様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 5 月 20 日現在

研究種目：基盤研究(B)
研究期間：2007 ～ 2008
課題番号：19300040
研究課題名（和文）　抽象化に基づく情間構造マイニング

研究課題名（英文）　Crossover Concept Mining Based on Conceptual Abstraction Hierarchies

研究代表者
原口 誠 （ MAKOTO HARAGUCHI ）
北海道大学・大学院情報科学研究科・教授
研究者番号：40128450

研究成果の概要：
潜在的なクラスタ間接続を，たとえ，外延的な重なりが小さな場合でも，内包的な繋がりが高い場合は有意義なクラスタ間接続関係を与える「情間」と捉える．こうした「情間」は，クラスタの外延と内包に対する評価関数を設定することにより，制約条件下の最適化問題として捉えることができ，具体的には，形式概念束探索における分枝限定アルゴリズムとして設計・開発・実証を行った．形式概念束は，概念の抽象階層と見なすことができ，主要なクラスタを繋ぐ適切な概念を抽象階層から見出すことを意味している．実証実験において，カテゴリとしては離れた異種文書群から，それらを跨ぐ，機能的な概念を情間（クロスオーバ概念）として抽出することに成功している．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	2,400,000	720,000	3,120,000
2008 年度	1,900,000	570,000	2,470,000
年度			
年度			
年度			
総　計	4,300,000	1,290,000	5,590,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：　クロスオーバ、内包的接続

１．研究開始当初の背景

急速に拡大した情報を有効に活用するにあたり関連する事物・情報をクラスタ化し，それらクラスタ間の関係分析により，情報世界をより簡略化・抽象化された構造物として要約し認識するニーズは高いと思われる．クラスタリングは外延的に事物を要約する手法だが，良いクラスタかどうかの基準の多くは事物の集合である外延に依存するために，内容の関連性を的確に抽出・発見する機能において十分ではないと思われる。すなわち、たとえ外延的な繋がりが顕在的でない場合でも、内包的に顕著な繋がりを検出する手法が、主要なクラスタを接続する顕在情報の発見のために必用になるとの考えに基づいて，本研究をスタートさせた．

２．研究の目的

密接に関連したものを纏まりあるもの，すなわち，クラスタとして抽出するクラスタリングは，雑多で個別的なデータからなるデータ群が持つ複雑さを軽減するために重要で

あり，多くの研究がなされてきた．一方，リンク情報などの関連性を用い，相互に密結合されたページやノードの塊を見つける，コミュニティやクリークの研究も，クラスタリング同様に，一般には膨大な個別的ノード群から意味のある部分を見出すための技法であると思われる．

類似性や関連性を与えるものは，クラスタリングにおいては特徴に関わる類似性尺度，また，コミュニティ検出においてはリンク情報，などの違いはあるが，構成要素数が最多もしくは比較的に多数なものを好む傾向にある．実際，小さなクラスタの数は膨大であり，そうした細々としたものを調べることにどれ程の意味があるのかという理由に基づくと思われる．また，データマイニングにおいても，ほぼ同様の理由により，生起頻度が比較的大な，頻出パターン検出が多かったと思われる．

『頻出するものは重要だ』は，重要さの基準・経験則としては十分な一般性を持つものであり，本研究はこれに異議を唱えるものでは決してない．しかしながら，非頻出なパターンで個体数において小規模なものの中にも重要なものが潜んでいる可能性もこれまた無視できないものと思われる．本研究では，

『非頻出であっても，主要なクラスタを繋ぐものは重要だ』

との考えに基づき，非頻出なものの中から重要なものを絞りこんで提示できるシステムをめざし，制約付のトップ N 分枝限定アルゴリズムによって十分に高速に抽出可能なことを示す．ここで，「主要なものを繋ぐ」と言っても，個体集合としてのクラスタを摂動させ，別の近傍クラスタをたどる行為などもあるかと思われる．ここでの「繋ぐ」とは，例え外延的に離れたもの（個体集合としての交わりが空であっても良い）でも，内包的な共通性が認められる場合は，重要な繋がりの可能性があると考える．つまり，

『異なる概念間にまたがる共通概念で非頻出なもの』

をクロスオーバ概念として定める．本研究においては，形式概念を検出目標物とした議論を行うが，これは，形式概念が外延的な個体オブジェクトと内包的な属性・特徴の組からなり，外延的にはレアだが，内包的な交わりとしてのクロスオーバを直接的に記述できるとの理由による．

３．研究の方法

O を**オブジェクト(個体)**の集合，F を**属性**の集合とする．任意のオブジェクト集合 $X \subseteq O$ について，X 中のすべてのオブジェクトに共有される属性集合を ΦX で表す．一方，任意の属性集合 $A \subseteq F$ について，A 中のすべての属性を有するオブジェクト集合を ΨA で表す．オブジェクト集合 $Z, W \subseteq O$ について，Z の共有する属性集合が W においても共有される，すなわち，$\Phi Z \subseteq \Phi W$ である時，**Zは W を含意する**と言い，$Z \rightarrow W$ と表記する．

$X = \{ x \mid X \rightarrow x \}$ なるオブジェクト集合 X を，**外延**と呼ぶ．つまり，外延とは，オブジェクト含意のもとで閉じた集合のことである．**内包**についても同様に，属性含意のもとで閉じた集合と定義できる．特に，任意の集合 $X \subseteq O$ と $A \subseteq F$ について，ΦX および ΨA はそれぞれ，内包と外延となることに注意する．**形式概念**とは，外延 X とその対応する内包 ΦX の組$(X, \Phi X)$，もしくは内包 C と対応する外延 ΨC の組$(\Psi C, C)$で定義される．以下の議論では，概念とその外延(あるいは内包)を同一視する．

形式概念の評価を行なうために，外延と内包に関する評価関数 $eval_O$ と $eval_F$ を考える．特に，ここでは，これら関数は集合の包含関係のもとで単調性を有すると仮定する．すなわち，$X_1 \subseteq X_2$ $(A_1 \subseteq A_2)$ ならば，$eval_O(X_1) \leqq eval_O(X_2)$ $(eval_F(A_1) \leqq eval_F(A_2))$とする．集合の要素数(サイズ)は，こうした評価関数のひとつであり，以下の議論ではこれを仮定する．

本研究では，非頻出な概念抽出を行うが，あまりにもレアすぎる，すなわち，個体概念に近いものが興味の外にあることも自明である．下記の形式化では，内包に対する制約により非頻出なものに限定し，その中で外延評価を最大化することにより，あまりにもレアなものを結果的に排除する．

目的関数(最適化)： 以下の制約を満たす外延 X の中で，$eval_O(X)$による評価値が上位 N のもの．

内包制約(必須)： 閾値 $\delta > 0$ に関して，$eval_F(\Phi X) \geqq \delta$ である．

空間制約(オプション)： X は以下を満たす．

(POS)：所与の正例オブジェクト集合S^+について，$S^+ \subseteq X$ である．

(NEG)：所与の負例オブジェクト集合 S について，$S^- \cap X = \phi$ である．

(SUB)：所与の関連属性集合 K について，$X \subseteq \Psi K$ である．

(POS) は，$I=\{z \mid S^{+} \rightarrow z\} \subseteq X$ であることを意味し，I は(概念の)外延束をボトムアップに探索する場合の起点を定めている．一方，(SUB)は，外延束の上限となる外延 ΨK を規定するものである．これは，$K \subseteq \Phi X$ なる制約と等価であり，抽出すべき概念は少なくとも K 中の属性をすべて含まなければならないことを要請し，これによりユーザの興味を反映させることができる．この様に，探索の対象は，I を下限，ΨK を上限とする部分束に限定される．

さて，所与の主要な概念 $(X_j, \Phi X_j)$ に対し，観点毎のクロスオーバ概念 C を下記の２条件を満たす $(\Psi C, C)$ として定める．ただし，観点は内包 K として与えるとする．

観点 K からみた類似性条件： $K \subseteq C$

外延条件： $\Psi C \cap X_1 \neq \phi$，　$\Psi C \cap X_2 \neq \phi$

類似性条件は，内包 K の拡大としてクロスオーバが観測できることを要請し，(SUB) $\Psi C \subseteq \Psi K$ と等価である．つまり，空間制約の特殊な場合として扱える．さらに，外延条件は，$x_j \in \Psi C \cap X_j$ なる例示を (POS) 制約として与えて，自動で満たされることに注意したい．これは，主要な概念間のクロスオーバと言っても様々であり，例示をトリガーとして持つクロスオーバを限定的に求めることを意味している．

さて，こうした概念探索をより高速に実行するために，内包制約による探索の枝刈りが有効となることを示す．内包制約は，外延の包含関係のもとでは逆単調性を有する．すなわち，$X_1 \subseteq X_2$ なる外延 X_1 と X_2 について，X_1 が内包制約を満たさない場合は，X_2 も同様に満たさない．よって，ボトムアップに深さ優先探索を行なう際，X_1 が制約を満たさないことがわかった時点で，そこから先に探索枝を張る必要はなく，直ちにバックトラックすることができる．

この様に，制約に基づいて探索空間を限定することが可能であるが，さらに効率的な探索を行なうには，概念の列挙法にも工夫が必要である．そのために，ここでは，拡張候補の動的順序付け，および，重複概念の生成を抑制するための規則を新たに導入する．

所与の制約を満たす外延 X と，あるオブジェクト $x \notin X$ を考える．いま，$\mathrm{cl}(X,x)=\{z \mid X,x \rightarrow z\}$ が制約を満たす外延である時，x は，X **の拡張候補**と呼ばれる．つまり，拡張候補 x は，X に追加することにより，制約を満足し，かつ，X を包含する外延を生成可能なオブジェクトである．探索過程において，x は，X の次の(より大きな)外延を求める際の新たな探索枝を形成する．

一般に X の拡張候補は複数存在するが，それらを用いて X から探索枝を張る順序は，探索効率に大きな影響を与える．ここでは，属性集合 $\Phi(X \cup \{x\})$ の大きさの昇順で処理していくものとする．こうした順序は，各 X 毎に決まるものであるから，これを**拡張候補の動的順序付け**と呼び，X における順序を $<_X$ と表記する．

X を x を用いて拡張することで，新たな外延 $\mathrm{cl}(X, x)$ が生成されるが，そこには $X \cup \{x\}$ が含意する y，すなわち，$X,x \rightarrow y$ なる y が x と共に含まれる．ここで，属性集合 $\Phi(X \cup \{x\})$ が小さいことは，$X \cup \{x\}$ がより多くの y を含意する期待が持てることを意味し，結果として，より大きな外延 $\mathrm{cl}(X, x)$ の生成が期待できる．つまり，拡張候補の動的順序付けは，より大きな外延を，より早期に見つけることを狙ったヒューリスティックであり，これにより，後に述べる分枝限定による枝刈り効果を高めることができる．

探索の基本戦略は，上述した拡張候補の動的順序付けに基づく深さ優先探索である．初期(ルート)ノードは，(POS) 制約から定まる外延 $\{z \mid S^{+} \rightarrow z\}$ とし，その拡張候補の動的順序に従って候補を選択しながら，探索木を深さ優先で拡張していく．探索過程で選択された候補の系列 $c_1,...,c_k$ は，ルートから外延 $\{z \mid S^{+}, c_1,...,c_k \rightarrow z\}$ に至るパスを表す．つまり，$S^{+} \cup \{c_1,...,c_k\}$ は，この外延のひとつの**生成元(generator)**となる．しかし，一般に，各外延の生成元は複数存在することから，効率良い探索を行なうためには，外延の重複生成を抑制する機構が必要となる．

そのために，ここでは拡張候補を，**右候補**と**左候補**のふたつに分類する．前者は実際に探索木を拡張する際に用いる候補であり，後者は重複生成をチェックするための候補である．

いま，外延の系列 $X_0=\{z \mid S^{+} \rightarrow z\}, X_1, \ldots, X_k$ を考える．ここで，$X_i = \{z \mid S^{+}, c_1,...,c_i \rightarrow z\} (1 \leqq i \leqq k)$であり，$c_i$ は，X_i を生成する際に X_{i-1} で選択された候補とする．この時，X_{k+1} を生成する際に X_k で選択されたある候補 c_{k+1} に関して，X_k での候補 r が $r \in \{c_1,...,c_k\}$ あるいは $r <_{X_{k+1}} c_{k+1}$ ならば，r は**左候補**と呼ばれる．また，これら以外の候補を**右候補**とする．

以下に示す枝刈り規則は，こうした候補の

区別によって，外延の重複生成を安全に排除する．

逆含意に基づく枝刈り： 拡張を試みる外延を X とし，その右候補 r と左候補 l を考える．この時，$X, r \rightarrow l$ ならば，X を r で拡張する必要はない．

これに加え，暫定的に見つかった解の評価値を利用することで，上位 N となる外延の生成が見込めない探索枝を安全に枝刈ることも可能である．

分枝限定に基づく枝刈り： 拡張を試みる外延をX，その右候補を r とする．この時($X_r = \{ w \mid X, r \rightarrow w \}$) $\cup$ {X_r における右候補} の評価値が，上位 N の暫定解における最小評価値よりも小さい場合，X を r で拡張する必要はない．

ただし，上位 N の暫定解がまだ見つかっていない場合は，この規則は無効となる．

４．研究成果

上述したアルゴリズムを Java で実装した．本節ではその実験結果について述べる．

ここでは，Web 文書クラスタリングのベンチマークとして公開されている *BankSearch* と呼ばれるデータセットを用いた．これは，11,000 の Web 文書(HTML 文書)から成り，次の 11 カテゴリからそれぞれ 1,000 文書を集めたものである：
``*Commercial Banks*'', ``*Building Societies*'', ``*Insurance Agencies*'', ``*Java*'', ``*C/C++*'', ``*Visual Basic*'', ``*Astronomy*'', ``*Biology*'', ``*Soccer*'', ``*Motor Sport*'' および ``*Sport*''．
前処理として，もとの HTML 文書からタグを取り除いてテキスト化し，そこから，**WordNet** にある形容詞・副詞，および，(一般的な)ストップワードを除去する．ステミング処理後，高頻度語および低頻度語を除いた 1,223 語を特徴語(属性)とした．ここで，各文書に付随するカテゴリの情報は，属性として一切用いていないことを強調しておく．なお，実験は，Dual-Core AMD Opteron processor 2222 SE を搭載した，主記憶 16GB の PC で行なった．

抽出概念例：
BankSearch 中の Web 文書
http://www.vbsquare.com/files/association
を与え，δ = 50 のもとで Top-3 概念を抽出した．以下は，得られた概念の一例である．

< { http://www.vbsquare.com/files/association/,
……
http://www.vbsquare.com/registry/tip471.html,
http://www.vb-helper.com/links.htm,
http://www.vbsquare.com/databases/dbclass/,
http://www.vbsquare.com/databases/learndb/,
http://www.vbsquare.com/mouse/context/ },
{ API, component, resource, …, tips, VB, graphic } >

外延は 35 の Web 文書から成るが，それらはすべて"*Visual Basic*"のリソースやチュートリアルに関するものであり，*BankSearch* においても，同一のカテゴリに属する文書であった．ただし，ここではこのカテゴリ情報を陽には用いておらず，あくまでも，文書中に現れる語のみに基づいて抽出した概念であることを再度強調しておく．この様に本手法では，事前に付与されたカテゴリ情報を陽に使わなくとも，意味的に妥当な概念の抽出が可能である．

次に，ふたつの Web 文書
http://www.citibank.com/uk/portal/consumer/helpdesk/tc/tc1.htm
http://vbtechniques.com/useragreement.asp
および，二つの関連語 claim, Internet を与え，δ = 50 のもとで，Top-1 概念の抽出を試みた．その結果，一例として次の概念を得た．

<{ http://www.citibank.com/uk/portal/consumer/helpdesk/tc/tc1.htm,
……
http://vbtechniques.com/useragreement.asp,
http://www.hrbs.co.uk/cashisatandcapply.htm,
http://www.hrbs.co.uk/panthertandconline.htm,
http://www.hrbs.co.uk/rewardsixtandcapply.htm,
http://www.lloyds.com/un/en/termsandconditions/category/article/ },
{ claim, Internet, accept, law, condition, …, reason, right, term, transfer } >

外延は 22 の文書から成り，これらは契約や条項に関するものである．また，これらは 4 つの異なるカテゴリ``*Commercial Banks*'', ``*Visual Basic*'', ``*Building Society*'' および ``*Insurance Agency*''に属するものであり，クロスオーバコンセプトの具体例となっている．

この様に，本手法は，クロスオーバコンセプトを含む様々なタイプの概念を抽出可能な極めて柔軟な枠組であると考える．

計算効率：

先に述べた通り，空間制約として与える正例・負例・関連語(属性)の集合は，ユーザの意図を反映しつつ探索空間を限定する．さらに，拡張候補の動的順序付けも計算効率の改善に寄与している．ここでは，先のクロスオーバコンセプト抽出において与えた正例と関連語について，その効果を具体的に示す．

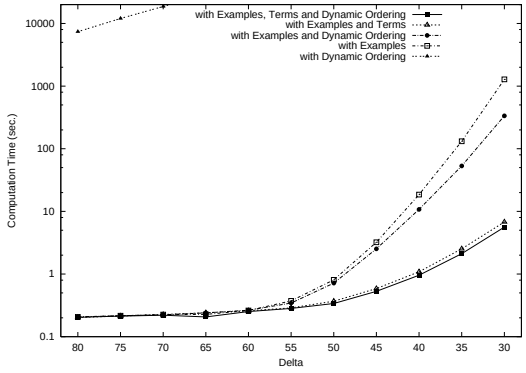

図 1. 正例・関連属性・動的順序付けによる計算時間の改善

図 1 より，正例・関連語(属性)，および，動的順序付けが，計算効率の改善に極めて効果的であることが見て取れる．正例と動的順序付けのみでも大きな改善が見られるが，関連語によって，さらに大幅な改善が得られる．こうした様子は，δ の値が低い領域で特に顕著であり，大規模データに対しても有望なものとなろう．

課題としては，非頻出性を第１制約として形式化したことから，一般には長大な内包が提示されるという難点がある．これを避けるために，ここでで与えたものと双対な手法（トップダウン法）を与えることができる．

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８件）

1. Yoshiaki OKUBO and Makoto HARAGUCHI, Finding Top-N Pseudo Formal Concepts with Core Intents, Proceedings of the 6th International Conference on Machine Learning and Data Mining – MLDM'09, 2009, 査読有(採択済).

2. 佐藤 憲二・大久保 好章・原口 誠・國藤 進，時系列データをもちいた形式概念分析法の提案，日本創造学会論文誌，12，175 – 187，2009，査読有．

3. Aixiang LI, Makoto HARAGUCHI and Yoshiaki OKUBO, Implicit Groups of Web Pages as Constrained Top N Concepts, Proceeding of the 2008 IEEE/WIC/ACM Internatinal Conference on Web Intelligence and Intelligent Agent Technology Workshops, 190 – 194, 2008, 査読有.

4. Aixiang LI, Makoto HARAGUCHI and Yoshiaki OKUBO, A Top N Closed Pattern Miner using Counterexamples, Proceedings of the International Workshop on Data Mining and Statistical Science - DMSS'08, 33 – 35, 2008，査読有.

5. Aixiang LI, Makoto HARAGUCHI and Yoshiaki OKUBO, Finding Top-N Formal Concepts Guided by Dynamic Ordering of Objects, Proceeding of the 6th International Conference on Concept Latticeand Its Applications - CLA'08, 17(poster), 2008，査読有.

6. 難波 徹郎・原口 誠・大久保 好章，遺伝子発現データからの接尾辞木に基づく疑似バイクラスタ抽出，統計数理，56，169 – 184，2008，査読有．

7. M. Haraguchi and Y. Okubo, An Extended Branch-and-Bound Search Algorithm for Finding Top-N Formal Concepts of Documents, Lecture Notes in Computer Science, 4384, 276 – 288, 2007, 査読有.

8. K. Sato, Y. Okubo, M. Haraguchi and S. Kunifuji, Data Mining of Time-Series Medical Data by Formal Concept Analysis, Lecture Notes in Computer Science, 4693, 1214 – 1221, 2007, 査読有.

〔学会発表〕（計１０件）

1. 原口 誠・大久保 好章，分枝限定法を用いた最適クロスオーバコンセプト検出法について，情報処理学会数理モデル化と問題解決研究会(第 72 回)，2008 年 12 月 17 日，大阪大学豊中キャンパス.

2. 大久保 好章・原口 誠，内包の核を考慮した疑似形式概念の Top-N 抽出，情報処理学会数理モデル化と問題解決研究会(第71回)，2008年9月18日，電気通信

大学.

3. 李 愛香・原口 誠, 内包による外延評価に基づく Top-N 形式概念文書クラスタの抽出, 人工知能学会全国大会(第22回, 2008年6月13日, 勤労者福祉総合センター.

4. 李 愛香・原口 誠, 2 次形式最小化に基づく動的順序付けを用いた形式概念探索, 情報処理学会第 70 回全国大会, 2008 年 3 月 13 – 15 日, 筑波大学・筑波キャンパス.

5. 伊藤 公人・谷口 剛・村上 梯治・有村 博紀・原口 誠, 塩基およびアミノ酸配列における共変異集合を列挙する高速アルゴリズム, 情報処理学会第 12 回バイオ情報学研究会, 2008 年 3 月 3 – 4 日, 九州大学・伊都キャンパス.

6. 李 愛香・原口 誠, 潜在的に重要なページを検出するためのフォーマルコンセプトアプローチ, 人工知能学会第 80 回知識ベースシステム研究会, 2008 年 1 月 15 – 16 日, NTT武蔵野研究開発センター.

7. 谷口 剛・原口 誠・伊藤 公人, グループ列分割に基づくインフルエンザウィルスのアミノ酸置換における時代的変化の解析, 情報処理学会第 11 回バイオ情報学研究会, 2007 年 12 月 20 – 21 日, 産業技術総合研究所臨海副都心センター.

8. 難波 徹郎・原口 誠・大久保 好章, 遺伝子発現データからの接尾辞木に基づく疑似バイクラスタ抽出, The International Workshop on Data-Mining and Statistical Science (DMSS2007), 2007 年 10 月 5 – 6 日, 統計数理研究所.

9. 難波 徹郎・原口 誠, 接尾辞木に基づいた疑似バイクラスタ抽出手法, 第 21 回人工知能学会全国大会, 2007 年 6 月 18 – 22 日, ワールドコンベンションセンターサミット.

10. 谷口 剛・伊藤 公人・五十嵐 学・村上 悌治・原口 誠, アイテム集合間の相関変化検出によるインフルエンザウイルス遺伝子データの解析, 第 21 回人工知能学会全国大会, 2007 年 6 月 18 – 22 日, ワールドコンベンションセンターサミット.

６．研究組織
(1)研究代表者
原口 誠（MAKOTO HARAGUCHI）
北海道大学・大学院情報科学研究科・教授
研究者番号：40128450

(2)研究分担者
富田 悦次（ETSUJI TOMITA）
電気通信大学・名誉教授
研究者番号：40016598

伊藤 公人（KIMIHITO ITO）
北海道大学・人獣共通感染症リサーチセンター・准教授
研究者番号：60396314

吉岡 真治（MASAHARU YOSHIOKA）
北海道大学・大学院情報科学研究科・准教授
研究者番号：40290879

大久保 好章（YOSHIAKI OKUBO）
北海道大学・大学院情報科学研究科・助教
研究者番号：40271639

(3)連携研究者
なし