

コールドスタート問題解決のためのソフトクラスタリング活用の提案

大西 佑紀[†] 菅谷 信介[‡]株式会社 ビズリーチ[†]

1. 概要

協調フィルタリングに代表される推薦アルゴリズムは、評価データの蓄積が乏しい新規ユーザや新規アイテムの推薦が難しいとされるコールドスタートという問題を本質的に抱えている[1]。しかし、Webサービスにおけるユーザやアイテムの流動性がますます高まるにつれて、その問題に対処するという現場でのニーズは非常に大きい。本稿では、推薦対象と被推薦対象に対しソフトクラスタリングを活用する推薦アルゴリズムを提案し、コールドスタート問題に対処する実務的な有効性を示した。

2. システム概要

本研究では転職サイト CareerTrek[2]のデータを利用した。求職者、求人の職種業種等の属性情報に加え、求人への閲覧、応募履歴などの行動ログの活用が可能である。同サービスにユーザ登録している求職者に対し、求人を推薦するシステムを想定する。本章では本システムにおける推薦アルゴリズムの処理の流れを説明する。本アルゴリズムの概要図を図1に示す。求職者、求人に対してそれぞれソフトクラスタリング処理を適用し、クラスタ内の既存会員の既存求人への評価データを基に、クラスタ間の選考スコアを算出する。詳細は本章で説明する。

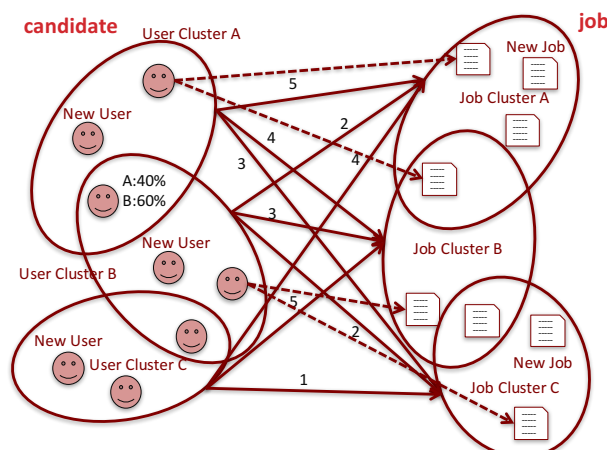


図1 アルゴリズム概要図

2.1. 処理の流れ

まず、全求職者 $C = \{c_1, c_2, \dots, c_{N_c}\}$ と全求人 $D = \{d_1, d_2, \dots, d_{N_d}\}$ それぞれについて行動履歴・属性情報に基づいたソフトクラスタリング処理を施し、各クラスタへの所属確率を算出する(2.2)。その後、各求職者クラスタ $R = \{r_1, r_2, \dots, r_{N_r}\}$ から各求人クラスタ $S = \{s_1, s_2, \dots, s_{N_s}\}$ への選好スコアを算出する(2.3)。選好スコアに基づき各求職者への推薦クラスタの選択を行い、選択されたクラスタ内の求人から求人のフィルタリングを行い、推薦求人 N 件を確定させる(2.4)。

2.2. ソフトクラスタリング

本節では、推薦対象・被推薦対象に対して使用したクラスタリング手法について述べる。

求職者のソフトクラスタリング手法

求職者の属性情報(職種・業種・年齢等)を基にクラスタリングする。各求職者の持つ各属性値を区分値に変換し、区分値をクラスタと見なす。求職者の持つ各区分値への所属確率を等分に配分することにより、求職者 c_i のクラスタ r_k への所属確率を $p(r_k | c_i)$ とする。

求人のソフトクラスタリング手法

求人のソフトクラスタリングには LDA (Latent Dirichlet Allocation) を使用した。LDA は一つの文書に複数のトピックが存在していると仮定し、そのトピックの分布を離散分布としてモデル化する手法であり[3][4]、自然言語処理の分野でよく用いられる。本システムにおいては、求人集合の潜在トピックを求人クラスタ S とし、求人のトピック混合比を、対応するクラスタへの所属確率とする。求人 d_j のクラスタ s_l への所属確率を $q(s_l | d_j)$ とする。

なお、本稿で提案するアルゴリズムはより一般的な枠組みとして、本節に述べる以外にも様々なソフトクラスタリング手法が適用可能である[5]。

2.3. 選好スコアの定義

求職者クラスタ R の各要素 r_k から求人クラスタ S の各要素 s_l への選好スコア $Score_{r_k, s_l}$ 、また各求職者 c_i から求人クラスタ S の各要素 s_l への選好スコア $Score_{c_i, s_l}$ を定義する。 $rate_{i,j}$ は求職者 c_i から求人 d_j への評価値を表す。

“A Proposal on Utilization of Soft Clustering to Solve Cold-start Problem.”

[†] Yuki Ohnishi – BizReach Inc.

[‡] Shinsuke Sugaya – BizReach Inc.

$$Score_{r_k, s_l} = \frac{\sum_i^{N_c} \sum_j^{N_d} p(r_k | c_i) q(s_l | d_j) rate_{i,j}}{(\sum_i^{N_c} p(r_k | c_i)) (\sum_j^{N_d} q(s_l | d_j))}$$

$$Score_{c_i, s_l} = \sum_k^{N_k} p(r_k | c_i) Score_{r_k, s_l}$$

本システムにおいては、求職者毎に算出された選好スコア $Score_{c_i, s_l}$ の高い上位 3 クラスタを推薦クラスタとして選択した。

2.4. 求人選択(フィルタリング)

前節で推薦されたクラスタ内の求人から推薦対象となる求人を絞り込む。本システムでは希望条件によるフィルタリングや、求人類似度によるフィルタリングを適用しており、業務要件に合わせて様々なフィルタを追加することが可能である。それぞれの詳細については本稿では割愛する。

3. 実験・評価

提案手法の性能を計測するために、比較実験を行った。本システムは求職者が推薦された求人に対して、“like”, “hate”をフィードバックできる仕組みが備わっている。手法の性能は“like”と“hate”の odds 比で計測するものとする。

3.1. 実験手法

提案手法と旧手法について比較実験を行った。旧手法とは、求職者の属性情報を基に人気求人を推薦するルールベースによる手法である。実験は 2 週間行い、手法間に優位な差が認められるかどうか確認した。

3.2. 実験対象

実験対象となるセグメントとして、コールドスタート問題の対象となる「新規ユーザ」と、「既存ユーザ」に分割する。上記 2 手法をそれぞれのセグメントにランダムに適用する。また、コールドスタート問題に対する提案手法の有用性評価のため、「新規ユーザ」は「求人に対して評価を行っていないユーザ」と定義する。また、実験対象はシステム登録している求職者であり、そのユーザ数は約 22 万人、推薦対象となる公開求人数は約 89,000 件(2016 年 9 月当時)である。

3.3. 実験結果・考察

図 2 に実験期間中の新規ユーザの odds 比の推移を示す。赤線が提案手法によるレコメンドを示し、旧手法と比較して odds 比が改善していることがわかる。また、図 3 に実験期間中の既存ユーザ向けの odds 比の推移を示す。こちらも同様に赤線が提案手法によるレコメンドを示し、旧手法と比較して、0.05～0.1 程度の odds 比の改善が見られる。また、odds 比算出の際に用いたデータ量は十分に取得しており、統計的に優位 ($p < .05$) であることをカイ二乗検定により確認している。

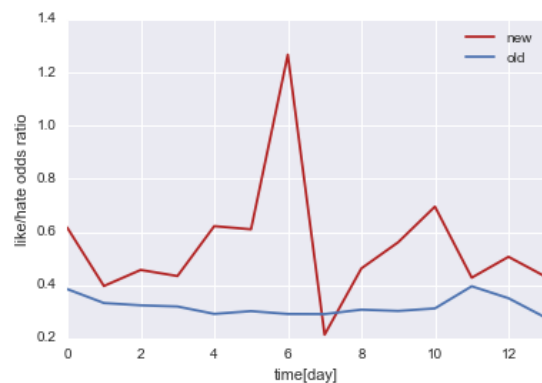


図 2 新規ユーザの odds 比推移(赤線:提案手法, 青線:旧手法)

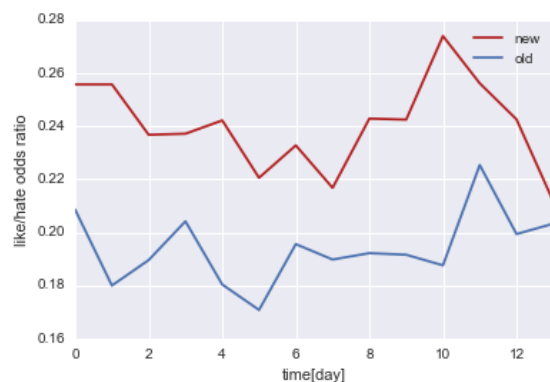


図 3 既存ユーザの odds 比推移(赤線:提案手法, 青線:旧手法)

4. おわりに

本稿では、コールドスタート問題に対処する推薦アルゴリズムを提案し、その性能を評価した。実験結果より、コールドスタートの対象となる新規ユーザへの推薦だけでなく、既存ユーザへの推薦性能向上という観点でも一定の成果を上げることができたといえる。

参考文献

- [1] F. Ricci, K. Rokach, B. Shapira (2015): “Recommender Systems Handbook Second Edition”, Springer
- [2] CareerTrek, <https://www.careertrek.com/>
- [3] Blei, D.M., Ng, A.Y., Jordan, M.I. (2003): “Latent dirichlet allocation”. Journal of Machine Learning Research **3**, 993-1022
- [4] 佐藤一誠 (2015): “自然言語処理シリーズ 8 トピックモデルによる統計的意味解析”, コロナ社
- [5] D. Reynolds (2008): “Gaussian Mixture Models”, Encyclopedia of Biometric Recognition, Springer, Journal Article