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２４年６月１８日現在

機関番号：62615
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 Method by Multiple Latent Topics

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

研究成果の概要（和文）：
本研究では、各種情報源から得られる時系列文書から、その背後にある潜在的なトピックを抽出するとともに、抽出されたトピックを用いた時系列文書の共有活用技術を提案することを目的としている。本研究では、この目的を達成するため、文書からの特徴語の抽出法、時間情報を考慮した潜在トピックモデルの構築、および、潜在トピックを用いた情報推薦法について研究を行い、時系列文書を効果的に共有活用する方法を提案した。

研究成果の概要（英文）：
The purpose of this study is to extract latent topics from text stream and to develop a method for sharing and utilizing stream documents. We studied this problem by (1) extracting feature phrases, (2) developing a latent topic model handling time information, and (3) applying it to information recommender systems. We proposed a information sharing and utilization method using latent topics.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	3,600,000	1,080,000	4,680,000
2008 年度	3,600,000	1,080,000	4,680,000
2009 年度	3,500,000	1,050,000	4,550,000
2010 年度	3,500,000	1,050,000	4,550,000
総　計	14,200,000	4,260,000	18,460,000

研究分野：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：テキスト処理、トピックモデル、機械学習

１．研究開始当初の背景

大量の文書から利用者が必要とする情報を効率的かつ効果的に取り出し利用する技術は、文書検索、文書クラスタリング、文書分類等さまざまな研究が行われている。また、Web の検索エンジンのように情報社会において基本的かつなくてはならない技術もある。これらの技術は、適宜更新されるものの、基本的にはデータベースに保存された静的な文書集合を扱っている。しかし、実際の文書は時事刻々変更されていく。そのため時間軸を考慮して文書間の関連を考慮して扱うことが必要になる。この研究の発足当初、時系列文書に対する解析・マイニング技術としては、ニュース記事を対象としたトピックの検出と追跡 (Topic Detection and Tracking) を目的とした研究が行われていた。この研究では、新聞記事を対象として、さまざまな事

件（イベント）に関する記事が時事刻々と配信される文書ストリームから事件ごとに記事をまとめたり、新しい事件に関する最初の記事を時系列文書の中から抽出する問題があつかわれた。これらの問題に対して、さまざまな手法が提案されたが、それらの多くは従来の文書クラスタリングの技術を応用したものが多く、文書の時間情報を陽に扱った技術はあまり見受けられなかった。

２．研究の目的

本研究では、さまざまな情報源から提供される時系列文書からそこに含まれる様々なトピックを抽出するとともに、抽出されたトピックを用いた時系列文書の共有活用技術を提案することを目的としている。トピック抽出に関しては、近年、統計的なモデルに基づいた文書からのトピック抽出法に関する研究が注目を集めている。本研究では、それらの最新の技術を用いて効果的かつ効率的な時系列文書の共有活用法を開発することを目指した。

３．研究の方法

本研究では、時系列文書を共有活用する方法を開発するために、問題を以下の３つの問題に分けて、研究をすすめた。

（１） 文書の特徴抽出

テキスト解析・マイニングにおいては、文書から特徴的な語を抽出する必要がある。一般には、文章中に現れる語の頻度(bag of words)に基づいて、文書の特徴を表すことが行われる。しかし、解析精度を上げるためには、文書のジャンルに適した特徴語の抽出も必要になる。そこで、本研究では、学術論文を対象の文書とした場合のその特徴である著者や所属といった語を抽出する方法について検討した。

（２） 時系列文書のトピックモデル

時系列文書を関連するグループにまとめるためのクラスタリング法について検討を行った。近年 Latent Dirichlet Allocation (LDA)と呼ばれる潜在トピックを用いた文書モデルが提案された。本研究では、このモデルを拡張し、時間情報を考慮した統計モデルの構築を試みた。

（３） 推薦システムへの応用

通常、時系列文書は情報量が多くなるため、利用者に必要な情報だけを選択的に提供することによって効果的に情報を共有することが可能になる。そこで、本研究で考案した時系列文書の解析法を情報推薦システムに応用し、その効果を試みた。

４．研究成果

（１） 文書の特徴抽出

文書は一般に文章中に現れる語の頻度に基づいて処理される。しかし、語には品詞以外にもその役割がある。たとえば、学術論文のような文書では、図１に示すようにタイトルページには著者や論文タイトルが含まれている。このような語やフレーズの役割を抽出することによって、その文書の特徴をより精度よく扱うことが可能になる。

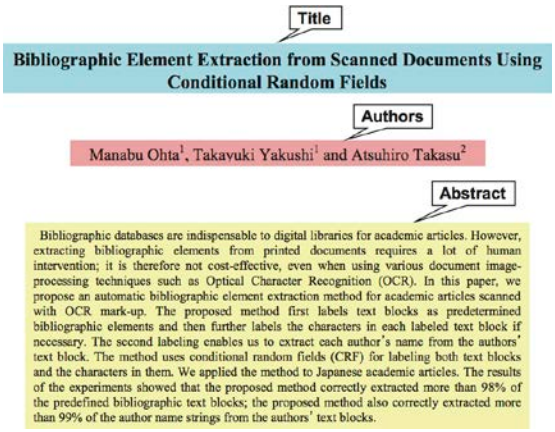

図 １　学術論文のレイアウト

本研究では、対象を学術情報のタイトルページに固定し、そこに含まれる著者、タイトル、アブストラクト等を抽出する方法について検討した。本研究で提案した方法は、以下に示す手順に従って情報を抽出する。

①タイトルページを行もしくは単語単位に分割し、１ページを行や単語の列としてみなす。

②行や単語単位の領域を囲む矩形領域を、長さ、幅、隣接する矩形領域との距離などを特徴とする特徴ベクトルで表す

③チェーンモデルの Conditional Random field のモデルをラベル付き矩形領域の列より求める

④ラベルのついてない矩形領域の列に③で求めた CRF を適用し、各矩形領域のラベル付けを行う。

⑤同一ラベルを持つ矩形領域をまとめることによって、タイトルや著者の領域を抽出する。

提案手法を情報処理学会論文誌の６８９件の論文のタイトルページを使って評価した。まず、これらの論文のタイトルページより行や単語を抽出し、手で各領域に正しいラベル（著者、タイトル等）を付与した。つぎに、280 件の論文を使って CRF を学習し、残りの 409 件の論文を用いて情報抽出精度を測定した。この実験では、著者、タイトル、アブストラクトの抽出精度を測定した。図２はラベルごとの抽出精度に加え、論文全体からすべての要素を正しく抽出できたかどうかを示す精度を表している。表にあらわされているように、多くのラベルが 99%以上の精度で抽出されていること、また、論文全体でも97%程度の精度で情報抽出が可能であること

がわかる。

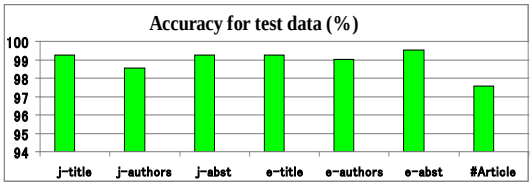

図 2　情報抽出制度

（2）時系列文書のトピックモデル

情報検索やテキストマイニングでは、文書は通常そこに含まれる単語の集合として扱われる。そして、その結果得られる単語頻度ベクトルを特徴ベクトルとして、ベクトル間の類似度を文書の類似度として検索や解析が行われる。通常、文書に含まれる単語の種類は膨大であるため、単語頻度ベクトルも高次元のベクトルとなる。

高次元ベクトルは測定される類似度が必ずしも適切でなかったり、その処理コストが高いといった問題があり、より低次元の特徴ベクトルに変換する方法が提案されてきた。たとえば、特異値分解を使って低次元ベクトルに変換する Latent Semantic Index や確率モデルを用いて低次元ベクトルに変換する probabilistic latent semantic index などが提案されてきた。これらの方法の特徴は、文書集合からデータ全体を効果的に表す factor の集合を抽出し、個々の factor を特徴とするベクトルに変換するところにある。

近年、ベイズ統計に基づいた統計モデル Latent Dirichlet Allocation (LDA)が提案され、情報検索やテキストマイニングに用いられている。LDA は確率的に文書を生成する言語モデルで、文書中の各語は潜在トピックより生成されるものと仮定されている。同一文書中の潜在トピックは、多項分布 θ に従ってランダムに生成され、各語は、トピック t ごとに定められた多項分布 ϕ_t に基づいて生成される。LDA は Bayesian model であり、これらの多項分布は、それぞれ Dirichlet 事前分布に従って生成される。図 3 は LDA のグラフィカルモデルを表している。

LDA は、情報検索やテキストマイニングの問題で優れた性能を示すことが報告されており、さまざまな問題への適用が試みられている。LDA を本研究の問題である時系列文書に適用するにあたってひとつの問題がある。それは、LDA が時間情報を陽に扱っていないことである。時系列文書では、特定のトピックはある時期にまとめて出現することが多いと予想される。しかし、LDA は時間情報を扱っていないため、時間的に離れた文書でも、時間的に隣接した文書でも、トピックが同様の確からしさで出現することになる。本研究では、各文書にタイムスタンプを付与し、このタイムスタンプも潜在トピックから生成するモデルを提案した。特に隣接したタイムスタンプを文書で共起させることによって、特定の潜在トピックは、対応する特定の時期により高い確率で出現するモデルを構築した。

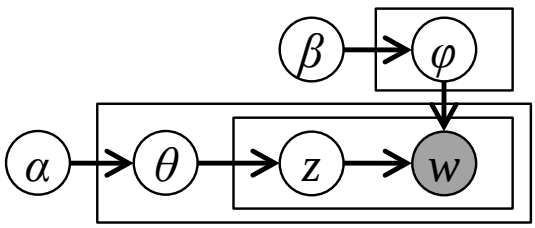

図 3　LDA のグラフィカルモデル

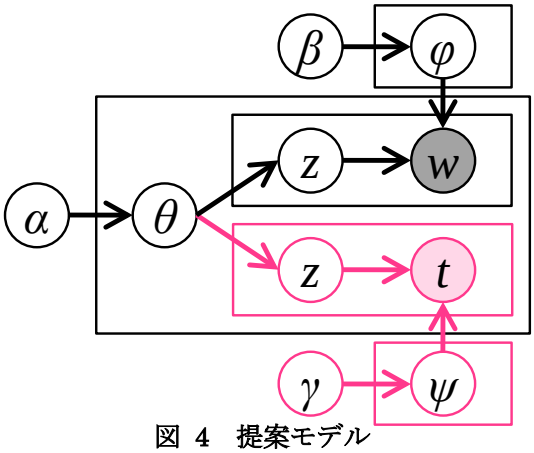

図 4　提案モデル

のグラフィカルモデル

図 4 は、このモデルのグラフィカルモデルを表している。図 3 の LDA と比較すると、単語 w に加えて、タイプスタンプ t が生成されている部分が異なっている。タイムスタンプも単語同様、潜在トピック z より生成され、潜在トピックは単語の潜在トピックを生成する多項分布 θ と同一の分布から生成されている。

本研究では、このモデルを使って 56,755 件の日本語の新聞記事コーパスの解析をした。図 5 は、その結果を表している。横軸は

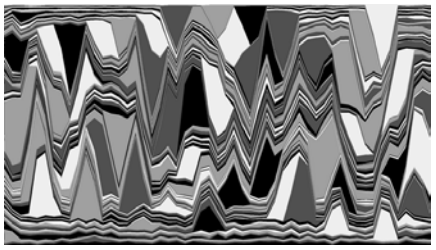

図 5　トピックの時間分布

実験に用いた新聞記事が発行された時間を

表している。一方、縦軸は、各トピックの各時間での割合を表している。幅の広いトピックは、その時期に集中的に新聞記事に取り上げられたトピックを表している。時間情報を陽に扱わないLDAのトピックの出現割合の時間変化と比較すると提案手法のほうが、より特定時期に集中してトピックが現れることが判明した。

（3）推薦システム

時系列文書は一般に増加の一途をたどるため、必要な時に必要な情報を利用者に届けることが重要になる。このような機能を提供する情報フィルタリングの研究では、これまでに様々な方法が提案されてきた。近年は、利用者のモデリングを重視した情報推薦システムの研究が注目を集めている。情報推薦システムでは、まず、利用者の嗜好を表す利用者プロフィルを利用者の行動から獲得する。このプロファイルを用いて利用者の必要とする情報を選択して届けることになる。しかし、利用者から得られるログは限られているため、利用者に類似した他の利用者(neighbor)を探し、その利用者のログも活用して情報の選択を行う協調フィルタリングの研究が活発に行われている。

類似する利用者を見つけるために、潜在トピックを使って膨大な利用者を比較的少数のグループに分ける方法が提案されてきた。本研究では、特に利用者からの行動ログが限られている場合に、そこから類似利用者を効果的に発見する方法を中心に研究を進めた。これまでは文書を対象として扱ってきたが、この研究では、さらに本や映画などの文書以外のアイテムも対象とした。提案手法は、利用者およびアイテムを同時に潜在トピックを用いて比較的少数のグループに対応づけ、グループごとの嗜好を統計モデルで表した。具体的には、利用者が特定アイテムに評点を与えたものを観測データと考え、(利用者、アイテム、評点）の３項組を観測データの最少単位と考えた。そして、この３項組みの集合を訓練データとしてモデルを学習する。図６に、このモデルのグラフィカルモデルを示す。図中の$r_1, \cdots, r_n$は利用者がアイテムにつけた評価値を表している。利用者はn個の視点からアイテムを評価することを想定している。z_u, z_vは、それぞれ利用者グループ、アイテムグループを表す潜在トピックを示している。θ、ϕは、これらの潜在トピックを生成する多項分布を表しており、$v_1, \cdots v_n$はトピックから評価値を生成する確率分布を表している。LDA と同様に各確率分布には対応する事前分布が定義されている。

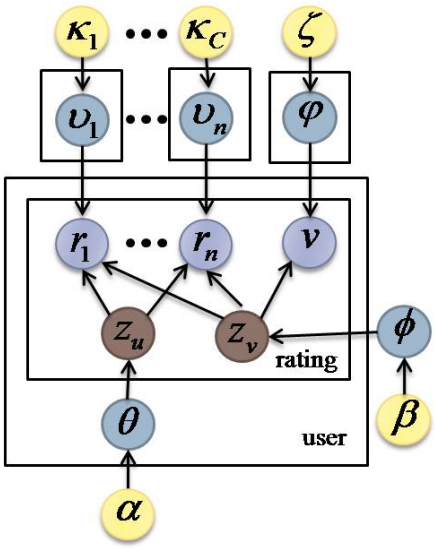

図 6　推薦システムのグラフィカルモデル

このモデルを、この分野の代表的な評価コーパスである Yahoo! Movie データに適用して、その性能を測ったところ、特に利用者のログデータが少ない場合に提案手法の性能が従来手法として優れていることがわかった。

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 4 件）

① 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、情報処理学会論文誌データベース、査読有、2 巻、2009，pp. 126 - 136.

② 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、査読有、2 巻、1 号、2009，pp. 93 - 103.

③ 篠原 靖志，高須 淳宏:
″効率的能動学習のためのサポートカーネルマシーン”，電子情報通信学会論文誌，J91-D，10 号、2008，pp. 2497-2506.

④ 正田 備也，高須 淳宏，安達　淳
″混合ディリクレイ分布を用いた文書分類の精度について″ 情報処理学会論文誌：データベース、査読有、 48 巻，SIG 11 号，2007 ， pp. 14-26.

〔学会発表〕（計 13 件）

① Atsuhiro Takasu: " A Multicriteria Recommendation Method from Data with Missing Rating Scores" Intl. Conference on Data and Knowledge Engineering (ICDKE 2011)，査読有，

2011.9.6, Milan, Italy.

② Atsuhiro Takasu, Saranya Maneeroj: "A Recommendation Algorithm Using Positive and Negative Latent Models" IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011), 査読有, 2011.4.15, Paris, France.

③ Atsuhiro Takasu: "Cross-lingual keyword recommendation using latent topics" Intl. Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2010), 査読有, 2010.9.15, Barcelona, Spain.

④ Pakapon Tangphoklang, Saranya Maneeroj, Atsuhiro Takasu: "Advanced Representative and Dynamic User Profile Based on MCDM for Multi-Criteria RS" Intl. Conference on Information Systems (Information Systems 2010), 査読有, 2010.3.18, Porto, Portugal.

⑤ Tomonari Masada, Daiji Fukagawa, Atsuhiro Takasu, Tsuyoshi Hamada, Yuichiro Shibata, Kiyoshi Oguri: "Dynamic Hyperparameter Optimization for Bayesian Topical Trend Analysis" ACM Conference on Information and Knowledge Management (CIKM 2009), 査読有, 2009.10.5, Hong Kong.

⑥ Quang Minh Vu, Atsuhiro Takasu, Jun Adachi: "A Versatile Record Linkage Method by Term Matching Model Using CRF" Database and Expert Systems Applications (DEXA 2009), (LNCS 5690), 査読有, 2009.9. 12, Linz, Austria.

⑦ Tomonari Masada, Atsuhiro Takasu, Tsuyoshi Hamada, Yuichiro Shibata: "Bag of Timestamps: A Simple and Efficient Bayesian Chronological Mining" Joint Conference on Asia-Pacific Web Conference and Web-Age Information Management (APWeb-WAIM 2009), 査読有, 2009.4.3, Suzhou, China

⑧ Quang Minh Vu, Atsuhiro Takasu, Jun Adachi: "Name Disambiguation Boosted by Latent Topics from Web Directories" IEEE/WIC/ACM Intl. Conference on Web Intelligence (WI2008), 査読有, 2008.12.15, Sydney, Australia.

⑨ Manabu Ohta, Atsuhiro Takasu: "CRF-based Authors' Name Tagging for Scanned Documents" Joint Conference on Digital Libraries (JCDL 08), 査読有, 2008.6.27, Pittsburgh, USA.

⑩ Atsuhiro Takasu, Kenro Aihara: "Information Extraction from Scanned Documents by Stochastic Page Layout Analysis" ACM Symposium on Applied Computing (SAC 2008), 査読有, 2008.3.26, Fortaleze, Brazil.

⑪ Manabu Ohta, Shun Yamasaki, Takayuki Yakushi, Atsuhiro Takasu:"Authors' Names Extraction from Scanned Documents" Intl. Conference on Digital Information Management (IEEE ICDIM'07), 査読有, 2007.10.23, London, UK.

⑫ Takaharu Takeda, Atsuhiro Takasu: "UpdateNews: A News Clustering and Summarization System Using Efficient Text Processing" ACM IEEE Joint Conference on Digital Libraries (JCDL 2007), 査読有, 2007.6.20, Vancouver, Canada.

⑬ Tomonari Masada, Atsuhiro Takasu, Jun Adachi: "Citation Data Clustering for Author Name Disambiguation" Intl. Conference on Scalable Information Systems (INFOSCALE 2007), 査読有, 2007.6.12, Suzhou, China.

６．研究組織

(1)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計・教授

研究者番号：90216648

(2)研究分担者

なし

(3)連携研究者

相原　健郎（AIHARA KENROU）

国立情報学研究所・コンテンツ科学研究計・准教授

研究者番号：90300706