導入が適切とは限らない，と考えられる．そこで，閉集合を自然に拡張した密な部分関係を集合論的に定式化することを目指す．
2. 弱閉集合間の順序代数的関係：閉集合間には集合の包含関係を用いた順序関係が成立し，閉集合全体は完備束をなす．弱閉集合に対して，それと対比できる形の順序代数的構造を与える．閉集合間には集合の包含関係を用いた順序関係が成立し，閉集合全体は完備束をなす，という特徴があり，頻出閉集合はこの特徴を巧妙に利用している．そこで，弱閉集合に対しても閉集合と同様の束理論を構成することを目指す．
3. 弱閉集合の不動点意味論：閉集合が数学的に強すぎることは，閉集合がある関数を属性値の集合に対して２度適用しただけ得られる不動点として特徴付けられる点にある．２項関係の特徴ある部分関係を求める手法はデータマイニングでも多く利用されており，EM アルゴリ

ズムや Page ランクアルゴリズムはその典型である．それらはある関数を任意有限回繰り返して得られる不動点を求めている．弱閉集合もそのような不動点で特徴づけることを目標に理論を構成する．閉集合が数学的に強すぎること，実用の際に細かすぎることは，閉集合がある関数を２度繰り返し適用しただけ得られる不動点として特徴付けられる，ということを帰結してしまう．一般に，不動点は関数を任意有限回繰り返し適用して漸近的に得られるものであり，データマイニングや機械学習のアルゴリズムはそのような繰り返しを持つものが多い．そこで弱閉集合を求めるアルゴリズムの不動点帰納が弱閉集合を求めることに一致するかどうかを検討する．

4. 弱閉集合を列挙するアルゴリズムの開発と実用性検討：弱閉集合を列挙する独自アルゴリズムの開発を目指し，実用性を検討する．構成した弱閉集合の理論をベースに，弱閉集合を求めるアルゴリズムを設計する．ちょうど閉集合に対して，完備２部グラフを求めるアルゴリズムが適用されたように，弱閉集合に対してもグラフ理論的なアルゴリズムの適用を試みる．

参考文献

[1] Wille, R.: Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. In: B. Ganter et al.: Formal Concept Analysis. Foundations and Applications, Springer, 2005.
[2] Pasquier, Y. Bastide, R. Taoil, and L. Kakhal: Efficient Mining of Association Rules using Closed Itemset Lattices, Inf. Sys. 24(1), 25-46, 1999.
[3] Uno, T. : A Fast Algorithm for Enumerating Bipartite Perfect Matchings, LNCS 2223, 367-379, 2001.
[4] Shin, K., Jung, J., Lee, S. and Kang, U. BEAR: Block Elimination Approach for Random Walk with Restart on Large Graphs, Proceedings of the 2015 ACM SIGMOD, 1571-1585.

４．研究成果

弱閉集合の集合論的定式化

閉集合をグラフ理論のことばを用いれば，２部グラフの極大完全部分グラフという解釈が可能である．すなわち，図１(a)の２項関係は，２つの頂点集合 M={M1，M2，…，M5}と A={A1，A2，…，A5}を考えたときに，●が頂点間に辺が引かれていることを表していると解釈すれば，２部グラフを表現していると解釈できる．さらに(b)に示す閉集合は，この２部グラフの極大完全グラフ(極大クリーク)，つまり，(a)が表すグラフの部分グラフで，M 側の頂点と A 側の頂点の組すべてについて辺があり，そこに入っていない頂点をどのように付加しても完全グラフにならないようなグラフになっている．この議論で利用している数学的道具は集合だけである．そこで，弱閉集合をグラフ理論の概念を参考にして構成することとした．

２部グラフではないグラフにおいて完全グラフ(クリーク)の定義を弱めた概念に k-Plex がある．k-Plex とは，各頂点ごとに高々k 本の辺を追加すると完全グラフになるようなグラフである．そこで k-Plex の定義を２部グラフ用に修正した．すなわち，２項関係を表現する２部グラフが２つの頂点集合 M と A の間の辺だけに着目するため，２部グラフに対しては(k,l)-Plex を，M 側の頂点については高々k 本，A 側の頂点については高々l 本の辺を追加すれば完全グラフになるもの，と定義する．そこで弱集合を，２項関係を２部グラフと解釈したときに，極大な(k,l)-閉部分集合となる部分関係として定式化した．以下では弱集合を(k,l)-閉集合とよぶ．例えば，図２の赤枠で囲んだ部分は(1,2)-閉集合である．

	1	2	3	4	5
A	1	1	1	0	0
B	1	1	1	0	0
C	1	1	1	1	0
D	1	1	0	1	1
E	0	0	0	1	1

図２ (1,2)-閉集合

弱閉集合を列挙するアルゴリズムの開発

閉集合の高速列挙アルゴリズムである LCM を修正することにより，(k,l)-閉集合の列挙するアルゴリズムを設計した．このアルゴリズムは，理論上は高速な列挙が不可能な場合もあるが，現実の小規模データに対しては十分高速に列挙することを確認している．

アルゴリズムの設計方針は，LCM に倣っている．すなわち LCM では，閉集合が，A の冪集合から M の冪集合への関数↓と M の冪集合から A の冪集合への関数↑を用いると，合成関数↓↑の不動点として表現できることを利用して，A の部分集合を{A1}，{A1,A2},{A1,A2,A3},…と線形的に増加させながら↓を適用し，得られた M の集合に↑を順次適用して不動点を求めるという逆探索を行う．このアルゴリズムの正しさを保証するのは合成関数↓↑の単調性である．

(k,l)-閉集合の場合は，関数↓と↑の定義を修正して↓k と↑l とした上で，同様に A の部分集合を線形的に増加させながら逆探索を行う．ただし，合成関数↓k↑l は必ずしも単調とならないので，↓k↑l を任意有限回適用した閉作用素を考えて，適用することになる．以下にも述べるが，(k,l)-閉集合は↓k↑l の不動点にならない場合があるので，提案アルゴリズムは，最悪の場合前回探索を行ってしまう．

弱閉集合間の順序代数的関係と弱閉集合の不動点意味論

研究計画段階では弱閉集合の順序代数的関係と不動点意味論を別項目として立てたが，成果としては一体であるので，ここでまとめて述べておく．

結論を述べれば，弱閉集合間に集合としての包含関係で順序を入れたとしても，完備束にはならない．また，弱閉集合は↓k↑l を任意有限回繰り返し適用したとしても漸近的な不動点にはならない．実際，下のような２項関係において，枠は(1,1)閉集合を示しているが，これに↓k↑l を繰り返し適用してゆくと「徐々に」右下に移動してゆく．このような例が上述のアルゴリズムの理論的な評価を妨げていることが判明した．

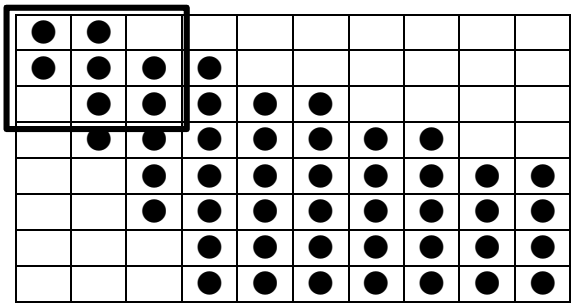

図３ (1,1)-閉集合と↓1↑1 の繰り返し適用が不動点を構成しない例

弱閉集合を列挙するアルゴリズムの実用性検討

実データとして，日本を訪問した外国人旅行者の GPS を用いた移動記録を集めたデータを借用することができたので，弱閉集合の考え方を経路データ向けに修正した上で，頻出する(k,l)-閉集合を列挙し，外国人旅行者がよく辿る特徴的な経路を列挙することに成功した．

データの前処理として，移動記録は緯度と経度からなるので，離散値属性に変換するために，住所表記を利用する．住所表記は都道府県から始まり，丁目まで階層的に構成されているので，適切な粒度の属性を選択することが可能である．２項関係データは，各 M の要素対して A の要素の有限集合を対応させていると解釈することができる．一方，経路データは，M を旅行者の集合，A を住所の集合とした場合，M の要素である旅行者に対して，住所の列が得られる．住所の列を集合とみなしてしまうと，旅行者の経路はわからなくなってしまうので，住所の列を列のまま扱うことにする．そのため，(k,l)-閉集合の定義を A 側については集合ではなく列と修正しておく．さらに，列挙アルゴリズムも LCM ではなく，列の列挙である PrefixSpan アルゴリズムを(k,l)-閉集合用に修正して利用した．

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Siqi Peng and Akihiro Yamamoto	４．巻 12323
２．論文標題 Mining Disjoint Sequential Pattern Pairs from Tourist Trajectory Data	５．発行年 2020年
３．雑誌名 Lecture Notes in Artificial Intelligence	６．最初と最後の頁 645-648
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-61527-7_42	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Takuya Kida, Tetsuji Kuboyama, Takeaki Uno, Akihiro Yamamoto	４．巻 109
２．論文標題 Special issue on Discovery Science	５．発行年 2020年
３．雑誌名 Machine Learning	６．最初と最後の頁 1145-1146
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10994-020-05883-7	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Madori Ikeda and Akihiro Yamamoto	４．巻 24
２．論文標題 Extending Various Thesauri by Finding Synonym Sets from a Formal Concept Lattice	５．発行年 2017年
３．雑誌名 Journal of NLP	６．最初と最後の頁 323-349
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.24.323	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yasuaki Kobayahsi, Hiromu Ohtsuka, Hisao Tamaki:	４．巻 89
２．論文標題 An improved fixed-parameter algorithm for one-page crossing minimization	５．発行年 2017年
３．雑誌名 LIPICS	６．最初と最後の頁 25:1-25:12
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　1件）

１．発表者名
Siqi Peng and Akihiro Yamamoto

２．発表標題
Applying ZBDD for Triadic Concept Analysis

３．学会等名
人工知能学会 人工知能基本問題研究会（第116回）

４．発表年
2021年

１．発表者名
日野 遼人，山本 章博

２．発表標題
消失イデアルのGroebner基底を計算するFarr-Gao のアルゴリズムの機械学習としての性質

３．学会等名
人工知能学会 人工知能基本問題研究会（第115回）

４．発表年
2021年

１．発表者名
Siqi Peng and Akihiro Yamamoto

２．発表標題
Introduction of Triadic Concept Analysis and Its Possible Improvement

３．学会等名
人工知能学会 人工知能基本問題研究会（第114回）

４．発表年
2020年

１．発表者名
久保田 稜，小島 健介，小林 靖明，山本 章博

２．発表標題
可換マッチング問題の固定パラメーター容易性に関する研究

３．学会等名
人工知能学会 人工知能基本問題研究会（第112回）

４．発表年
2020年

１．発表者名
Siqi Peng and Akihiro Yamamoto

２．発表標題
Mining Dijoint Sequential Pattern Pairs from Tourist Trajectory Data

３．学会等名
23rd International Conference on Discovery Science（国際学会）

４．発表年
2020年

１．発表者名
Siqi Peng and Akihiro Yamamoto

２．発表標題
Improvement of sequential pattern mining based on (k,l)-frequency and generative probability

３．学会等名
人工知能学会 人工知能基本問題研究会（第111回）

４．発表年
2020年

１．発表者名
里見 琢聞, 小林 靖明, 山本 章博

２．発表標題
文字列データの線形最小汎化問題に対するアルゴリズム

３．学会等名
人工知能学会 人工知能基本問題研究会（第109回）

４．発表年
2020年

１．発表者名
小島 健介(研究協力者), 呉 可天(研究協力者)

２．発表標題
二部グラフにおける(k,l)-Plexのための形式概念解析の拡張

３．学会等名
人工知能学会 人工知能基本問題研究会（第108回）

４．発表年
2019年

１．発表者名 江良 佳朗，山本 章博，熊田 孝恒
２．発表標題 ドライブデータからの運転手間の相違を表す属性のDTWによる発見
３．学会等名 人工知能学会 人工知能基本問題研究会（第106回）
４．発表年 2018年

１．発表者名 久保田稜，小林靖明，山本章博
２．発表標題 整数計画法による木構造データ間のアラインメント距離の計算
３．学会等名 人工知能学会 人工知能基本問題研究会（第106回）
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	小林　靖明 (Kobayashi Yauaki) (60735083)	京都大学・情報学研究科・助教 (14301)	
研究分担者	久保山　哲二 (Kuboyama Testuji) (80302660)	学習院大学・付置研究所・教授 (32606)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計1件

国際研究集会	開催年
The 20th International Conference on Discovery Science (DS 2017)	2017年～2017年

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関