トピック相関による確率的潜在意味解析の拡張

An extended method of probabilistic latent semantic indexing by topic correlations

柴山 直樹 Naoki Shibayama
古田 一雄 Kazuo Furuta

*1 東京大学工学系研究科システム量子工学専攻
Department of Quantum Engineering and Systems Science, The University of Tokyo

Probabilistic Latent Semantic Indexing (pLSI) is a popular method for collaborative filtering which is based on simple matrix algorithm. This method has extendibility and scalability due to its simplicity of algorithm. Because of this simplicity, however, pLSI is not enough efficient for clustering or directory representation which are useful techniques of searching massive data. In this paper, we extends pLSI model to study the correlation between hidden topics. This expansion gives pLSI a power to analyze higher structure in topics.

1. はじめに

協調フィルタリングのベースとして用いられることが多い確率的潜在意味解析(probabilistic Latent Semantic Indexing : pLSI)は、単純な行列計算で実装できるスケールしやすいアルゴリズムである。その理論の単純さから拡張性が高く、研究が盛んだが、反面、学習される背景トピックには相関構造や階層構造といった高次構造を仮定しない。そのため、トピックのディレクトリ型表現やネットワーク型表現といった大規模データ検索に対して有効な方法には直接は対応しないため、これらを実装するには別の理論と組み合わせる必要がある。

そこで本研究では、pLSI モデルのトピック分布に相関を考慮し、pLSI の利点を保ったまま、トピック間の高次構造を学習するよう拡張した相関潜在意味解析(correlated Latent Semantic Indexing : cLSI)を提案する。cLSI では、確率的生成モデルに立脚した理論的に自然なネットワーク型検索や可視化を提供できると考えられる。

2. 提案手法概要

pLSI やその拡張は、離散確率変数の生成モデルをデータから学習する手法であり、適用できるデータの範囲は本来幅広い。ここでは説明をわかりやすくするため、文書内のカウントデータを例にしてモデルの説明を行う。

以後、*d*:文書、*w*:単語、*z*:トピックと表記する。また x の確率を p(x)と表記し、y による条件付き確率を p(x|y)と表記する。

2.1 モデル

pLSI は単語の確率的生成モデルとして理解することができるため、cLSI との比較は、生成モデルを用いると容易である。グラフィカルモデルで二つのモデルを比較したものを図1に示す。cLSI では、トピックに相関を導入するため、トピックを連続的に生成するモデルを仮定する。

2.2 学習

モデルの学習は、pLSI と同じく対数尤度(エントロピー)の最

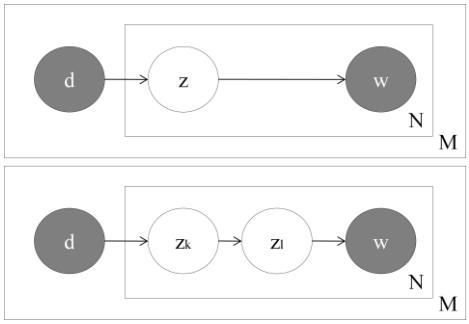

図 1 ：グラフィカルモデル （上）pLSI （下）cLSI

大化によって行う。cLSI の対数尤度関数は次のようになる。

$$L = \frac{1}{N}\sum_{d,w} n(d,w)\log\left(\sum_{k,l} p(d \mid z_k)\,p(w \mid z_l)\,p(z_k, z_l)\right)$$

この式は pLSI と同じく解析解が求まらないため、[Hofmann01]と同じ手順を用いて学習を行う。具体的には、対数尤度を、完全データ集合が得られたときの対数尤度で近似し、Expectation Maximization(EM)アルゴリズムを用いて最適化を行う。

ここで得られる最適化は、確率表現を用いた近似的なイタレーションとして導出される。このイタレーションを行列計算で表現したものをアルゴリズム1に示す。

行列計算に表現しなおすことは、過読性が上がると共に、並列化によるスケールアウトを促す意味で重要である。ここでパラメータは $D_{dk}=[p(d|z_k)]$、$W_{wl}=[p(w|z_l)]$、$Z_{kl}=[p(z_k|z_l)]$とし、周辺化分布は $Z_{kl}=[\sum_l p(z_k|z_l)]$とする。またデータは $X_{dw}=[n(d,w)/N]$とする。結合分布 $M_{dw}=[p(d,w)]$は次のように表現される。

$$M_{dw} = D_{dk} Z^{kl} W_{wl} \rightarrow X_{dw}$$

記法にアインシュタインの縮約規約を用いている。

2.3 計算量

アルゴリズム1は $O(\mathrm{size}(d)\cdot\mathrm{size}(w)\cdot\mathrm{size}(z))$の計算量である。ただし、データが疎行列の場合、データの充填率 p に左右される。全体の計算量は次のようである。

$O(\mathrm{size}(d)\cdot\mathrm{size}(w)\cdot\mathrm{size}(z)+\max(d,w)\cdot\mathrm{size}(z)^2)$

連絡先:柴山直樹, 東京大学工学系研究科 古田・管野研究室, 文京区本郷 7-3-1, shibayama@bs.t.u-tokyo.ac.jp

アルゴリズム 1：cLSI の EM アルゴリズム

$$E_{dw} = \frac{X_{dw}}{M_{dw}}$$

$$Z_{kl} = Z_{kl} D_{dk} E^{dw} W_{wl}$$

$$D_{dk} = D_{dk} \frac{E_{dw} W^{wl} Z_{kl}}{Z_k}$$

$$W_{wl} = W_{wl} \frac{E_{dw} D^{dk} Z_{kl}}{Z_l}$$

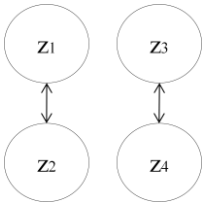

図 2 人工データの背景トピック相関

3. 評価実験

cLSI を評価するために二つの実験を行った。ひとつはトピックの相関構造の学習評価であり、もう一つは未知データへの予測性能評価である。人工データを作成し、これらを評価した。

3.1 人工データ

現実にあるデータでは、真のトピック分布やトピック間の相関を知ることができず評価が難しい。そのため、ここでは次のような人工データを作成し、入力データとして用いた。

図2のように4つのトピックを用意し、二つずつがクラスタを構成している場合を仮定した。同じトピック同士の相関は別トピック同士の相関の2倍とし、データの大きさは size(d)=500,size(w)=12 とした。$p(w|z)$は比較しやすいよう図4の上段のようにし、$p(d|z)$に関しては一様乱数を用いた。

3.2 背景構造の学習

図3に背景トピックの分布の学習結果を示す。真のトピック分布との KL ダイバージェンスは 0.075 であった。また、トピックを構成する単語生成確率である $p(w|z)$の学習結果を図4の下段に示す。トピック分布と単語生成確率を共にうまく学習していると言える。

3.3 未知データのモデル化

1-hold out の交差検定を行い、対数尤度と perplexity の比較を行った結果を表1に示す。pLSI に比べて、予測性能も変わらないことが示せた。

4. 課題と展望

本実験では、トピック分布の構造学習を評価するため、正しいトピック数を用いたが、自然データでは、それを事前に知ることはできない。また、トピックの独立性の仮定による自由度の制限がないため、pLSI でも問題であった過学習はより一層の問題となりうる。pLSI の研究ではアニーリングや階層ベイズ化などによる過学習の排除がなされているが、pLSI の特徴である行列計算を維持したままの過学習への対応は難しい問題となる。

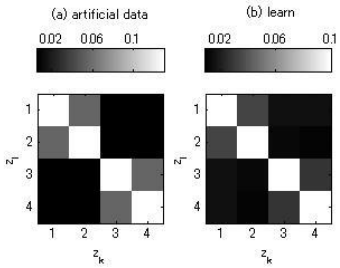

図 3 トピック相関の学習 （左）元データ （右）学習結果

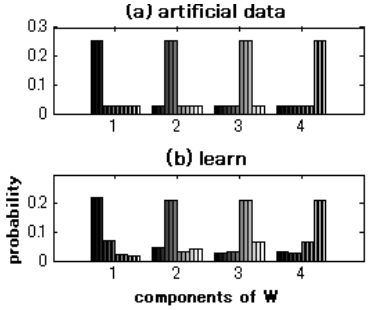

図 4 単語生成確率の学習 （上）元データ （下）学習結果

表 1 モデル化精度の評価

	log likelihood		perplexity	
pLSI	−2.2678	±0.0006	9.6579	±0.0001
cLSI	−2.2680	±0.0006	9.6671	±0.0001

これらを加味した過学習への対応と評価は今後の課題である。

また、相関構造以外のトピックの高次構造、たとえば、トピックの階層化を表現する確率的生成モデルと学習手法も cLSI と同じ導出方法を用いて導くことができる。それらについても実験・評価を行う必要がある。

5. まとめ

本研究では、pLSI を拡張し、トピック分布に相関を考慮した cLSI の提案を行った。

cLSI はトピック間相関を学習可能な確率的生成モデルの一種であり、pLSI では理論的な整合性を保ったまま実装することが難しい、トピックのネットワーク表現に対応することが可能となる。また、学習は EM アルゴリズムを用いて最適化され、pLSI と同じく行列計算を用いて表現される。

実験では、人工データを用いて、トピック間相関の学習と、予測性能について評価した。

最後に、課題と展望では過学習の問題とその対応について、議論した。

参考文献

[Hofmann 2001] T.Hofmann: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, 42(1):177-196, 2001.

[Blei 2003] D.M.Blei: Latent Dirichlet allocation, *The journal of Machine Learning Research*, 3:993-1022 , 2003.

[Blei 2006] D.M.Blei,J Laffety: Correlated topic models , *Advances in neural information processing systems*, 18:147, 2006.