学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　ジェイ ホンジェ

審査担当者　主　査　特任教授　原口 誠
副　査　教　授　有村 博紀
副　査　教　授　杉本 雅則

学位論文題名

Study on Discovery and Exploration Systems Considering User's Intention
（ユーザの意図を考慮した発見・探索システムの研究）

大規模データに容易にアクセス可能な今日において, 様々な知識発見や情報検索・推薦に関する研究が精力的に遂行されている. 記憶装置の大規模化と処理装置の高速化に伴い, 多くの人にとって等しく有用な情報・知識の発見・検出・学習については既に一定の成功を収めている. 一方, 個々のユーザの興味や発見・検索意図によっては, より個別的なクラスターを扱う必要があり, それらが一般的なものと乖離する程, 小規模もしくは中規模の局所クラスターとして出現する. そうした多数の局所クラスターの中で, どのクラスターがユーザの潜在的な意図と合致するかは明示的に与えられているとは限らず, 可能性のあるものを列挙する, もしくは誘導・推測する手法が必要となる. 本論文では, グラフにおける局所クラスターを一定の密度要求のもとに擬似クリークとして検出できる列挙器 (第 1,2 章), ユーザが潜在的に興味を持つものに誘う機能を持った概念の連想検索 (第 3 章), 異なるデータ領域間での属性の関連付けを, ヒントと呼ばれるユーザ毎の観点の例示に基づいて推論できるシステム (第 4 章) を提案し, ユーザの多様な要求に応える検索・発見機能を持つシステムを構築する際に必要となる技法を提案・検証している.

第 1 章 "Enumerating Maximal Clique Sets with Pseudo-Clique Constraints" においては, オーバーラップするクリークを構成要素とするクリーク族を効率良く全列挙するアルゴリズムを与え実験的に評価している. クリークは最高密度を持つサブグラフとして良く知られ, スパースグラフに対して現実的なアルゴリズムが開発されてきた. 一方, そうしたクリークは互いに重なりあうことが多く, そのことが妥当な解の総数を増加させる一つの要因となっていた. この問題を解決するために, 本研究においては, 極大クリークを部品化した頂点を持つクリークグラフにおけるクリークで, 部品クリークの和集合が元のグラフにおける疑似クリークになるものを求める方式を提案している. 部品となるクリークの総数が問題になるが, ユーザが目標とするサイズ下限制約やクリーク間の集合論的相関の下限制約などを用いることにより, 全列挙が可能なクリークグラフを形成することに成功している. 例として, タンパク質相互作用ネットワークや SNS における友達関係グラフ (約 100 万頂点) を用い, 他の手法との比較の上その有効性を検証している.

第 2 章 "Enumerating Pseudo-Cliques with Density Lower Bound" においては, グラフサイズの線形時間で構築可能な j-核を用いて, 疑似クリーク列挙手法のさらなる深化を実現している. 第 1 章における疑似クリークは, クリーク族から構成される疑似クリークであり, その部分としてクリークを中核に持つものであった. 一般的にはしかし, 中核をなす頂点集合は必ずしもクリークとは限らず, いくつかの辺が欠けた疑似クリークであっても良い. 経験則を搭載した高速検出手法はいくつか既

に知られているが, 列挙の完全性が保障され, かつ 100 万頂点程度のグラフに対して現実的に有効な手法は未だ得られていなかった. 本章では, 疑似クリークに接続数下限制約 (j-核性) を持たせ, 一定の密度が保障された疑似クリークのクラスを考える. そうした密度保障付きの疑似クリークを候補頂点の追加により形成するプロセスにおいて, 非接続数上限制約から決まる一定の距離内の頂点のみが追加候補となることから,j-核性極大疑似クリークはそうした候補を加えた頂点集合の j-核 の部分集合となる. この性質を用い, 密度要求を満たさない多くの疑似クリークを探索の早期において安全に棄却でき, 密度要求を満たす極大疑似クリークを漏れなく列挙できることを示している. さらに, 約 90 万頂点のグラフに対し, 大幅な高速化が達成されることを実験的にも示している.

第 3 章 “Associative Search by Shifting Concepts via Bridges” では, ユーザーが与える初期クエリーが表す概念をシフトさせることにより, ユーザが潜在的に興味を持つものに誘導する連想検索技法について述べている. 一般に良く用いられている (文書の) 連想検索では, 文書と語の双対性を利用し, 探索プロセスの各段階でユーザに語を選択させ, ユーザーがその初期段階において陽には気づいていない文書群 (語彙群) に誘導する. 得られる語彙群は初期クエリーが定める概念に対しその兄弟概念になっている. 本研究では, ユーザのクエリーと整合的で, かつ, 多分野に関し言及している人を良いメディエータと呼び, 個々の言葉ではなくユーザの検索時の興味とマッチしそうな人を選択させる. メディエータは初期クエリーではカバーできない概念を持つので, これを汎化により取得し, 汎化概念をさらに特殊化することにより兄弟概念を結果的に誘導するシステムを設計・実装している. 技法的には汎化と特殊化のための概念束探索を, またメディエータの多分野性をエントロピーで計量している. 実験は文書データに対し行い, いくつかの興味深い結果を実際に得ている.

第 4 章 “Feature Association Discovery by Linear Algebraic Inference” では, 異なるデータ領域間での属性の関連付けを, ユーザの例示毎に行うシステムについて述べている. 具体的には, 異なる行 (個体) と列 (属性) を持つ領域データをデータ行列の形で 2 つ与え, ユーザーの立場から相似・類似関係にあると思われる属性の対の例示を複数個与える. 技術的には非負行列分解をシンプルに用いている. すなわち, トライ分解は行の相関と列の相関を同時に表現していることから, 行の相関により同じ次元の行空間に縮約し, この縮約データをさらに列の相関により同じ次元の列空間に縮約する. このトライ分解は各データ行列毎に独立に行うが, 結果として, 同じ次元の共通空間に元の属性 (列) を射影することが可能となる. こうしたトライ分解によりユーザが与えたヒントと呼ばれる属性対が共通空間で近接したベクトルに写像されるべく, 属性の 2 部グラフラプラシアンで記述できる正則化項を加味した最適化問題に帰着させている. 手法の検証実験は, 異なる言語間の単語の対応関係をヒントから求める問題として実行し, グーグル翻訳により構築したアンサーセットに対し, 一定の正答率を持つ結果を得ている.

第 5 章では本論文をまとめ, 残された課題, 特に, 第 4 章の属性の関連付け問題の精度を上げるための事前圧縮技法として第 1・2 章の疑似クリーク検出を使う可能性について論じている.

以上を要するに, 密な部分グラフ列挙, 概念の連想検索, 属性の関連付け問題を解くための技法を開発・検証し, 大規模データの時代におけるユーザの多様な要求に応える知識発見・検索のために必要となる技法を明らかにしている. 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.