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　５月１５日現在

機関番号：１７３０１
研究種目：若手研究（Ｂ）
研究期間：2010～2011
課題番号：２２７００１５０
研究課題名（和文）　統計学的ライムを利用した情報ナビゲーション

研究課題名（英文）　Information Navigation using Statistical Rhymes

研究代表者
正田　備也（MASADA TOMONARI）
長崎大学・大学院工学研究科・准教授
研究者番号：60413928

研究成果の概要（和文）：本研究は、「意味的な関連性によるのではない単語の共起関係であっても，統計学的に有意な頻度で生じているならば情報収集の手掛かりとして有用性を持つ」という仮定に基づいている。この、統計学的に有意な頻度で生じる共起を、「統計学的ライム」と呼ぶ。そして、ベイズ的な確率モデルを使い、統計学的に有意な頻度で生じている単語の共起関係を抽出することを目指した。最終的に、論文末尾や研究者の Web サイトに現れる書誌情報を、著者名・論文タイトル・学術雑誌名・発表年など異なる書誌フィールドへと教師無し学習によって自動分割する、新しい LDA（潜在的ディリクレ配分法）タイプのトピック抽出法を提案できた。また、提案のモデルの分割精度を半教師付き学習により改善することに成功した。

研究成果の概要（英文）：This project is based on the following assumption: Words that co-occur in statistically significant frequency can be used as a guide in useful information navigation system even when those co-occurrences are not based on semantic similarity or relatedness. We call such co-occurrences *statistical rhyme*. We have been trying to extract statistical rhymes with Bayesian probabilistic models. We consequently succeeded in proposing a new LDA(latent Dirichlet allocation)-like topic extraction method that can give a segmentation of word token sequences appearing in bibliographic data, which we can observe in references section of academic papers or in publications section of researchers' Web sites. Our method split each bibliographic data into the segments each corresponding to different data field, e.g. authors, paper title, journal, pages, publication year, etc. Further, we improved segmentation accuracy by making the inference semi-supervised.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	2,600,000	780,000	3,380,000
２０１１年度	500,000	150,000	650,000
総　計	3,100,000	930,000	4,030,000

研究分野：知能情報学
科研費の分科・細目：総合領域、情報学
キーワード：データマイニング、確率モデル、ベイズ理論、トピックモデル、並列化

１．研究開始当初の背景

生まれたときから土砂降りの情報を浴び続ける世代にとって、ネットから得られる情報はあまりに大量で、個々の情報の意味を噛みしめては次に進むというように、意味というスピードの遅いロジックによって消化できる量ではない。意味よりも速く情報を手繰り寄せる新たなロジックが必要と思われる。

実際、特にソーシャルメディアでは、単に最新の情報だから、単に友達の友達が書いた情報だから、等々、情報の意味内容にとって外的な関連性（メタデータ上の関連性）を頼

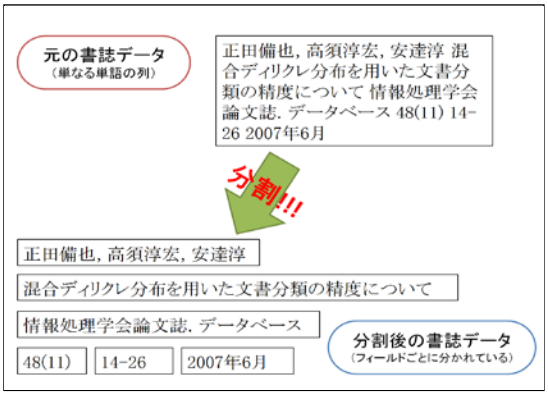
りに大量の情報を手繰り寄せてはじめて、砂漠化した情報の深い砂嵐の中に、オアシスのように特定の意図やイメージなど、意味内容的なものが浮かび上がることが多い。

例えば、ラップにおけるライムは、単に読みが同じという理由で、意味的な関連は乏しい言葉を頻繁に共起させる。同様に、意味的な関連性とは別の関連性ではあるが、統計学的に有意な頻度で生じる関連性によって情報を手繰り寄せるしくみを、本研究では提案しようとした。研究課題名にある「統計学的ライム」とは、情報の意味内容にとって外的な関連性のことを言おうとした概念である。

２．研究の目的

われわれは、上記のフィロソフィーを礎に、データ間の外的関連性、つまり、データのデータ＝メタデータに基づく関連性を利用して、膨大なデータの中から一定の法則性を発見することを、研究の目的とした。試行錯誤はあったが、最終的に以下のような、極めて具体的な問題設定にたどりついた。

学術論文の末尾の参考文献一覧や、研究者が作成した発表論文の Web ページからは、膨大な書誌データ（個々の論文の著者名、タイトル、発表学会や発表雑誌、ページ数、発表年など）が取得可能である。しかし、これら書誌データをデータベースに登録するには、単なる単語の列にすぎない書誌情報を、異なる書誌フィールドへと分割（セグメンテーション）しなければならない。



そこで、本研究ではこの問題に**教師無し**の解法を与えることを目標として設定した。

この問題は、単なる単語の列としての書誌データの中に、書誌フィールドというメタデータを見つけ出そうとしている。その際、個々の書誌データがどの分野に属する論文のデータであれ（コンピュータ・サイエンスで言えば、アーキテクチャ分野であれ、人工知能分野であれ、システム・ソフトウェア分野であれ、等）著者名フィールドにはこの単語が現れやすい、タイトルフィールドにはこの単語が現われやすいなど、各分野の意味内容的区別とは無関係な統計学的に有意な頻度で生じる現象を的確にとらえることが重要である。この意味で、本研究の元々の意図、つまり「統計学的ライム」の利用という意図に沿った問題設定になっている。

書誌フィールドとしては、著者名・タイトル・発表雑誌名などが考えられるが、これらのフィールドへと各書誌情報を分割するにあたって、本研究では、以下のような厳しい状況設定を仮定した。

1. どのような書誌フィールドが存在しているかは、分からない。
2. さらに、各書誌データに書誌フィールドがどの順番で現われるかも、分からない。
3. 分かっているのは、書誌データがいくつの書誌フィールドに分割されるか、その**個数だけ**である。
4. 書誌フィールドの個数は分かっているが、すべての書誌データがその個数に分割されるとは限らない。ページ数など一部の書誌フィールドを含まない書誌データもあるかもしれない。

関連技術の現状としては、書誌データを異なるフィールドへ分割する問題は、すでに何度も解かれており、その結果としてCiteseer、Google Scholar、Microsoft Academic Search などのシステムが実際に運用されている。しかしながら、上記のように**ほとんど事前の知識を前提しないで**、書誌フィールド分割問題を解く試みは、本研究が初めてと思われる。

３．研究の方法

本研究では、LDA（latent Dirichlet allocation：潜在的ディリクレ配分法）を拡張した確率モデルを利用することで、書誌情報分割問題に、教師無しの解法を与えた。各トピックが異なる書誌フィールドに対応すると考えれば、LDA 的なトピックモデルを書誌フィールド分割問題に適用できる。

具体的には、Chen ら[Chen+ NAACL HLT2009]が提案した確率モデルを利用している。ただし、Chen らは本研究が解こうとしている問題とは異なる問題を解くために、このモデルを提案した。したがって、同じ確率モデルを使うにしても、使い方を問題に合わせて変更する必要があった。詳細は、研究代表者による論文[Masada+ WISS2010, Masada IJOCI2011, Masada+ ICADL2011]を参照されたい。一言で言えば、データの性質に合わせて、ディリクレ事前分布のハイパーパラメータの設定方法を大きく変え、トピックの分布が dense になるように設定した。

評価実験では、書誌フィールドへ分割されていない大規模な書誌データ集合を持ち合わせていなかったので、DBLPのデータ（http://dblp.uni-trier.de/xml/）、および、MEDLINE のデータ(MEDLINE®/PUBMED®, a database of

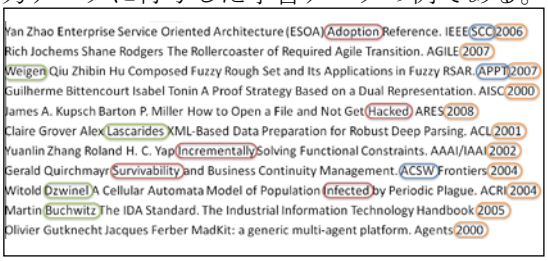
the U.S. National Library of Medicine)から、人工的に未分割の書誌データを作成した。提案手法の分割精度は、Fスコア（precisionとrecallの調和平均）で評価した。

先に述べたように、本研究で解くことを目指した問題は、すでに解かれている問題である。そして、いくつかの書誌データベースが、すでに実用レベルで利用可能であるから、書誌フィールドへの分割精度はすでにかなり高いはずである。しかし、本研究では、書誌フィールドの個数しか分からない、という極めて乏しい前提知識の下で、どの程度の分割精度が出せるか、明らかにしようとしている。これは、例えば、広く使われている作法に則らず作成された書誌情報が混在する書誌データ集合を分析する、等の状況に対応しており、必ずしも非現実的でない。むしろ、予想外のふるまいをするデータの分析はデータマイニングの重要な課題である。

とは言え、やはり書誌フィールドの個数しか分からない状況では、高々80%強の精度しか達成されなかった。この、純粋に教師無し学習を適用した結果は[Masada+ WISS2010]にて発表されている（なお、この論文の拡張版[Masada IJOCI2011]では、教師無し隠れマルコフモデルとの比較がおこなわれている）。

そこで、研究方法を再考し、分割のための手がかりを少し与え、半教師付き学習にすることで、精度を向上させることを目指した。[Masada+ ICADL2011]

次図が、半教師付き学習を実現するため入力データに付与した学習データの例である。

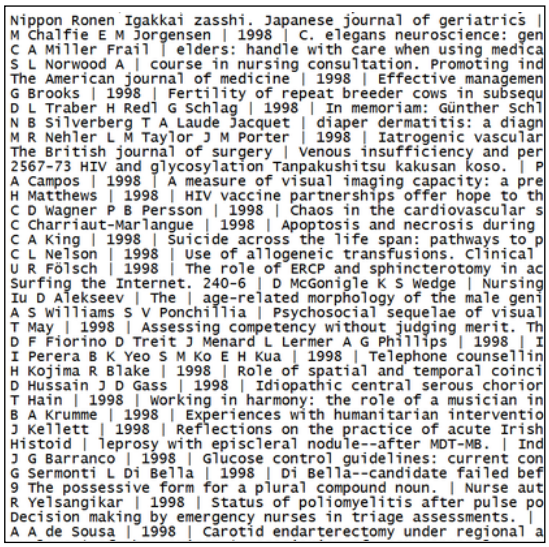


前もって、著者名にしか現れない単語、タイトルにしか現れない単語など、それぞれのフィールドにのみ現れる単語を、既存の書誌データ集合から辞書として抽出する。　そして、入力データの中にこの辞書に含まれる単語が現われていれば、そのすべてについて、自動的に対応する書誌フィールドを教師信号としてラベル付けしてしまう。

もちろん、このようなラベル付け方法では、誤った教師信号も混在する。例えば、MEDLINEデータから作成したデータセットでは、13%の単語トークンに上図のように教師信号として書誌フィールドを付与できたが、そのうちの 1/40 程度は誤りであった。しかし、提案手法は、元は教師なし手法なので、誤った教師信号を修正する可能性も期待でき、ここが実験のポイントとなる。

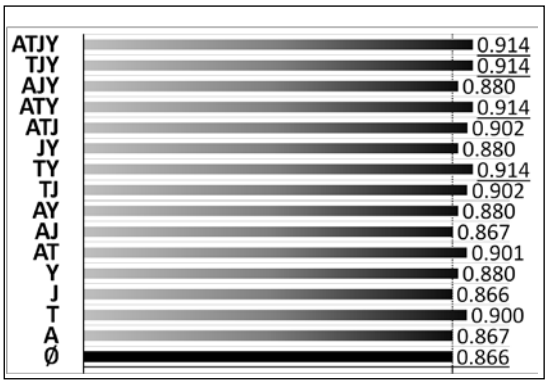
４．研究成果

まず、分割結果の例を下に示す。この図で、各行は別々の書誌データを表わしている。そして縦棒が、提案手法によって推定された書誌フィールドの切れ目である。つまり、縦棒を除いたものが元の入力データである。100%正解とはいかないが、かなり高い精度で分割が果たされていると分かる。なお、この図に示したものは、半教師付き学習にした後の提案手法による分割結果である。DBLP を元に作



成したデータセットでは F スコアで 91%の分割精度、MEDLINE を元に作成したデータセットでは 94%の精度を実現できた。

下のグラフは、DBLP を元に作成したデータセットでの評価結果である。最高で 91.4%の分割精度を達成できている。

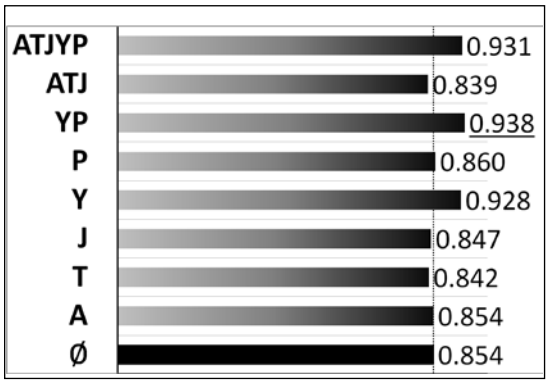


上のグラフで、左列の A,T,J,Y などの記号は、教師信号として、どの書誌フィールドの正解を与えたかを示している。A が著者名、T が論文タイトル、J が雑誌名、Y が発表年に対応する。例えば「TY」という記号が付された棒グラフは、論文タイトルと発表年についてのみ教師信号を与えてパラメータ学習した場合の精度である。この 0.914 という値が、「ATY」や「TJY」と変わらないことは、A や J、つまり、著者名や雑誌名などのフィールドに関する教師信号を与えても、分割精度の改良に寄与しないということである。逆に言

えば、論文タイトルにだけ出てくる単語とはどのような単語か、また、発表年にだけ出てくる単語とはどのような単語か、という情報が分割の精度改良に本質的に寄与するということである。なお、「∅」は教師無し学習の場合の分割精度である。また、この DBLP から作成したデータについては、各書誌データが 4 つのフィールドに分割されることは既知（つまり、トピック数は 4）としている。

例えば、実験では、発表年に関しては「1900」から「2012」の 4 ケタの数字である単語トークンに自動的に「発表年」という教師信号を与えている。もちろんこれが間違うこともある。論文タイトルに年号のような 4 ケタの数値が現われることもあるからである。しかし、実験結果は、この教師信号が有用であることを示している（「∅」0.866→「Y」0.880 へ改良）。実際、他の書誌フィールドに現れているにも関わらず、誤って発表年と教師信号を付与されたトークンの、約半分について、提案手法は正解書誌フィールド（発表年ではないフィールド）を与えている。

もう一度、上の棒グラフを見ると、タイトルにのみ現われる単語について付与された教師信号のほうが、大きな効果をもたらしている（「∅」0.866→「T」0.900 へ改良）。ただ、下図の MEDLINE データの場合の結果を見ると、発表年について教師信号を与えるほうが、データセットによらず重要と考えられる。



このグラフは、MEDLINE を元に作成したデータでの実験結果を示している。DBLP の場合と異なり、ページ数（「P」）というフィールドを含めて、計 5 つのフィールドへと各書誌データが分割されると想定している。つまり、トピック数は 5 としている。発表年（「Y」）を用いると、教師無しの場合の 0.854 に比べ、0.928 と分割精度が飛躍的に向上する。これは DBLP の場合と同様である。

しかし興味深いのは、単独ではあまり分割精度向上に寄与しない（0.854→0.860）ページ数（「P」）が、発表年と組み合わさると、1 ポイント（「Y」0.928→「YP」0.938）の精度向上をもたらす点である。この結果は、書誌フィールド分割問題が、フィールド間の関係性にも影響されながら解かれるべき問題であることを示唆していると思われる。

今後の課題としては、推定計算の高速化が挙げられる。現段階でも、事後分布のパラメータ推定は、OpenMP を用いて並列化されている。推定計算はギブス・サンプリングによっており、厳密に言えば、並列化により並列化前と同等の計算ではなくなるが、今回の実験では、並列度が高々12（Intel Core i7 の仮想コア数）のため、推定結果にはほぼ影響がなかった。しかし、逆に言えば、この程度の並列化にとどまっているため、計算の高速化は十分と言えない。例えば、研究代表者は過去に LDA の CVB という推定方法を GPU を用いて高速化している[Masada+ IEA/AIE2009]。この手法をそのまま適用することは、推定方法がギブス・サンプリングと CVB とで異なるため不可能であるが、GPU を用いた推定の高速化を果たせれば、より大規模な書誌データ集合への対応も可能となるだろう。これが今後の課題である。

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１件）

① 正田備也、潜在的置換による書誌要素の教師無し分割（〔学会発表〕①の拡張版）、IJOCI、第 2 巻、第 2 号、49-62 頁、2011 年、査読有

〔学会発表〕（計２件）

① 正田備也、柴田裕一郎、小栗清、潜在的置換による書誌要素の教師無し分割、WISS2010（中国・香港）にて発表、査読有、シュプリンガー・レクチャー・ノーツ・イン・コンピュータ・サイエンス 6724 巻、254-267 頁、2010 年 12 月 12 日

② 正田備也、高須淳宏、柴田裕一郎、小栗清、潜在的置換による書誌要素の半教師付き分割、ICADL2011（中国・北京）にて発表、シュプリンガー・レクチャー・ノーツ・イン・コンピュータ・サイエンス 7008 巻、60-69 頁、2011 年 10 月 25 日（論文採択率 25%）

〔その他〕

以下は、本研究の成果を含む内容が表示されている、研究代表者の Web サイトである。
http://diversity-mining-lab.wikispaces.com/

６．研究組織

(1)研究代表者

正田 備也（MASADA TOMONARI）
長崎大学・大学院工学研究科・准教授
研究者番号：60413928