LF-003

分散・ヘテロなデータからのトピック全体構造の学習

Learning Global Topic Structure from Distributed and Heterogeneous Data

松村 憲和† 森永 聡† 山西 健司†
Norikazu Matsumura Satoshi Morinaga Kenji Yamanishi

1. はじめに

複数のリモートサイトに蓄積されているテキストデータに含まれる情報を統合して、全体としてどのようなトピックが存在しているかを精度よく分析するフレームワークを提案する。その際、サイトとの通信量を出来るだけ少なくしてプライバシーを保護することを条件とする。

現実の環境では、単一のサイトに蓄積されているデータのみならず、複数サイトに分散蓄積されているデータをセンタで統合して分析することにより、全体を俯瞰する情報を発見することが重要となる場合がある。例えば、組織別に蓄積されている活動報告書群から、組織全体における活動内容を分析する場合などが相当する。この場合、各データセット間では使用語彙集合が異なるなど、サイト間で非均質性（ヘテロ性）が存在するのが通常である。また、通信量やプライバシー等の問題から、各サイトの生データ全体を一箇所に集めること自体が現実的ではないことも多い。このような状況、すなわち、1)ヘテロ性を持つ分散蓄積されたデータに対して、2)生データを一箇所に集めることなく、3)なるべく少ない通信量で、4)一箇所に集めた時と同程度の精度で分析する、ことは非常に重要である。

本報告では特に、分散蓄積されているテキストデータから、全体としてどのようなトピックが存在しているかの分析を、上記条件のもとで行うフレームワークを提案し、実データを用いてその有効性を検証する。ここで「トピック」とは特定の事象や活動について述べたテキスト群と定義する。本フレームワークの特徴は、テキストデータの A)複数のトピックが混在する構造を有限混合モデルでモデル化し、B)各サイトからはデータの集約を通じて情報を受取とり、C)これを基に生成される中間生成分布からのサンプリングを通じて全体構造を学習することによって、上記 1)－4)を実現するところにある。（図 1）

関連研究として、Clifton et al.の分散 EM アルゴリズム[1]、Yamanishi の分散協調学習[2]、Morinaga et al.の分散協調マイニング[6]、トピック検出[5]がある。[1],[2]の手法では、サイトとの通信を繰り返すことによって非常に高精度の分析を実現するが、上記条件の 3)が満たされず、プライバシーの面等において問題がある。[6]の手法はサイト間の類似性に事前知識が必要であり、適用ケースが限られる。[5]の手法は有限混合モデルによるトピック分析を実現するが、分散ヘテロな状況は想定されていない。

2．トピック全体構造のモデル化

サイト A におけるテキストデータの語彙集合を $W^A=\{w_1^A,w_2^A,\cdots,w_{\dim A}^A\}$ とし、そこでの各テキストデータは文書ベクトル $x^A=(x_1^A,x_2^A,\cdots,x_{\dim A}^A)$ で表すものとする。例えば x_i^A は語彙 w_i^A がその文書内に現れたら 1、現れなければ 0 の値をとる。$\dim A$ はサイト A 内で使用された語彙の数を表す。各サイトで語彙集合が異なるため、サイト間で文書ベクトルの何番目の成分がどの語彙に対応するかについて整合性はない（ヘテロ性）。このヘテロ性に対処するために、本フレームワークでは、各サイト l の語彙集合 W^l をセンタに送り、センタにおいて統合語彙集合 W を構成した上で、$w^l \to w$ の属性対応表を作成する。対応表に基づき、特定語彙が各サイトのベクトル成分の何番目に当たるかを整合させることによって、x を統合語彙集合の下で定義された文書ベクトルとみなす。以下では、文書ベクトルは統合語彙集合の下で定義されているとする。

i 番目のトピックを述べている文書ベクトル x は確率分布 $p_i(x|\theta_i)$ に従うものとモデル化する。θ_i は実数値のパラメータベクトルである。x がどのトピックに属しているか未知の場合は、K 個のトピックからなる有限混合分布 $p(x|\theta,K)=\sum_{i=1}^{K}\pi_i p_i(x|\theta_i)$ に従うものとモデル化する。π_i は i 番目のトピックの出現確率である。x の実現値の集合が与えられた場合に、未知のパラメータ $\theta_1,\cdots,\theta_k$ と $\pi_1,\cdots,\pi_k$ を推定することが、トピック構造を学習することに相当する[5]。ここでは、実現値の集合が各サイトに分散蓄積されているとする。各サイトの集約情報が集められてトピック全体構造を学習する場所をセンタとよぶ。

有限混合モデルの学習には一般に EM アルゴリズム[7]が使用されるが、それには全データがセンタに集められている必要がある（これを一括型 EM 法とよぶ）。これに対して Clifton et al. [1]は、サイト、センタ間で情報の送受信を複数回繰り返しながら一括型 EM 法を厳密に実行するアルゴリズムを提案している（これを分散 EM 法とよぶ）。

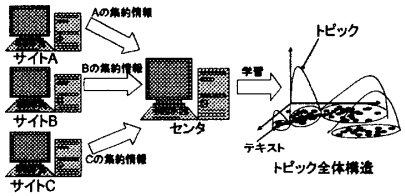

図 1: 分散・ヘテロなデータからのトピック全体構造の学習

3. 提案学習手法

ここでは一括型 EM 法と同程度の学習精度で実現するサイトからセンタへ一方的に情報が流れる学習法を提案する。

3.1 サンプルデータ法

サンプルデータ法では、1)各サイトが自分のデータに基づき有限混合モデルのパラメータをそれぞれ学習し、2)そのモデルからサイトのデータ数の比に従う少数のサンプリングデータを発生させ、3)そのデータをセンタに送り、4)センタでそれらのデータから有限混合分布を学習するという手順でトピックの全体構造を学習する。これにより、生データを集めることなく少ない通信量で分析が実現できる。

3.2 局所統合法

局所統合法では、1)各サイトが自分のテキストデータに基づき有限混合モデル $p^l(x|\theta^l)$ のパラメータを学習し、2)

† 日本電気株式会社, NEC Corporation

学習結果のパラメータをセンタに送り、3)センタで中間生成分布 $p(x|\theta)=\sum_{l=1}^{s}\frac{n_l}{N}p^l(x|\theta^l)$ を構成し、4)この分布に従ってサンプリングを行う。ここに n_l をサイト l のデータ数、N を全サイトのデータ数、s を全サイト数とする。5)センタでサンプリングデータから有限混合分布を学習するという手順で、トピックの全体構造を学習する。これにより、生データを集めることなく少ない通信量で分析が実現する。

ここで、サンプルデータ法と局所統合法はサンプリングが行われる場所が各サイトかセンタかの違いのみであるので、分析精度等は同じとなることに注意せよ。また、各サイトの学習結果はサンプル生成のみに用いられるため、センタ、各サイトが同じタイプの有限混合モデルである必要は無い。つまり、各サイトで既に独自のトピック分析を行っているのであれば、サイト間の整合性や都合を考慮することなく、結果をそのまま利用できる。

4. 手法評価

4.1 通信量の比較

サイト、センタ間の通信量に関して、分散 EM 法、サンプルデータ法、局所統合法の比較を行う。表 1 は通信量に関する 3 手法の比較を表す。e は EM アルゴリズムの反復回数、n' はサンプリングデータ数である。表より、分散 EM 法と比較してサンプルデータ法は、サイトへの送信情報を必要としないこと、テキストデータの dim は数百～数千となるため、$n_l' < e \cdot K \cdot \dim$ となるケースが多く通信量が少ないことから、優位であることがわかる。局所統合法も、サイトへの送信情報を必要としないこと、ほぼ全てのケースで $K_l < e \cdot K$ となり通信量が少ないことから、分散 EM 法より優位である。プライバシー保護に関しても、サンプルデータ法、局所統合法は、送受信する情報量が分散 EM 法より少ないことから、優れているといえる。

表 1: 3 手法の通信量の比較表

手法	センタへ送る情報	サイトへ送る情報	センタへの通信量
分散 EM 法	指定された統計量	全体のトピック	$O(e \cdot K \cdot \dim^2)$
サンプル法	サンプリングデータ	なし	$O(n_l' \cdot \dim_l)$
局所統合法	有限混合分布パラメータ	なし	$O(K_l \cdot \dim_l^2)$

4.2 分析精度の比較（広報データによる実験）

NEC の広報データ（2002 年～2004 年まで 1186 件のタイトル部分）を用いて、全データをセンタに集めて一括型 EM 法を行った結果と、提案手法の結果を比較した。各 p_i は次元毎に独立なベルヌーイ分布で、各サイトには各年度の広報データが蓄積されているとする。一括型 EM 法と提案手法によるそれぞれの学習結果に対して、各データを最大事後確率を持つコンポーネントに振り分けることでクラスタリングを行い、結果の類似度を ARI[4]により計算した。ARI は最大 1 平均 0 であり、クラスタリング結果が類似している程値が大きく、0.5 程度で十分類似していると判定してよい。

図 2 は提案手法でサンプリングデータ数を 10～700 と変化させた時の類似度を示したグラフである。サンプリングデータ数が 600 辺りで ARI が約 0.5 に収束する。これにより、サンプリングデータ数を増やすことで提案手法の結果は一括型 EM 法に一定のレベルまで近づくことが確認された。図 3 はサイト、センタおよび一括型 EM 法において出

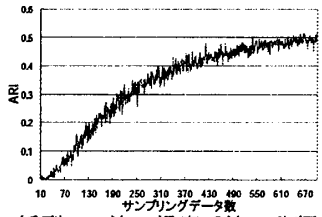

図 2: 一括型 EM 法、提案手法の分類結果類似度

	トピック1	トピック2	トピック3	トピック4	トピック5	トピック6	トピック7	トピック8
サイトA 2002年	NEC システム 分野	世界 業績 LSI	サービス BIGLOBE ブロードバンド	Express サーバ CPU				
サイトB 2003年	NEC ソリューション 分野	世界 次世代 カラー	サービス BIGLOBE インターネット	Express モデル CPU	システム アクセス VALUMO		ノート Mate PC	
サイトC 2004年	NEC 事業 中国	技術 設立 世界	BIGLOBE 配信 動画	Express Itanium モデル	効率化 業務 DiosaGlobe			UNIVERGE IP SV
センタ 提案手法 サンプリング数600	NEC 分野 中国	世界 技術 LSI	BIGLOBE サービス ブロードバンド	サーバ Express CPU	構築 業務 基幹	シリーズ ラインアップ VPN		UNIVERGE SV 低減
一括型EM法	NEC 分野 事業	世界 技術 開発	BIGLOBE サービス ブロードバンド	Express CPU サーバ	構築 業務 効率化	製品 NAS 運用性	ノート パソコン PC	

図 3:各サイト、センタ、一括型 EM 法のトピック特徴語

現確率の大きかった順にトピックの特徴語を[3]の手法で上位 3 位まで求めたものである。出現確率の低いトピック 6 以下はノイズの影響を受けているが、NEC、世界、BIGLOBE、Express の 4 トピックは、各サイト、センタのどちらにも出現しているのが分かる。これは、各サイトのトピックが適切に学習されたことを示している。また、トピック「構築」は各サイトではトピックと認識されていないが、一括型 EM 法と提案手法のセンタでは認識されている。これは全体構造を学習して初めて得られる知見であり、一括型 EM 法より通信量を削減した提案手法においても再現できたことは注目すべきである。

5. まとめ

複数のリモートサイトに蓄積されているテキストに対して、トピック全体構造を学習する手法の提案を行った。本手法は、既存手法と比較して通信量・プライバシーの面で優位であり、全データを一箇所に集めた場合と比べて同程度の精度を実現することを示した。また、個別サイトを分析するだけでは得られないような知見も本手法を用いることでトピック全体構造の中に得られることを示した。

参考文献

[1] C.Clifton, M.Kantarcioglu et al. Tools for Privacy Preserving Distributed Data Mining. *SIGKDD*, vol.4/2, 1-7, 2003.

[2] K.Yamanishi. Distributed Cooperative Baysian Learning Strategies. *Information and Computation*, vol.250, 22-56, 1999

[3] K.Yamanishi and H.Li. Mining Open Answers in Questionnaire Data. *IEEE Intelligent Systems*, 58-63, 2002

[4] L.Hubert and P.Arabie. Comparing partitions. *Journal of Classification*, vol.2, 193-198, 1985.

[5] S.Morinaga and K.Yamanishi. Tracking Dynamics of Topic Trends Using a Finite Mixture Model. *KDD*, 811-816, 2004.

[6] S.Morinaga, K.Yamanishi, and J.Takeuchi. Distributed Cooperative Mining for Information Consortia. *KDD*, 619-624, 2003

[7] 樺島祥介, 上田修功. 統計科学のフロンティア１１ 計算統計 I . 157-167, 2003