1F4-4

トピック情報を利用したユーザの知識推定

Capturing users' domain knowledge based on topic model

片山太一 Taichi Katayama　小林のぞみ Nozomi Kobayashi　牧野俊朗 Toshiro Makino　松尾義博 Yoshihiro Matsuo

日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation

For personalization, it is important to capture various information about users. We assume that one of the key information is users' knowledge. In previous studies, the difficulties of general terms have already been focused on. In this study, we focus on technical terms. Users have different background knowledge about different topics. Therefore, we assume that the difficulty of a technical term should be determined depending on each user's knowledge level and the domain which the word belongs to. Therefore, we propose a method to capture users' domain knowledge regarding the difficulty of a word which changes depending on a user and a target domain. We use topic adaptation for Latent Dirichlet Allocation(LDA). Experimental result shows that our proposed method is effective especially in estimating the knowledge of the users who have a lot of knowledge.

1. はじめに

近年，blog や wiki，SNS，Social Bookmark などの様々なソーシャルサービスによって Web の質が変化し，これまで情報資源に対して閲覧者であった利用者が作成者へと変わり，これにより Web における情報資源は日々激増することになった．このような情報爆発時代では利用者が求めている情報を素早く探すことが重要となっているが，日々増え続ける情報資源の中から探すのは困難であり，利用者は検索窓を前にして途方に暮れることも多くなっている [9].

そこで近年，全てのユーザに対し同じ結果を提示するのではなく， ユーザそれぞれの背景を考慮にいれて情報を選択・提示することで，ユーザが欲しい情報を容易に見つけられるようにすることが重要になっている．[4] によると，これらの個人に適した情報を選択し，個人に適した方法で表示するサービスをパーソナライゼーションといい，パーソナライゼーションを行うために必要となるユーザに関するプロファイル情報を抽出する技術をユーザプロファイリング技術という．

いままでのユーザプロファイリング技術に関する研究は，ユーザの興味関心 [7] やユーザの属性 [8] を抽出するものがほとんどであった．しかし，検索システムなどの情報提示のサービスでは，興味関心や属性だけでなく背景知識も重要な項目の一つであると考えられる．

例えば，検索システムを利用して情報を探す場合を考える．ユーザの検索クエリに対する知識を推定できれば，検索クエリに対して専門的な知識を有するユーザに対しては，専門的な用語で書かれている情報を，専門的な知識を持たないユーザに対しては，平易な用語で書かれている情報を提示することが可能になる．また，ユーザの知らない知識に対して注釈を出すことが可能になり，ユーザが欲しい情報を容易に得られることができる．

上記のような情報提示サービスを実現するためには，ユーザの知識を知り，ユーザの知識に見合った情報を提示する仕組みが必要であると考えられる．まず，ユーザの特定分野の用語に関する知識を知ることに着目した．用語の知識を知ることは，用語の知る知らないを推定する問題からどの程度詳しいのかまで推定する問題まで様々である．そこで，本稿では用語の知識を知ることの基本である用語を知っているか否かを推定することを行う．従って，特定分野における専門用語を意味を知っている用語とそうでない用語に分けることを本研究の目的とする．

以下，2. では関連研究とその課題について述べ，3. では 2. で述べた課題を解決する手法について説明する．そして，4. で実験の詳細について述べ，5. で評価の結果と考察，6. で結論と今後の課題について説明する．

連絡先: 片山太一, 日本電信電話株式会社 NTT メディアインテリジェンス研究所, 神奈川県横須賀市光の丘 1-1, 239-0847, katayama.taichi@lab.ntt.co.jp

2. 関連研究と課題

ユーザが用語を知っているか否かの推定に関する研究は，一般的な用語を対象としたものがほとんどであった．たとえば [1] は，一般的な英語の用語に対して知っているかどうかの予測をしている．[1] では，英語の基本語彙 12,000 語に対し，英語母国話者を含むチームが人手で 12 段階の難易度を付与した語彙リスト [11] を利用し，ある英用語を知らないとその英用語よりも難しいものはすべて知らないという考えのもと，ユーザの知らない用語を入力とし，項目応答理論 [6] *1を用いて，知らない用語を予測する手法を提案している．[2] は，一般的な日本語の用語に対して，人手で 7 段階の難易度を付与した語彙リスト [12] を利用し，[1] と同様の考えのもと，ユーザの知っている用語と知らない用語を入力とし，難易度の境界を決定することで知識を推定している．

これらの関連研究と本研究で大きく異なるところは，専門用語を扱うことである．本稿では，専門用語と一般的な用語（一般用語）を以下のように定義する．専門用語とは，ある業種や分野で専門的に使われている用語である．一般用語とは，一般的に使用されている用語である．専門用語には，関連研究のように用語の難易度が付与されているリストがなく，人手で作成するにはコストが膨大にかかるため，人手で作成された難易度付きの辞書を仮定することは難しい．

もし，関連研究のように難易度付きの辞書を仮定できたと

*1 日本語では，項目反応理論，テスト理論などとも呼ばれる．

しても，専門用語には単純に解けない課題がある．それは，一般的な難易度とユーザにとっての難易度は異なるという点である．一般用語の多くは一般教養として学ぶことが多く，ユーザによって学ぶ順番が大きく異なることはない．しかし専門用語は，分野ごとに様々なジャンルで細分化されており，ユーザによって学ぶジャンルの違いや順番が異なる．従って，専門用語では，あるジャンルについては詳しいが，あるジャンルについては詳しくないといったユーザもいるため，知識の量が同じくらいでも，同じような知識を持っているとは限らない．例えば，スポーツの分野について考える．野球には詳しいが，サッカーには疎いユーザにとっては，野球の用語よりもサッカーの用語の方が比較的難しいと感じる．しかし，サッカーには詳しいが，野球に疎いユーザには，サッカーの用語の方が簡単で，野球の用語の方が難しいと感じる．このように，難易度付きの辞書を仮定できたとしても，ユーザごとに感じる難易度は異なるため，[1] や [2] の手法では専門用語の知識を推定することは困難である．

3. 提案手法

本稿では，[2] のように少数のユーザの入力からユーザの知っている用語と知らない用語の境界を決定することで専門用語の知識を推定する手法をとる．ユーザの入力は，他の手法として書いた文章や様々なログを入力とするアプローチも考えられる．しかし，ユーザが知らない用語はテキスト中に出現しないため，知らない用語を推定する別の処理を加える必要がある．そのため，まずはユーザに少数のデータを入力してもらうアプローチを取る．

2. で述べた課題である一般的な難易度とユーザにとっての難易度が異なる点については，一般的な難易度順に並べられた専門用語をユーザごとにジャンルの詳しさを反映させた難易度順に並び替えることで解決する．手法としては，ユーザのジャンルをトピックと考え，LDA のトピック適応を用いて並び替える．

知識推定の処理全体について説明する．まず処理の事前準備として，専門分野の一般的な難易度を付与したリストを用意し，各トピックにおける用語の生起確率を求める．[3] によれば，用語の頻度と親密度には高い相関関係があることが報告されている．この用語の親密度はどのくらいその用語になじみがあるかという尺度であり，ほとんど難易度と等価であると考えられる．そこで，[3] の考えに基づき，用語の生起確率を難易度とする．次に，用意したリストの中から少数の専門用語を取り出し，ユーザに提示し，知っているかどうかの入力をしてもらう．入力された情報をもとに，用語の生起確率を計算し直し，その値を用いて専門用語を並び替え，知っている用語と知らない用語の境界を求める．境界よりも，簡単であると推定した結果を知っている，そうでない用語を知らない用語として出力する．

以下，3.1 では，3.2 で並び替えるための事前準備として，LDA による各トピックにおける用語の生起確率について述べ，3.2 では，トピック適応を用いた並び替えについて説明し，3.3 では，境界の求め方について説明する．

3.1 各トピックにおける用語の生起確率

ユーザのトピックの詳しさを反映させるために LDA のトピック適応を用いる．そのための事前準備として，LDA を用いて各トピックにおける用語の生起確率を求める．LDA [5] とは，文書のような離散データ集合の生成モデルである．

LDA を利用して，トピック $z(1 \leq k \leq K)$ における用語 w の生起確率を求める方法を説明する．モデルの学習として，Gibbs サンプリングに基づく学習を行うと，パラメータの更新式として式 1 が与えられる．

$$p(z_{mn} = k|z^{-mn}, w) \propto (n_{mk}^{-mn} + \alpha)\frac{n_{wk}^{-mn} + \beta}{n_{k}^{-mn} + V\beta} \quad (1)$$

z_{mn} は文書 m 内で n 番目の用語のトピック，n_{mk} は文書 m 内でトピック k である用語数，n_k はトピック k である全用語数，n_{wk} はトピック k である用語 w の数，V は全文書中に出現する語彙数を示している．α，β はハイパーパラメータである．

式 1 に基づき生成モデルを学習させ，学習させた値を 2 式に入力することで，トピック z における用語 w の生起確率が求まる．

$$p(w|z = k) = \frac{n_{wk} + \beta}{n_k + V\beta} \quad (2)$$

3.2 トピック適応による並び替え

トピック適応とは，トピック別に学習された複数のモデルを持ち，ユーザの入力に対して，最適な混合比を求めて混合する方法である．トピック適応を用いることで，詳しいトピックの混合比が大きくなり，詳しくないトピックの混合比が小さくなるため，ユーザの詳しさを反映させた難易度順に並び替えることができる．

一般的なトピック適応は文脈や文章が入力であるのに対して，本研究では，知っている用語のリストが入力である．そのため，学習したモデルは頻度を考慮しているが，入力は知っている・知らないの 2 値になるため，頻度を考慮することができない．そこで，学習モデルと合わせるために，学習データの頻度を用いて考慮する．具体的には，ユーザの知っている用語が学習データの頻度の数だけ出現する文書を仮想的に作り出し，これを入力としてトピック適応を行う．

トピック適応に必要な混合比率の学習にも Gibbs サンプリングに基づく学習を行うとパラメータの更新式として式 3 が得られる．

$$p(z_{hn} = k|z^{-hn}, w) \propto (n_{hk}^{-hn} + \alpha)p(w|z = k) \quad (3)$$

z_{hn} は仮想文書 h の n 番目の用語のトピック，n_{hk} は仮想文書 h 内でトピック k である用語数を示す．$p(w|z = k)$ は式 2 で算出したもの，α は式 1 で使用した値を使う．

式 3 に基づきモデルを適応させ，適応させた値を式 4 に入力することで，仮想文書 h のもとでのトピック z における混合比 γ_k が求まる．

$$\gamma_k = \frac{n_{hk} + \alpha}{\sum_{k=1}^{K}(n_{hk} + \alpha)} \quad (4)$$

混合比 γ_k と学習済みの生起確率 $p(w|z = k)$ を掛け合わせることで，仮想文書 h のもとでの用語 w の生起確率を求める．

$$p(w|h) = \sum_{k=1}^{K} \gamma_k p(w|z = k) \quad (5)$$

式 5 で算出された生起確率をトピックを反映したユーザごとの難易度とし，この値を用いて並び替える．

表 1: 今回使用した用語の例

rss	cgi	ブログ	キャッシュ	JSON	検索エンジン	マッシュアップ
FTP	OSI 参照モデル	IP アドレス	ADSL	PPPoE	アクセスポイント	パブリックドメイン
DBMS	ソースコード	エンコード	インストール	パーティション	公開鍵暗号	マークアップ言語
主記憶装置	仮想メモリ	ハブ	ハードウェア	MBR	磁気ディスク	マザーボード

表 2: 被験者内訳

知っている用語の数	男性 (人)	女性 (人)	合計 (人)
100 個以下	3	3	6
101〜200 個	4	5	9
201 個以上	3	2	5
合計	10	10	20

3.3 境界推定

本手法では，ユーザが知っている用語と知らない用語の境界推定の手法として [10] の手法をとる．具体的には，知っている用語と知らない用語の難易度の境界付近では，知っているか否かに多少のばらつきがあるという考えのもと，簡単な順に並べた時に知らない用語が二つ以上連続する用語の難易度と，難しい順に並べた時に知っている用語が二つ以上連続する用語の難易度との中間点を境界とする．

少数のユーザの入力から，上記の手法で境界を求め，この境界を用いてユーザの知識を推定する．

4. 評価実験詳細

本評価実験において，LDA によるトピック適応を用いて用語の難易度を並び替えることが，ユーザの知識推定において有効であることを示す．

4.1 対象分野と専門用語のリスト

本実験では，対象分野をコンピュータとし，専門用語を抽出した．コンピュータ分野を選んだ理由は，初心者から熟練者まであらゆる人が興味を持ち，かつユーザの知識の差が存在すると思われるためである．コンピュータ系の web テキストの中から，web 系・通信系・ソフトウェア系・ハードウェア系の用語を中心に約 400 語を人手で抽出した．表 1 に抽出した用語の一例を示す．

4.2 評価データ

評価用データとして， 4.1 で用意した専門用語に対して，20 名の被験者により，専門用語を知っているかどうかのラベルを付与した．被験者 20 名の内訳を表 2 に示す．被験者は，表 2 から見ても分かる通り，男女それぞれ 10 人ずつ用意し，知っている語の数もなるべく偏りがないようにした．なお，被験者には用語が知っているか知らないかを判断する際の基準として，下記を提示した．

> 明確ではなくても，その用語について一つでも意味を定義できる用語は知っている用語である

4.3 LDA の学習データと学習条件

LDA によるトピック推定用学習データとして， 4.1 で用意した専門用語が 2 種類以上出現するブログを約 4 万記事用意した．学習条件として，潜在トピック数 $K = 20$，ハイパーパラメータ $\alpha = 0.5$，$\beta = 0.5$，反復数 1000 回として LDA によるトピック推定のモデル学習を行った．

また混合比の学習条件は，反復数を 100 回とし，それ以外はトピック推定と同様の学習条件で学習を行った．混合比の学習に用いるユーザの入力語は， 4.1 で用意した専門用語のリストからランダムに 10 語選んだ．

*2

4.4 比較手法

[2] のように語の一般的な難易度を利用することで，用語の知っている・知らないを推定する手法を比較手法とする．ここでも，[3] の考えに基づき，用語の出現確率を用いて一般的な難易度を付与する．具体的には，難易度は数字が大きいほど簡単であると定義し，大規模テキストから算出される出現確率をそのまま難易度とした．難易度 $d(w)$ の算出式を式 6 に示す．

$$d(w) = \frac{n(w)}{\sum_w n(w)} \tag{6}$$

$n(w)$ は用語 w の大規模テキスト中の頻度を示している．難易度 $d(w)$ の算出で扱う大規模テキストは， 4.2 の LDA のトピック推定用学習データとして用意した約 4 万記事のブログを用いる．

この難易度を用いて，3.3 で述べた手法で境界を求め，知っている用語と知らない用語を推定した．

4.5 評価尺度

本実験は，正解率の指標で評価する．ユーザの知っている用語の集合を T，ユーザの知らない用語の集合を F，システムが知っていると推定した用語の集合を P，システムが知らないと推定した用語の集合を N とすると，正解率は以下の式で算出される．

$$\text{正解率} = \frac{|P \cap T| + |N \cap F|}{|P| + |N|} \tag{7}$$

5. 評価結果と考察

知っている用語の数でユーザを並び替え，それぞれの正解率をプロットした結果を図 1 に示す．図 1 の結果から比較手法はユーザの知っている用語の数により精度のばらつきが大きいものの，提案手法は精度のばらつきが小さくなっていることが分かる．これは，どのようなユーザでも提案手法の方が精度を安定して推定できることを示している．

図 1 の結果から多くの知識を有するユーザに対して，提案手法が大きく差をつけていることが分かる．多くの語を知っているユーザの大半は， 2. で述べたように，あるジャンルには詳しいが，あるジャンルには詳しくない．このようなユーザに対して，提案手法であるトピック適応によって，ユーザがどの

*2 潜在トピック数，ユーザの入力語数については，予備実験の結果最も良かった数を，そのほかの数値はデフォルト値を用いた．

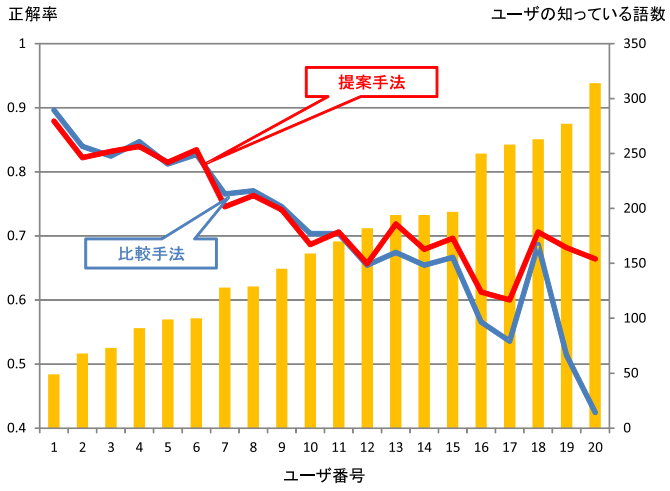

図 1: それぞれのユーザの推定結果

トピックに詳しいのかを捉え，難易度順にうまく並び替えることができ，比較手法から大きく精度向上ができたと考えられる．最も知っている語が多いユーザでは，約 24%ほどの精度向上を確認することができた．

一方，知っている用語の数が 100 語以下のユーザは，比較手法とほとんど差がなかった．これらのユーザが知っている用語の多くは，”ブログ”や”PC”などのコンピュータ分野の基礎的な用語であり，一般的な用語になりつつある専門用語であるため，ユーザによる難易度の違いが少ない．そのため，トピック適応を行って並び替えたとしても，並び替える前と大きく変化する用語があまりなかったことが原因と考えられる．

6. 結論と今後の課題

本稿では，ユーザの専門用語に対する知識を推定する技術について報告した．知識を推定することを用語を知っているかどうかを推定することと定義し，この問題に取り組んだ．先行研究で扱われてきた一般用語とは異なり，専門用語ではユーザごとに詳しいジャンルが変わることを課題として挙げ，LDA のトピック適応を用いて解く手法を提案した．20 名の被験者に対して，約 400 語の専門用語の知識を提案手法と比較手法によってそれぞれ推定を行った．評価実験の結果，提案手法により，多くの知識を有するユーザに対して，最大 24%もの正解率の向上が見られた．これにより，提案手法は多くの知識を有するユーザに対して有効であることが確認された．

以下に今後の課題を述べる．本研究では，専門用語をコンピュータ分野の用語約 400 語に限定して評価を行った．そのため，今回検討した手法が，医療・法律などの他分野においても適用可能であるか検証する必要がある．また，ユーザの入力の用語の選択方法等もランダムに固定して，評価実験を行った．そのため，ユーザの様々な入力に対して，頑健なシステムであるか検討する必要もある．

これらの課題に取り組んだ後，1. に述べたような検索システムを実現するために，ユーザの知っている用語と知らない用語を用いて，ユーザが理解できる情報を提示する技術について取り組んでいく．

参考文献

[1] 江原遥，二宮崇，清水伸幸，中川裕志,”Web ページ中のユーザが知らない語を予測する読解支援システム”, 人工知能学会全国大会予稿集，2F1-4，June 2010

[2] 天野成昭, 頭の中に用語はいくつある？, 情報処理 37(4), 351-354, 1996-04-15,

[3] 近藤公久，天野成昭,「日本語の語彙特性」データベース：有効性と問題点, 電子情報通信学会技術研究報告. TL, 思考と言語 100(335), 1-8, 2000-10-05

[4] 土方嘉徳, 情報推薦・情報フィルタリングのためのユーザプロファイリング技術, 人工知能学会論文誌,Vol.19,No.3,2004

[5] D. Blei, A. Y. Ng and M. Jordan, Latent dirichlet allocation,Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[6] 大友賢二，項目応答理論，2001，大修館書店

[7] 堀幸雄，中山堯，“ユーザの検索要求に基づいた興味関心の定量的評価”，情報知識学会誌 Vol.16， No.2， pp.33-38 (2006)

[8] 大倉　務，清水伸幸，中川裕志，”スケーラブルで凡庸なブログ著者属性推定手法”，情処学自然言語処理研報，2007

[9] 喜連川 優, 情報爆発のこれまでとこれから (小特集　情報爆発が創り出すサイバーフィジカルな情報処理), 電子情報通信学会誌 vol94,No.8,pp.662-666(2011)

[10] 語彙数推定テスト,http://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/goi-test.html

[11] Standard Vocabulary List 12,000,http://www.alc.co.jp/eng/vocab/svl

[12] 天野成昭他 (1999) 『NTT データベースシリーズ日本語の語彙特性（第 1 期）』三省堂