様式 C-7-2

自己評価報告書

平成２２年５月２６日現在

研究種目：基盤研究（Ｂ）
研究期間：2007～2010
課題番号：19300032
研究課題名（和文）　時系列多重トピックモデルによる情報共有法の研究
研究課題名（英文）　Study on Information Sharing Based on Multiple Topic Model for Text Stream
研究代表者
高須　淳宏（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

研究代表者の専門分野：情報工学
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：情報共有、多重トピックモデル、テキスト処理、機械学習

１．研究計画の概要

本研究は、複数の人間が係わるプロジェクトで生成・収集される各種情報を共有し活用するための情報共有システムの構築法を考案することを目的としている。通常の情報検索は、これらの情報を文書集合と考え、その中から目的にあった部分文書集合を効果的に検索することが主たる目的とする。しかし、プロジェクトの過程で得られる情報は問題解決の過程を反映したものであり、それらを時系列情報としてとらえることによってプロジェクトメンバに問題解決の過程をremindさせたり、同種のプロジェクトに携わる人間に問題解決へのヒントを与えるシステムの構築が期待できる。そこで、本研究では、特に時間情報を考慮した時系列文書の処理技術に焦点をあてて、情報共有システムを構築することをめざしている。

本研究では、以下の２つの課題を中心に研究を進めている。

（１）時系列文書モデルの提案とモデルの獲得支援法

まず、時間とともに文書に記述されているトピックがどのように変化しているかを把握するための時系列文書モデルを構築し、次にそのようなモデルをデータから学習するための効率の良いアルゴリズムを開発することが中心的な課題となる。

（２）情報共有システムの試作と評価

上記のモデルを用いて、情報共有システムを構築し、手法の評価を行う。本システムでは、従来の情報検索機能に加え、潜在トピックに基づいた時系列文書モデルも用いてトピック変遷マップを提示する機能を実装する。そして、このような多様な情報閲覧・提示機能が情報共有システムに及ぼす効果について分析する。

２．研究の進捗状況

本研究の中心課題である時系列文書モデルの構築と情報共有システムの試作を行っている。

（１）時系列文書モデル

時系列文書モデルにおいては、各文書はその内容と生成時間の組とみなされる。文書の内容については、基本的に、従来の研究と同様に文書を bag of words(BOW)とみなしている。本研究では、潜在トピックより BOW と時間の両方を生成する複数の確率モデルを考案し、その効果について検討を進めてきた。本研究で考案したモデルは、BOW の生成については、いずれも近年注目されている文書生成モデル Latent Dirichlet Allocation と同様に Dirichlet 事前分布と多項分布に基づいた確率モデルとなっている。一方、文書の生成時間については、①時間を一定の幅で区切り、この離散的な時間を用いてトピックと文書生成時間の関連を表すモデルと②時間軸上での連続的な確率分布を用いてトピックと文書生成を関連づけるモデルを考案した。また、時系列データからこれらのモデルを獲得するためのモデル推定アルゴリズムを開発した。トピックトラッキングの評価用コーパスを用いてこれらのモデルを評価したところ、時間軸に対して単峰的な分布関数を持つトピックについては高い精度で抽出できることがわかった。

（２）情報共有システム
　共有情報共有システムではさまざまな形式で記述された情報を扱う必要がある。これらの情報を上記のモデルで扱うためには、フォーマットの統一や情報の抽出が必要になる。そこで、文書から各種の属性を抽出するための情報抽出法の研究を進めた。ここでは、文書のレイアウトや構文構造に基づいて重要な情報を抽出すためのページ文法を提案しその効率的な構文解析アルゴリズムを開発した。
　次に、文書に現れる重要な情報を文書間で結び付けるための近似マッチングアルゴリズムを開発した。この研究では、類似度を計算するための統計的なモデルを考案し、そのモデルのパラメタをベイズ学習するためのGibbsサンプリングに基づくアルゴリズムを開発した。

３．現在までの達成度
②おおむね順調に進展している。
　当初の計画どおり時系列文書モデルを構築することができた。また、情報共有システムについては、より効果的にトピックを抽出し、その変遷を見つけるために、文書からの情報抽出に関する研究を行った。この課題は、計画立案段階では重視されていなかったが、トピック抽出の性能向上に効果があると考え取り組んだ。その結果、システムの試作に若干の遅れがあるが、全体として、おおむね順調に進展している。

４．今後の研究の推進方策
　今後は情報共有システムの試作を中心に研究を進めることを計画している。
（１）トピック変遷マップ構築機能
本研究で考案したモデルにもとづいてトピック変遷マップを提示する機能を実装する。また、トピック変遷マップ上で、各トピックにおける主要なキーワードの変化を追跡することによってトピック中の主題の変換を抽出することを試みる。
（２）プロジェクト参加者とトピックの関連度の抽出
プロジェクトに参加する各メンバと関連の強いトピックを結びつける機能を実装し、メンバに新規情報を推薦することを試みる。

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３件）
①竹田隆治、高須淳宏：複数文字列検知に基づいたSplogフィルタリング手法，情報処理学会論文誌　データベース，Vol. 2, pp.93-103, 2009，査読有

②A. Takasu, D. Fukagawa, T. Akutsu, Latent Topic Extraction from Relational Table for Record Matching, Lecture Note in Computer Science, Vol. 5808, pp. 547 - 560, 2009, 査読有

③薬師貴之、太田学、高須淳宏：ＣＲＦを用いた学術論文ＯＣＲテキストからの自動書誌要素抽出、情報処理学会データベース、Vol. 2, pp.126-136, 2009 査読有

〔学会発表〕（計１０件）
①T.Tekeda, A. Takasu: UpdateNews: A News Clustering and Summarization System Using Efficient Text Processing, ACM IEEE Joint Conference on Digital Libraries, 2007.6.10, Vancouver, Canada.

② A. Takasu, D. Fukagawa, T. Akutsu, Statistical Learning Algorithm for Tree Similarity, International Conference on Data Mining, 2007.10.29, Omaha, NE, USA

③M.Ohta, A.Takasu, CRF-based Authors' Name Tagging for Scanned Documents, ACM,IEEE Joint Conference on Digital Libraries, 2008.6.18, Pittsburgh, PA, USA

④ M.Ohta, A. Takasu: Bibliographic Element Extraction from Scanned Documents Using Conditional Random Filed, International Conference on Digital Information Management, 2008. 11. 13, London, UK.

⑤A. Takasu, Bayesian Similarity Model Estimation for Approximate Recognized Text, International Conference on Document Analysis and Recognition, 2009.7.29, Barcelona, Spain.

⑥M. Ohta, T. Hachiki, A. Takasu: Using Web Resource for Support of Online Browsing of Research Papers, IEEE International Conference on Information Reuse and Integration, 2009. 8. 12. Las Vegas, USA.