3E1-2

潜在トピックを用いた専門家推薦手法と知的創造活動への影響評価

An Expert Recommendation Method using Latent Topics and its Evaluation for Intellectual Activities

薦田和弘　　大澤幸生
Kazuhiro Komoda　Yukio Ohsawa

東京大学工学系研究科
School of Engineering, the University of Tokyo

In expert recommendation, it is important for a system user to utilize the result so that he can make his intellectual activities better. In this paper, using latent dirichlet allocation, we recommend experts by considering one's specialty of a topic, the similarities and relationships between the system user and the experts. We evaluate how these elements affect the user's intellectual activities by utilizing the attributes and behaviors of the recommendation candidates as well as their ranking list.

1. はじめに

大学での勉学・研究活動，会社等の組織での業務，日常生活においては，問題解決に必要な知識を収集し活用することが重要となる．専門家発見は，問題解決，質問への回答，話題についての詳細な情報提供等を行える人材として，対象の問題について適切な能力と知識を持つ専門家を特定する分野であり，情報検索の観点から企業内の専門家を発見する研究 [Balog 06]，ユーザに合わせた専門家推薦 [Smirnova 11] などが行われている．これらの研究では，単語検索による専門家発見自体が目的であり，ユーザが推薦結果を活用する方法は不明確である．

一方，環境中の状況・事象から形成される多層的な系列（シナリオ）を選択していく主体的な意思決定を支援するため，生起確率に応じて各事象にポテンシャルの概念を導入するモデルが提案されている [大澤 06]．近い生起確率を持ち高頻度で共起する事象同士を近傍に配置するポテンシャル場において，シナリオ間の遷移の容易さを表現でき，異なるシナリオ間でユーザが情報を収集し意思決定を行う指針を得られる．

本稿では，ヒューマンネットワークを対象にポテンシャルモデルを適用し，ユーザと，特定の話題に関する専門家の間の知識伝達を支援するシステムの構築を行う．具体的には，潜在的ディリクレ配分法 (LDA) によってユーザを潜在的なトピックで表現し，ユーザのグループ分けや，ポテンシャル場に応じた遷移の容易さを検討する．本研究は，専門家の順位付けに留まった従来の専門家推薦の研究を，ユーザによる推薦結果の活用に発展させる上で重要であると考える．

2. 提案システム

本章ではシステムの概要と提案手法について説明する．

2.1 システム概要

システムの概要を示す．処理の流れは以下の通りである．

(1) コーパスの作成

対象とする文書集合が日本語の場合，京都大学情報学研究科で開発された MeCab を用いて形態素解析し，以降の処理に必要な形態素（言語として意味を持つ最小単位）を抽出する．各文書は，抽出された形態素を出現順に並べたバスケットと見なせる．(2) 以降の初期条件として，$k = 1, h_k \gg 1$ とする．

(2) 生成確率の高い文書の選定

LDA（2.2 節）によって計算された文書の生成確率に基づき，高さ h_k について，$\exp(-h_k)$ 以上の生成確率 $p(d)$ を持つ文書 d を全て選出しノードとする．ノード数は M_1 とする．

(3) 文書間距離に基づくクラスタリング

(2) で選出された文書について，距離の近い上位 $M_2(= M_1)$ 組をリンクで相互に結合しクラスタリングを行う．距離の計算には topical difference [Weng 10] を用いる．文書 d と文書 d' が正規化された確率分布 $DT'_{d\cdot}, DT'_{d'\cdot}$ で表現できるとき，topical difference は，$D_{JS}(d,d') = \frac{1}{2}(D_{KL}(DT'_{d\cdot}||M) + D_{KL}(DT'_{d'\cdot}||M))$ を用いて

$dist(d,d') = \sqrt{2 * D_{JS}(d,d')}$

で表現される．ただし，M は確率分布 $DT'_{d\cdot}$ と $DT'_{d'\cdot}$ の平均，$D_{KL}(P||Q) \equiv \sum_e P(e) \log \frac{P(e)}{Q(e)}$ である．

(4) クラスタの描画

(3) を第 k レベルのクラスタリングとし，$h_k \leftarrow h_k + \delta, k \leftarrow k+1$（ただし $0 < \delta \ll 1$）として (2) に戻る．

(5) 専門家推薦とポテンシャル場の可視化

トピック t に対する文書 d の専門性を DT'_{dt} とし，この値が大きい文書をユーザに推薦する．また，(2) から (4) の手順を一定回数繰り返し，$p(d)$ に応じた文書の色付けを行い，文書間の topical difference に基づいた文書間の繋がりやすさをポテンシャルモデル（2.3 節）として可視化する．ユーザによる文書間の関係把握や，推薦結果の活用を促す．

2.2 LDA(Latent Dirichlet Allocation)

ユーザが興味を持っている潜在的なトピックを表現する方法として，本稿では代表的なトピックモデルである LDA を用いる [Blei 03]．LDA は，一文書に複数トピックが含まれることを表現できる，文書生成過程の確率的モデルであり，文書 d がトピック z の多項分布，トピック z が単語 w の多項分布で表現される．文書-トピック多項分布 DT とトピック-単語多項分布 TW は各々ディリクレ事前分布を仮定する．

モデルの推定方法としては Gibbs Sampling を用いる [Griffiths 04]．Gibbs Sampling では，サンプリングしたい確率分布 $f(\boldsymbol{x}) = f(x_1, \cdots, x_M)$ について，各ステップで 1 つの変数 x_i の値を置き換える．その際，残りの変数の値を固定し

連絡先: 薦田和弘，東京大学工学系研究科，〒 113-8656　東京都文京区本郷 7-3-1，03-5841-2908，6643030505@mail.ecc.u-tokyo.ac.jp

表 1: 論文に関する情報

ユーザ ID	エネルギー	論文タイトル
u17	517.22	Experience, Engagement, and Shikake
u14	347.92	Understanding and Applying Trigger Piggybacking for Persuasive Technologies
u7	353.78	A Logic-Based Methodology for the Formalization of Shikake Principles and Examples
u9	488.24	Might Avatar-Mediated Interactions Rehabilitate People Suffering from Aphasia?

た条件での，対象変数の条件付き分布 $f(x_i|\boldsymbol{x}_{-i})$ に従って抽出した値を得る．LDA においては最終的に以下の式が得られ，この確率に従って順次サンプリングを行うことで各単語のトピックが得られる．

$$P(z_i|\boldsymbol{z}_{-i},\boldsymbol{w}) \propto \frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta}\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha} \approx p(w_i|j)p(j|d_i)$$

ここで，全ての文書を繋げた一つの長い単語列における位置を i，トピックを区別する添字を j とする．右辺第一項がトピック j での w_i の確率，第二項が文書 d_i でのトピック j の確率を表す．

2.3 ポテンシャルモデル

ポテンシャルエネルギーの，チャンス発見の文脈での応用について言及する．

2.3.1 チャンス発見におけるポテンシャル

事象 X の生起確率 $p(X)$ に対して，X のポテンシャル値が $E = -\log p(X)$ であると仮定する [大澤 06]．本稿では，各ユーザをノードとし，その間の繋がりを考慮するため，各ユーザの生成確率を新たに導入する．各ユーザ u を一つの文書 $d(u)$ とすると，2.2 節の LDA モデルで求めた $p(t|d(u))$（t はトピック）に基づき，$d(u)$ の文書集合全体での実現しやすさを

$$p(d(u)) = \prod_{n=1}^{N}\sum_{t=1}^{T} p(t|d(u))p(w_n|t)$$

と書ける．$E = -\log p(d(u))$ と 2.1 節の h_k が対応する．

3. 提案システムの適用例

本章では提案システムを論文集合に適用した例を示す．

3.1 システム適用対象

2013 AAAI Spring Symposium の Shikakeology ワークショップの 19 論文を対象とし，各論文の第一筆者を要旨から作成された文書で表現する．対象ユーザの集合を S' とすると，$|S'| = 19, |V| = 640$ の文書集合を得た．

3.2 結果の検討

表 1 に論文に関する情報を示す．文書の専門性を DT'_{dt} で比較し，コミュニケーションや学習の専門家を u9 とする．実際に，u9 は失語症の人々に対してアバターを用いたリハビリテーションを提案している．次に，図 1 で，経験や仕掛学 (Shikakeology) を論じた u17 と u9 の間での連携の可能性を考える．ポテンシャル図において色が濃いノードはエネルギーが低い谷に相当し，他のノードとの連携が容易でないと考えられるため，現状では両者の間に障壁があると解釈できる．

図 2 では，u17 から u9 に至る経路上の u14 と u7 のエネルギーが上昇しており，u17,u14,u7,u9 のエネルギーの差は小さく，u17 と u9 の連携はより容易になると期待される．これは，2.3.1 節より $p(d(u))$ を下げること，即ち専門性 $p(t|d(u))$ が高いトピック（論理的な定式化や手法）については典型的でない単語を増やし，低いトピック（理解やエンゲージメント）については典型的な単語を増やすことに対応する．

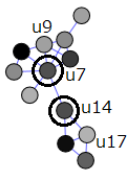

図 1: 初期状態

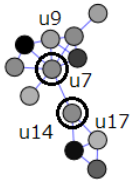

図 2: u14 と u7 のポテンシャルを下げた状態

4. おわりに

本稿では，潜在トピックを用いて単語レベルの差異に依存しないユーザ間距離表現と専門家発見を実現し，ポテンシャル場での可視化によりユーザが専門家推薦の結果を活用できる可能性を示した．プログラミングエラー解消のような自分の問題解決や，求職・転職活動において自分の能力を適切な集団に売り込む際に専門家に効果的にアプローチする戦略を立てることが可能になると期待される．

今後の課題として，対象とするユーザ集合を適切に選定し，手法の有効性について多くの被験者を通して実験を行うことが挙げられる．

参考文献

[Balog 06] K. Balog, L. Azzopardi, M. de Rijke: Formal Models for Expert Finding in Enterprise Corpora, ACM SIGIR (2006).

[Smirnova 11] E. Smirnova and K. Balog: A User-Oriented Model for Expert Finding, Springer (2011).

[大澤 06] 大澤幸生: チャンス発見のデータ分析, 東京電機大学出版局 (2006).

[Weng 10] J. Weng, E. P. Lim, J. Jiang and Q. He: TwitterRank: Finding Topic-sensitive Influential Twitterers, WSDM (2010).

[Blei 03] D. M. Blei, A. Y. Ng and M. I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research (2003).

[Griffiths 04] T. L. Griffiths, and M. Steyvers: Finding scientific topics, Proceedings of the National Academy of Sciences (2004).