様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２３年　５月２８日現在

機関番号：１４５０１
研究種目：基盤研究（Ｂ）
研究期間：２００８～２０１０
課題番号：２０３００００３８
研究課題名（和文）　統計モデリングとデータマイニングに基づくネットワーク化知識の創出と活用
研究課題名（英文）　Generating networked knowledge based on statistical modeling and data mining techniques and its applications
研究代表者
江口　浩二（EGUCHI KOJI）
神戸大学・大学院システム情報学研究科・准教授
研究者番号：５０３２１５７６

研究成果の概要（和文）：
本課題は、統計モデリングとデータマイニングを用いて、断片的に散在した情報コンテンツからネットワーク化知識を創出し、人間の知的活動に活用する手段の確立をめざす。この目的のもと、主に以下に示す観点から研究に取り組んだ。
(1) トピックモデルによるテキストデータからの関係構造の抽出：
種々の構造を有するテキストデータから潜在構造を統計的に推定し、固有表現間の関係やデータ間の関係を抽出する問題などに適用する研究
(2) ネットワークデータに対する頂点クラスタ推定とその応用：
ネットワークデータの潜在的な構造を統計的に推定し、リンク予測等の実問題に適用する研究
(3) ネットワークデータに対するパターン発見：
辺または頂点に数値属性の集合が対応付けられたネットワークなどのように複雑な構造をもつデータから、特徴的なパターンを列挙するアルゴリズムに関する研究

研究成果の概要（英文）：
We aim to construct networked knowledge from scattered pieces of information using techniques of statistical modeling and data mining, and apply that knowledge to solve problems in human intellectual activities. To achieve these objectives, we mainly carried out the following research tasks:
(1) Research on extracting relational structures from text data using topic models:
We statistically estimated latent topics from some sorts of structured text data, and used the latent topics in some applications, such as extracting relational structures between entities or between data.
(2) Research on inferring node clusters from network data and its applications:
We statistically estimated latent communities from a network, and applied the estimated communities to real-world problems, such as link prediction.
(3) Research on discovering patterns from network data:
We developed algorithms to discover characteristic patterns from complex structured data, such as a single network in which nodes or edges are associated with a set of numerical attributes.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	3,800,000	1,140,000	4,940,000
２００９年度	3,800,000	1,140,000	4,940,000
２０１０年度	3,500,000	1,050,000	4,550,000
総　計	11,100,000	3,330,000	14,430,000

研究分野：情報学
科研費の分科・細目：情報学　メディア情報学・データベース
キーワード：統計モデリング、データマイニング、トピックモデル、ネットワーク分析、グラフマイニング

１．研究開始当初の背景

(1) 情報検索タスクなどを想定した統計的テキスト処理技術の主要な流れの一つとして、確率的トピックモデル（混合メンバーシップモデルとも呼ばれる）が注目を浴びつつある。特に、潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）が知られており、文書を単語の集合と見なし、そのような文書の集合から潜在的な話題構造を推定するのに用いることができる。また、様々な拡張が可能であり、例えば、固有表現が特徴づけられただけの構造をもつテキストから、潜在的な話題構造を推定し、それに基づいて固有表現の関係の度合いを重みとして表現したネットワークを抽出することができる。

(2) 確率的トピックモデル（混合メンバーシップモデル）は、グラフデータに適用することも可能で、コミュニティすなわち互いに関連する頂点からなるクラスタを推定するのに用いることができる。このように推定されたモデルはグラフデータの探索的分析（exploratory analysis）の手段となるだけでなく、リンク予測などの問題に応用できる。

(3) また、先に述べたような、統計モデリングによって属性が付与されたネットワークを分析する手段として、グラフパターン発見手法が有効である。データマイニング分野で相関ルール発見手法に端を発し、グラフパターン発見手法が発展してきたが、従来のグラフパターン発見の研究では、頂点に離散確率分布が対応するような多構造ネットワークからのパターン発見や、重み付きネットワークにおける近似パターンの発見については、それまでほとんど研究されてこなかったため、新たな観点に基づく拡張が必要となる。

２．研究の目的

(1) 本課題は、人物名や地名を表す固有表現とその他の一般記述、あるいはWikipediaにおける辞典項目と解説記述とカテゴリラベルなどのように、ある種の構造をもつテキストデータを対象に、それぞれの構成要素を互いに依存関係をもつ確率変数とみなすことで、妥当な統計モデルの設計と推定を試みる。

(2) また、確率的トピックモデル（混合メンバーシップモデル）をグラフデータにおけるコミュニティすなわち互いに関連する頂点からなるクラスタの推定のために用いて、推定されたモデルをリンク予測などの問題に適用し、その有効性を評価する。

(3) 統計モデリングによって属性が付与されたネットワークを想定して、頂点に離散確率分布が対応するようなネットワークからのパターン発見や、重み付きネットワークにおける近似パターンの発見のため、新たなグラフパターン発見手法を開発する。

(4) 以上のように、本課題は統計モデリングとデータマイニングによって、散在した情報の断片から、人間の知的活動に直接活用可能な「知識」を発見し、活用する手段の確立を目的とする。

３．研究の方法

(1) トピックモデルによるテキストデータからの関係構造の抽出：

① 学術固有表現抽出技術を適用して人名等のWho型、地名等のWhere型の固有表現がタグ付けされたテキストデータを想定し、そのような単語型を観測変数と見なし、潜在変数として表現したトピックとの依存性をモデル化する。ギブスサンプリングによってモデルの未知パラメータを推定する。得られたトピックに基づいて固有表現間の関係の定量化を実現し、また、ネットワークとして表現することで複雑な関係を可視化する。

② ブログポストの潜在的トピックに着目して、ブログポスト間のハイパーリンクで不適切なものを検出し、除外することにより、情報伝播ネットワークを抽出する。ハイパーリンクの両端に位置するブログポストの潜在トピック分布を推定し、その分布間距離に基づいて分布が十分異なると判定された場合にアフィリエイト・リンクなどの不適切なハイパーリンクであると仮定する。さらに、適切なハイパーリンクに基づいて情報伝播を表すネットワークを抽出する。

(2) ネットワークデータに対する頂点クラスタ推定とその応用：
① ネットワーク（グラフ）データにおいて、コミュニティを潜在変数で表現したモデルを仮定し、部分的に観測されたデータからモデルの未知パラメータを統計的に推定する。推定手法にはギブスサンプリングを用いる。推定されたコミュニティに基づいて未観測リンクの予測を実現する。
② カテゴリ木構造の各頂点に文書群が割り当てられたテキストデータに対して、トピックを潜在変数として表現し、カテゴリ木構造を考慮しつつ潜在トピックを推定する手法を開発する。葉ノードから内部ノードに至るパスが生成される確率を混合比とし、葉ノードにおけるトピック分布の混合分布として内部ノードのトピックを表現する。推定にはギブスサンプリングを用いる。

(3) ネットワークデータに対するパターン発見:
① 頂点や辺に数値属性の集合が付与された単一グラフを対象とする。数値属性集合を伴うグラフパターンに対する標準形を定義した上で、ラベル付きグラフパターンの列挙手法と高密度クラスタに基づく定量的アイテム集合列挙手法を有機的に連動させ、効率的な特徴的パターン発見アルゴリズムを構築する。
② グラフ自身及びグラフの各構成要素に対し、その重要性や信頼性、意義などを表す重みが付与された、外部及び内部重み付きグラフを想定し、そのグラフデータベースを対象とした特徴的パターン発見について検討を行う。ユーティリティーに基づく特徴的アイテム集合発見の考えをグラフへと拡張し、外部及び内部重みに基づく統合評価関数を開発することにより特徴的パターンを定義する。その上で、上界値に基づく枝刈りメカニズムを飽和・極大パターン発見手法へと組み込むことで、効率的な発見アルゴリズムを構築する。

４．研究成果
(1) トピックモデルによるテキストデータからの関係構造の抽出：
① 学術文献における専門用語間の関係性を定量化する問題において、潜在トピックに着目し、そのモデル化方法、推定方法、語間類似度の計算方法、および、トピック数による性能の違いを明らかにした。
② ブログポストの潜在的トピックに着目して、ブログポスト間のハイパーリンクで不適切なものを検出し、除外することにより、情報伝搬ネットワークを的確に抽出する手法を開発した。また、実際の日本語ブログデータを用いた実験によって、提案手法の有効性を示した。

(2) ネットワークデータに対する頂点クラスタ推定とその応用：
① 部分的に観測されるネットワークから潜在的な構造を統計的に推定し、それを用いて未観測のリンクを予測する手法を開発した。当該手法に用いることによる、生物学的ネットワークにおけるリンク予測の有効性を評価するとともに、それに加えて文献から得られた知識を統合することによる効果を確認した。
② カテゴリ木構造における各頂点に文書群が割り当てられたテキストデータコレクションに対して、カテゴリ木構造を考慮しつつ潜在トピックを推定する手法を実現した。また、階層的テキスト分類すなわち新たに追加された文書をカテゴリ構造上の頂点に割り付ける問題に適用した。

(3) ネットワークデータに対するパターン発見：
① 頂点または辺に定量的アイテム集合をもつ単一グラフを対象とした頻出パターン発見アルゴリズムを実現した。また、テキスト属性付きネットワークデータに対してテキスト属性に潜在するトピックの分布を推定し、その構造的なパターンを効率的に獲得するシステムを実現し、評価を行った。
② グラフの構成要素とグラフそのものに重みが付与された、内部及び外部重み付きグラフを対象に、種々の観点での特徴的パターンを効率的に獲得するアルゴリズムの開発に成功した。

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 5 件）（総計 34 件）
① 三好 裕樹, 尾崎 知伸, 江口 浩二, 大川 剛直, “定量的アイテム集合付き単一グラフからの頻出パターンマイニング”, 人工知能学会論文誌, Vol.26,

No. 1, pp. 284-296, 2011. 査読有.
http://dx.doi.org/10.1527/tjsai.26.284

② 信田 正樹, 尾崎 知伸, 大川 剛直, “内部および外部重みを考慮した頻出部分グラフマイニング”, 情報処理学会論文誌 データベース, Vol. 3, No. 2, pp. 1-12, 2010. 査読有.
http://id.nii.ac.jp/1001/00070136/

③ 横山 正太朗, 江口 浩二, 大川 剛直, “潜在トピックを用いたブログ空間からの情報伝搬ネットワーク抽出”, 電子情報通信学会論文誌, Vol. J93-D, No. 3, pp. 180-188, 2010. 査読有.
http://search.ieice.org/bin/summary.php?id=j93-d_3_180&category=D&year=2010&lang=J

④ 麻生 竜矢, 江口 浩二, “学術文献の潜在トピックに着目したタンパク質相互関係に関する知識の抽出”, 情報処理学会論文誌 データベース, Vol. 2, No. 2, pp. 86-95, 2009. 査読有.
http://id.nii.ac.jp/1001/00060731

⑤ Atsuhiro Takasu, Daiji Fukagawa, and Tatsuya Akutsu, “Latent Topic Extraction from Relational Table for Record Matching”, Discovery Science: Proceedings of the 12th International Conference on Discovery Science, Vol. LNAI 5808, pp. 449-456, 2009. 査読有.
http://dx.doi.org/10.1007/978-3-642-04747-3_38

〔学会発表〕(計 5 件) (総計 21 件)

① 江口 浩二, “統計的言語モデルと情報検索” (チュートリアル講演), 第 3 回データ工学と情報マネジメントに関するフォーラム, 2011 年 2 月 27 日, 静岡県伊豆市. 査読無.

② 林 幸記, 江口 浩二, 高須 淳宏, “カテゴリ階層構造を考慮した確率的トピックモデルとその応用”, 情報処理学会第 200 回自然言語処理研究会・第 101 回情報基礎とアクセス技術研究会, 2011 年 1 月 28 日, 東京都世田谷区. 査読無.

③ 江口 浩二, “統計モデリングとデータマイニングに基づくネットワーク化知識の創出と活用”, 2010 年度科研・合同シンポジウム：言語処理技術の深化と理論･応用の新展開, 2010 年 9 月 28 日, 東京都文京区. 査読無.

④ 蜷川 陽, 江口 浩二, “大規模ネットワーク構造の確率的グループモデルに基づくリンク予測”, 情報処理学会第 17 回バイオ情報学研究会, 2009 年 5 月 26 日, 沖縄県国頭郡. 査読無.

⑤ 江口 浩二, 塩崎 仁博, “多重多型トピックモデルを用いたアノテーション付きテキストからのエンティティ検索”, 情報処理学会第 145 回データベースシステム研究会・第 91 回情報学基礎研究会, 2008 年 6 月 20 日, 北海道小樽市. 査読無.

〔図書〕(計 1 件)

① 江口 浩二, “文書クラスタリング”, 言語処理学事典, 共立出版, pp. 334-339, 2009 年 12 月.

〔その他〕
ホームページ等
http://www.prmir.scitec.kobe-u.ac.jp/

６．研究組織

(1)研究代表者
江口 浩二 (EGUCHI KOJI)
神戸大学・大学院システム情報学研究科・准教授
研究者番号：５０３２１５７６

(2)研究分担者
高須 淳弘 (TAKASU ATSUHIRO)
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：９０２１６６４８

大川 剛直 (OHKAWA TAKENAO)
神戸大学・大学院システム情報学研究科・教授
研究者番号：３０２２３７３８

(3)連携研究者
尾崎 知伸 (OZAKI TOMONOBU)
大阪大学・サイバーメディアセンター・特任講師
研究者番号：４０３６５４５８
(H20→H21：研究分担者)

宇野 毅明 (UNO TAKEAKI)
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：００３０２９７７