様式 C-7-2

自己評価報告書

平成２３年　４月２６日現在

機関番号：１４５０１
研究種目：基盤研究（Ｂ）
研究期間：２００８～２０１１
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．研究計画の概要

本課題は、統計モデリング技術とデータマイニング技術を駆使・拡張し、断片的に散在した情報コンテンツから、人間の知的活動に直接活用可能なネットワーク化知識を創出し、活用する手段の確立をめざす。この目的のもと、以下の 3 つの観点から研究に取り組んでいる。

(1) テキストデータからの関係構造の抽出：
不完全な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、単語間の関係やデータ間の関係を抽出する研究

(2) ネットワークデータに対するパターン発見：
頂点に離散確率分布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

(3) 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、リンク予測等の応用問題に適用する研究

２．研究の進捗状況

2008 年度から 2010 年度までに、以下の研究成果を得た。

(1) 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発見し、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。

(3) 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実証した。

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．現在までの達成度

①当初の計画以上に進展している。

（理由）

本研究課題の研究成果に関連して、34 件の雑誌論文（査読付き国際会議論文を含む）、21 件の学会発表、1 冊の図書（分担執筆）を発表している。これらのうち 4 件については招待講演・依頼講演であり、6 件については各賞を受賞している。

４．今後の研究の推進方策

2008 年度から 2010 年度までに実施してきた本研究課題を発展的に展開し、2011 年度以降は基盤研究(B)「大規模構造データに対する確率モデル推定に基づく知識の創出と活用」（前年度申請、課題番号 23300039）として、以下の研究項目を実施する。

(1) 内部構造付きテキストデータから潜在トピックを推測し活用する研究
(2) 多元ネットワークデータから潜在的頂点クラスタを推測し、リンク予測問題等へ応用する研究
(3) 外部構造付きテキストデータから潜在トピックを推測し、テキスト分類問題等へ応用する研究

５． 代表的な研究成果

（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 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. 査読有り.

[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. 査読有り.

[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. 査読有り.

[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. 査読有り.

[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. 査読有り.

〔学会発表〕（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. 査読無し.

[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. 査読無し.

[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. 査読無し.

[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. 査読無し.

[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. 査読無し.

〔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

http://www.prmir.scitec.kobe-u.ac.jp/