潜在トピックモデル　LDA　ユーザプロファイリング

Collaboration Projects

潜在トピックモデルを利用したユーザプロファイリング技術

webアクセスログを利用したユーザのプロファイリングを行うことを目的に，トピックモデルを応用したweb閲覧行動のモデル化技術を開発した．本技術は，事前にユーザのwebアクセスを，ユーザの意図を最も良く反映したURLだけを抽出することで，高い精度でのユーザプロファイルのモデル化を実現する．なお，本研究は，大阪大学サイバーメディアセンター内に設置した共同研究部門との共同研究により実施した．

サービス＆ソリューション開発部　藤本（ふじもと）　拓（ひろし）　秋永（あきなが）　和計（よしかず）
栄藤（えとう）　稔（みのる）
マーケティング部　金野（きんの）　晃（あきら）

1. まえがき

web閲覧ログの解析によるユーザのプロファイリングは，ターゲット広告，コンテンツ推薦といったwebアプリケーションの高度化，パーソナライズに有用な手段の1つである．

本研究では，文書分類分野で広く使われる，潜在トピックモデル[*1]を利用したwebユーザのプロファイリング方式の確立を目的とする．具体的には，広範なweb閲覧行動が記録されるプロキシログ[*2]を解析し，ユーザのweb閲覧行動をモデル化することで，ユーザプロファイルを生成する．

本研究では，文書解析に使われる潜在トピックモデルを，プロキシログの解析へ適用することで，ユーザプロファイリングを行う．このとき，より良いプロファイル結果を得るためには，潜在トピックモデルに投入するURLの集合が，ユーザの意図を良く反映したものでなくてはならない．そのため，大量のwebページへのアクセスを含むプロキシログから，どのようにユーザの意図を最も良く反映したURL系列だけを抽出するかが，本研究の主題となる．

文書分類分野では，大量の文書から，文書の意味を最も良く反映した単語の集合を抽出する課題に対して，単語の属性抽出による抽象化を行う辞書構成が有効であると言われてきた[1]．プロキシログの解析においても同様のアプローチが有効であると考えられるが，今までにこれを考慮した研究は行われていない．本研究では，与えられたログから，URLセッションごとに，ユーザの意図を最も反映した単語セットを生成する手法である，Cross-Hierarchical Directory Matching（CHDM）を提案する．CHDMは，同一URLセッション中で最も意味的な抽象度の高いURLへのアクセスが，最もユーザの意図を反映すると仮定する．この仮定に基づき，CHDMは，階層型URL辞書を利用することで，web閲覧ログから抽象度の高いURLのみを抽出する．辞書は，広範なweb空間のURLをカバーし，辞書に登録されたすべてのURLに対して意味的な階層関係を与えるものである．このような辞書は，例えばYahoo! JAPAN Directory[*3]などの，ディレクトリ型検索エンジンより生成可能である．

本技術の実現にあたり，さまざまな意図や趣味・嗜好に基づいた大量のwebアクセスログを用意する必要があり，そのデータの収集に大阪大

＊1　**潜在トピックモデル**：文書分類分野で広く利用される技術であり，文書が，複数の単語の分布で表される潜在トピックによって，生成されることを表現したモデル．

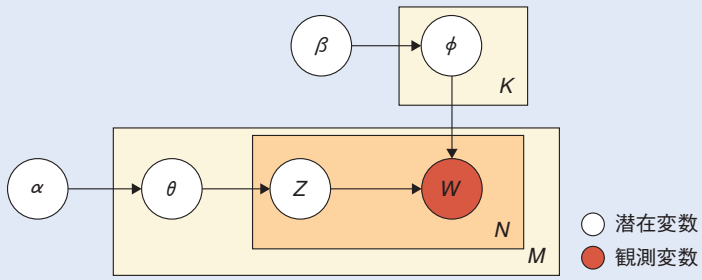
学の計算機環境を利用するため，大阪大学との共同研究とした．

2. LDAによる定式化

ユーザのプロキシログから，ユーザのweb閲覧行動をモデル化するため，本研究では潜在トピックモデルの1つであるLDA（Latent Dirichlet Allocation）[*4][2]を利用する．

プロキシログの解析にLDAを当てはめる場合，ユーザのweb閲覧行動に潜在トピック[*5]が存在すると仮定する．例えば，ユーザは「プログラミング」という潜在トピックの基で「C言語」というトピックを持つ学習サイトへアクセスする．このような仮定を置いた場合，文書をユーザ，単語をURLと置き換えることで，LDAを適用することが可能となる．文書解析とプロキシログ解析の比較を**図1**に示す．ユーザの各URLの閲覧頻度（以下，N）をLDAへ入力することで，各ユーザは潜在トピックの確率分布（以下，θ）で表現され，各潜在トピックは，URLの確率分布（以下，ϕ）で表現される．

本研究の目標は，ユーザのweb閲覧行動を最適にモデル化したθ，ϕを導出することである．したがって，高い精度でのモデル化には，プロキシログからLDAへの入力に最適なURLの集合（W）を生成することが重要になる．



	一般的な文書解析の場合	プロキシログ解析をLDAに当てはめた場合
α	θの事前分布を生成するためのパラメータ	同左
β	ϕの事前分布を生成するためのパラメータ	同左
θ	潜在トピックの確率分布	同左
ϕ	単語の確率分布	URLの確率分布
M	文書数	ユーザ数
N	1文書当りの単語の出現回数	ユーザの各URLの閲覧頻度
K	トピック数	同左
Z	単語の潜在トピック	URLの潜在トピック
W	単語の集合	URLの集合

※観測されたWを用いることで他の潜在変数の推定が可能

図1　LDAモデルにおける文書解析とプロキシログ解析の比較

3. 提案方式

3.1 単語セットの生成

本研究では，**図2**に示すようなプロキシログを想定する．プロキシログの各レコードは少なくとも，ユーザID，コンテンツ閲覧時刻，閲覧コンテンツのURLを含み，時刻順にソートされて記録される．さらに，各ユーザの一連のレコードの集合として，URLセッション（以下，「セッション」と呼ぶ）に分割され，セッションのID（セッション識別ID）も記録される．各セッションは，特定のタイムアウトによって分割される．図2では1～3の3つのセッションに分割されている．

文書分類における従来研究に従えば，単語の生成には，単語の属性抽出による抽象化が有効である[1]．本研究では，セッションからのURL生成においても同様の効果が得られると仮定し，セッションの中で最も上位概念にあたるURLの集合を抽出し，これを当該セッションから生成される単語セットとする．以下では，セッションごとに，LDAへの入力となる単語セットを抽出する手法を述べる．

図2には，さらにセッションと抽出される単語セットの関係が示されている．図は，3つのセッション1，2，3から構成される，あるユーザu1のログと，そこから得られる単語セットを示している．例えば，セッション1では，v1とv2へのアクセスが記録されるが，v1のみが抽出される．また，セッション2では，v3，v4，v5が抽出され，セッション3では，v1，v3が抽出される．各セッションから，実際にどのようにして最適な単語セットを抽出するかは，次節で述べる．これら3つの抽出結果

＊2　**プロキシログ**：プロキシサーバを介してwebページにアクセスした際に，プロキシサーバに蓄積されるアクセスログ．

＊3　**Yahoo! JAPAN Directory**：ディレクトリ型のwebページ検索エンジンの1つ．webページに最大18階層からなるカテゴリを付与することでwebページを分類し，カテゴリよりwebページの検索を可能としている．Yahoo!は，Yahoo! Inc.の商標または登録商標．

＊4　**LDA**：潜在トピックモデルの一形態であり，単語ごとに確率的にトピックが決定され，さらに文書ごとに確率的にトピックが決定されるモデル．

を合わせることで，最終的にユーザが各単語を何回アクセスしたかを導出する．

3.2 CHDM

セッションから，上位概念に抽象化したURLの集合を生成するため，本研究では，Yahoo! JAPAN Directoryを利用し，階層型URL辞書（以下，辞書）を生成した．辞書は，階層化されたカテゴリにより構成され，上位の階層ほど抽象度の高い概念を有し，また各カテゴリには複数のURLが登録されている．例えば，スポーツニュースカテゴリの下には，ワールドカップカテゴリが存在し，各カテゴリには対応するURLが登録される．そのため，登録されたURL間の意味的な階層関係を知ることが可能である．本研究では，上記述べたような辞書を利用したセッションの抽象化処理を，CHDMと呼ぶ．

CHDMの基本的な動作は，2つのステップに分類される．まず，セッションに含まれるURLから，辞書に登録されている単語セットを抽出する（マッチングステップ）．

次に，当該セッションにおいて最上位概念にあたる単語セットを抽出する．これは，辞書を利用することで，意味的な階層関係がある単語セットを発見し，各集合について，最上位概念のURLを抽出していくことで実現する（抽象化ステップ）．以上のCHDMの動作および辞書の定義は，文献[3]にてさらに詳しく述べている．

CHDMの動作例を，**図3**を基に説明する．図は，プロキシログに記録された，あるユーザのセッションを左に示し，辞書を右に示している．セッションには6つのアクセスが含まれ，辞書にはc1からc5の5つのカテゴリとv1，v3，v4，v5のURLが登録されている．

まずマッチングステップにおいて，時刻t1，t3，t4，t5，t6にアクセスされたURLから辞書にマッチしたURLを抽出する（図3①）．

次に，抽象化ステップでは，抽出されたURLから最上位概念に当たる単語セットを抽出する．まず，抽出された各URLに対応するカテゴリを辞書より抽出する（図3②）．得られたカテゴリについて，c3とc4，c3とc5は，それぞれ意味的な階層関係にある．そこで，これらについては，最上位概念にあたるc3に対応するURLのみを抽出する（図3③）．したがって，当該セッションから最終的に得られる単語セットは，カテゴリc2，カテゴリc3に対

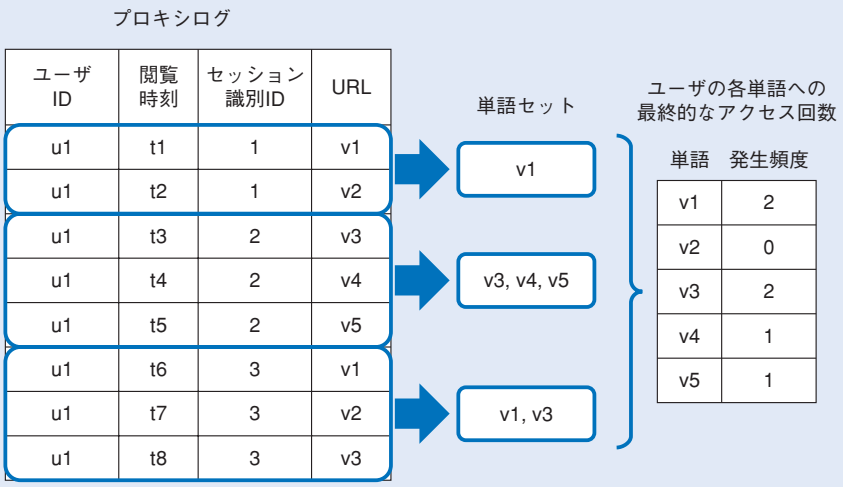


図2　セッションと単語の関係例

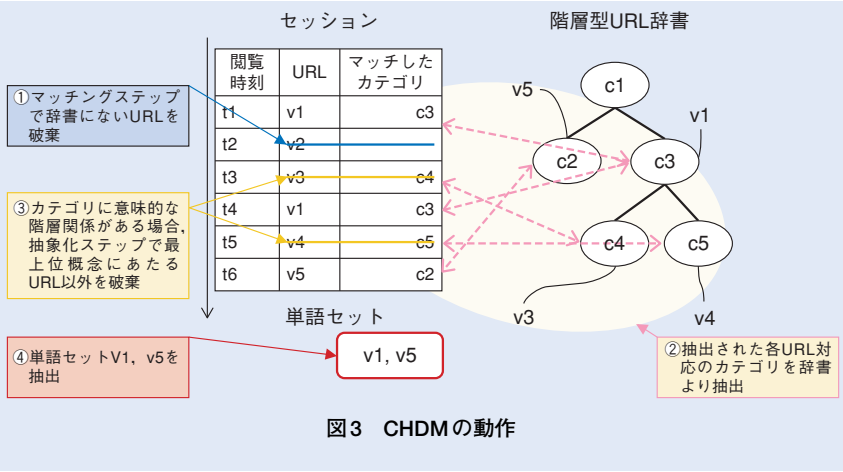


図3　CHDMの動作

＊5 **潜在トピック**：潜在トピックモデルにおいて，存在が仮定される潜在変数であり，単語の頻度分布により表現される．潜在トピックモデルでは，文書は，この潜在トピックの確率分布により表現される．

応するURL（v1，v5）となる（図3④）．

すべてのユーザについて，すべてのセッションから単語セットを抽出したあとは，その和集合として，LDAへ与える最終的なURLの集合（W）が与えられる．そして，各ユーザについて，各URLを閲覧したセッション回数の合計値として，各URLの閲覧頻度，つまりLDAへの入力であるNを導出可能である．

4. 性能評価

4.1 データセット

CHDMによって得られたモデルの精度を評価するため，大阪大学の学生7,537人のweb閲覧を記録したプロキシログを，2010年4月から7月の4カ月に渡って収集した．ログのサイズは40GB，レコード数は約1億3千万レコードである．またセッションを分割するタイムアウトは1,800secと設定した．これにより，合計で175,831のセッションを得た．また，2010年7月にYahoo! JAPAN Directory を巡回することで，57万のURLが登録された辞書を生成した．これらの辞書に登録されたURLのうち，プロキシログ上でweb閲覧ユーザ数が5以上のURLを4,550抽出した．辞書と上記175,831のセッションとの間でCHDMを適用した結果，全セッションの80％以上から単語セットが抽出された．

4.2 評価結果

CHDMで得られたモデルの精度を，以下2つの方式で得られた精度と比較する．まず，非抽象化方式で，これはCHDMを一切適用せずに生成したモデルである．次に，ディレクトリマッチ方式で，これはCHDMのマッチングステップのみを適用して生成したモデルであり，抽象化ステップの性能評価に利用する．また，評価指標にはパープレキシティ[*6]を用い，前半3カ月のログを用いて生成したモデルを，後半1カ月のログと比較することで，モデルの精度を評価した．

結果を**図4**に示す．図は，各方式について，潜在トピック数ごとのパープレキシティの変化を表す．CHDMは，他のすべての方式と比較して，良い性能を示している．特に，ディレクトリマッチ方式からもさらに10％程度性能が向上しており，抽象化ステップによる効果も大きいことが分かる．

ただし，CHDMは，辞書による抽象化により良いモデルが得られるという，発見的な仮定に基づいている．性能評価により，仮定が正しいことを示したが，別の辞書を利用した場合に，同様の結果が得られるという保証は無い．

5. ユーザプロファイルの可視化

CHDMにより得られたモデルを利用したユーザプロファイリング結果を示し，モデルの有効性を主観的に評価する．データは，前章で利用したものと同じデータを用い，潜在トピック数を24に設定した場合に，LDAから得られたモデルを用いた．

得られたモデルの有効性を確認するため，潜在トピックと，学生がもつ属性（専攻，学年）との関係について調べた．具体的には，24の潜在トピックを，「専攻（理系/文系）」と「学年（高学年/低学年）」の2つの軸に非線形射影し，これを可視化した．射影の詳細な方法は，文献[3]で述べている．

射影結果を二次元グラフ上にマップした結果を，**図5**に示す．各点は，ユーザを表し，x軸の正方向に配置

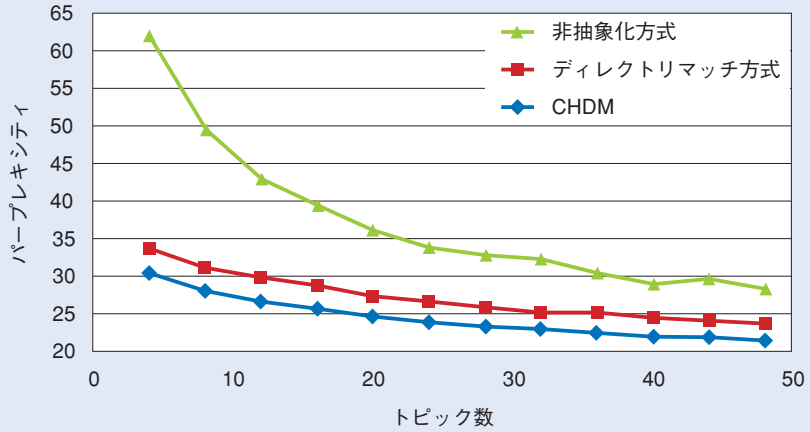
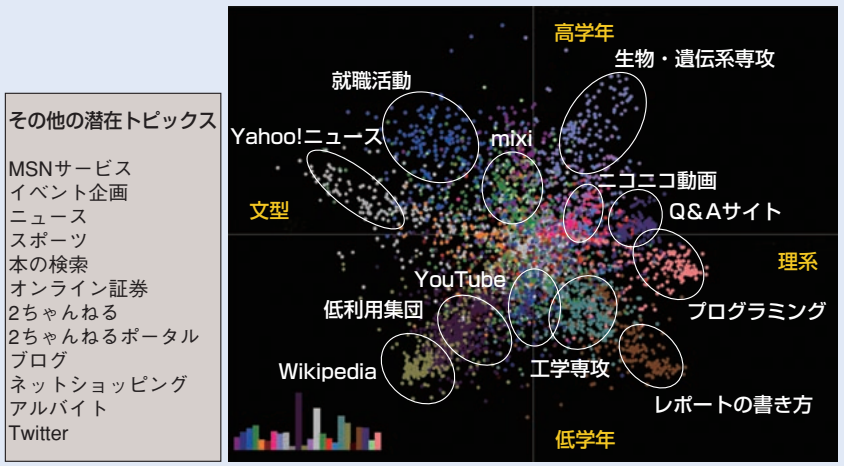


図4　パープレキシティ評価

＊6　**パープレキシティ**：2つの分布間の距離を測る尺度．ここでは，その2つがモデルと実測値であり，得られたモデルが実測値にどの程度当てはまるかを評価する指標として用いた．値が小さいほど，良いモデルとなる．

されるほど，当該ユーザの潜在トピックが理系的であり，負方向に配置されるほど文系的であり，y軸の正方向に配置されるほど高学年傾向であり，負方向に配置されるほど低学年傾向となる．また，24の潜在トピック数それぞれに対応する色を用意し，各ユーザの最も支配的な潜在トピックにより，各ユーザを24色のいずれかで色付けしている．また，24の潜在トピックに名前を付け，特にグラフの特定箇所に集まっているものについては，図中に潜在トピック名を記述し，それ以外のものは図左に記述した．さらに，図左下には，各潜在トピックに属する学生数を表す棒グラフを示した．



ニコニコ動画®：㈱ドワンゴの登録商標．
mixi™：㈱ミクシィの登録商標．
MSN®：米国Microsoft Corporationの米国およびその他の国における商標または登録商標．
Twitter：アメリカ合衆国または他の国々におけるTwitter,Inc.の登録商標．
YouTube™：米国Google Inc.の商標または登録商標．

図5　ユーザプロファイルの可視化

図では，同じ色の点が同じ場所に集まる傾向が見られる．これは，潜在トピックと属性値に強い相関関係があることを示している．特に，「就職活動」「生物・遺伝系専攻」「プログラミング」などの潜在トピックに，その傾向が強く見られる．したがって，生成したモデルから得られたプロファイリング結果は，学生の属性を良く反映しており，定性的に良いモデルが得られているといえる．

6. あとがき

本研究は，広範なユーザのweb閲覧行動をモデル化するため，URLセッションからの単語生成手法を提案した．提案方式では，階層型URL辞書を利用したURL系列の抽象化を行うことにより，高精度なモデルを得られる単語セットの抽出を可能とした．また，7,537人のユーザのプロキシログに対して提案方式を適用し，単語の予測精度の評価を行うことで，提案方式の有効性を示した．さらに，モデルから得られたプロファイリング結果を可視化し，主観的にモデルの有効性を示した．

提案方式は，使用する階層型URL辞書により性能が異なる．今後は，構成する辞書の違いによる性能比較を行い，より精度の良い辞書の構成手法について検討する．

文　献

[1] Z. Elberrichi, A. Rahmoun and M. A. Bentaalah：“Using WordNet for Text Categorization,” The International Arab Journal of Information Technology, Vol. 5, No.1, Jan. 2008.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan：“Latent dirichlet allocation,” The Journal of Machine Learning Research archive Vol.3, pp.993-1022, 2003.

[3] H. Fujimoto, M. Etoh, A. Kinno and Y. Akinaga：“Topic Analysis of Web User Behavior Using LDA Model on Proxy Logs,” ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, LNCS Vol.6634/2011, pp.525-536, 2011.