2012 年度卒業研究概要

ベクトル空間モデルを用いた知識共有コミュニティでのユーザマッチング

大谷紀子研究室
0932138 高橋柊

1. 研究の背景・目的

今日，Wikipedia を始め Yahoo!知恵袋，ソーシャルブックマークなど，Web 上にて多数のユーザにより創られる知識共有コミュニティが数多く存在する．知識共有コミュニティの活発化には投稿数の増加が必要不可欠であり，ユーザに対して投稿項目を推薦することは有用である．多くの場合，ユーザに投稿項目を推薦するための興味範囲特定には，投稿内容から算出される重要語が用いられる．しかし，テキストから重要語を得る手法は投稿内容の文体や文章長に大きく左右されるという問題点がある．ユーザの投稿項目への関与の有無という 2 値の情報を用いることにより，投稿内容や文章長に左右されることなく興味範囲の特定ができる．ユーザに対し，興味範囲の類似するユーザの投稿項目を提示することで，投稿項目の推薦が可能となる．

本研究では，ユーザの投稿項目への関与の有無という 2 値の情報を用いることで，ユーザの興味分野を取得する．また興味分野の情報を元に類似ユーザを検出することで，ユーザの新たな投稿を促進する．

2. マッチング手法

多くの知識共有コミュニティでは履歴情報や投稿ユーザ情報が開示されているため，ユーザの投稿履歴取得は比較的安易である．本研究ではベクトル空間モデル[1]を用い，ユーザの興味範囲を表す興味ベクトルを生成する．興味ベクトルの次元はマッチング対象となる全ユーザの投稿項目数とし，各成分はユーザが当該項目に投稿した場合は 1，投稿していない場合は 0 となる．したがって，興味ベクトルは通常 1 の割合が次元数に対し非常に少ない疎ベクトルとして与えられる．

興味範囲が重なっているユーザを抽出するために，ユーザ間で興味ベクトルの類似度を算出する．2 人のユーザU_iとU_jの類似度$sim(U_i, U_j)$は，両ユーザの興味ベクトル$\overrightarrow{u_i}$と$\overrightarrow{u_j}$のなす角の余弦値により求められる．

$$\boldsymbol{sim}\left(\boldsymbol{U_i}, \boldsymbol{U_j}\right) = \frac{\overrightarrow{\boldsymbol{u_i}} \cdot \overrightarrow{\boldsymbol{u_j}}}{|\overrightarrow{\boldsymbol{u_i}}||\overrightarrow{\boldsymbol{u_j}}|} \tag{1}$$

$sim\left(U_i, U_j\right)$は興味ベクトルの向きが一致しているとき最大値の 1 をとり，直交していれば最小値の 0 をとる．値が 1 に近いほどU_iとU_jの興味範囲は似ているといえる．ユーザU_iに対し，$sim\left(U_i, U_j\right)$が最大の余弦値をとるユーザU_jの投稿項目を提示することで，投稿項目の推薦が可能となる．

興味ベクトル間の類似度$sim\left(U_i, U_j\right)$では，興味範囲が似ていても投稿項目に重なりがなければ，類似ユーザとして判定されない．$sim\left(U_i, U_j\right)$が 0 であっても，$\overrightarrow{u_i}, \overrightarrow{u_j}$2 つの興味ベクトルと類似する$\overrightarrow{u_k}$を持つユーザ$U_k$が存在する場合，$U_i, U_j$間には類似性があると推測される．興味ベクトルを使用して，非階層クラスタリングの代表的アルゴリズムである K-means 法を用いユーザをクラスタリングする．セ

ントロイドには各クラスタの平均ベクトルを，距離計算には正規化した興味ベクトル間のユークリッド距離を用いる．

ユーザ集合をX，興味ベクトルをu，クラスタX_iはユーザ集合の網羅的で互いに素な部分集合，$\bar{x_i}$はX_iのセントロイドとしたときの処理手順を以下に示す．

① ユーザ集合をランダムにk個のクラスタに分割し初期クラスタを作成

② クラスタについてセントロイド$\bar{x_i} = \sum_{u \in X_i} u$を算出し，$\bar{x_i}$を正規化する

③ 興味ベクトル$u \in X$をセントロイド$\bar{x_i}$とのユークリッド距離が最小となるクラスタX_iへ移動

④ クラスタに変化がない，もしくは指定回数を超えたら終了しクラスタ$\{X_i\}$を出力．そうでない場合②へ戻る

同一クラスタX_iに含まれているユーザは，興味範囲が類似していると考えられる．ユーザに対し，同じクラスタ内にいるユーザの投稿項目を提示することで，投稿項目の推薦が可能となる．

3. 評価実験

2012 年 11 月現在提供されている日本語版 Wikipedia のダンプデータを用い，投稿ページがカテゴリに属している 1050 ユーザを対象とした評価実験を行った．投稿項目は，ユーザの投稿したページが属しているカテゴリ 315589 個とし，ユーザの興味ベクトルを生成した．また全ユーザの持つ興味ベクトルを対象とし，$k = 40$とした K-means 法を用いたクラスタリングを行った．

実験データから K-means 法で作成したクラスタの精度を評価する．評価手法には K-分割交差検定を用いる．投稿項目を 5 個に分割し，5 分の 4 を訓練用データとして興味ベクトルに，残った部分を検証用データとする．訓練用データを K-means 法によりクラスタリングする．検証用データの投稿項目ごとに，訓練用データから得られた各クラスタのユーザが投稿している割合を算出し，標準偏差をとったものを評価値とする．

表 1：評価実験結果（小数点第 5 位以下切り捨て）

	評価値	カテゴリ名	投稿ユーザ数
最大値	0.1384	日本海軍の運搬船	14
最小値	0.0433	栄市の企業	10

出力される評価値が大きいほど，同じクラスタに属しているユーザが，同じ項目に投稿しているといえる．評価値は，すべての投稿ユーザが同じクラスタに含まれる場合，最大値の 0.1558 をとり，すべてのクラスタが同じ割合の投稿ユーザを持っている場合最小値の 0 をとる．検証用データで投稿数が上位 5%に含まれる項目のうち，評価値が最大および最小となった項目の評価値，カテゴリ名，投稿ユーザ数を表 1 に示す．

4. 考察

表 1 の評価値 0.1526 をとる「日本海軍の運搬船」は 40 クラスへの投稿者中，88.58%が 1 つのクラスタに属しており，クラスタリングの精度が高い．評価値 0.0433 の「栄市の企業」は投稿者が複数のクラスタに分散して存在するため，クラスタリングの精度が低い．学習データに似ている項目が多く存在する場合，ユーザの興味範囲に基づくマッチングが可能であるといえる．また，評価値の低い項目に関して重み付けを行うことで，より正確なクラスタリングが可能であると考えられる．

参考文献

[1]大谷紀子，“情報検索におけるベクトル空間モデルの応用”武蔵工業大学環境情報学部紀要，第 5 号，pp.99-109，2004.