2006 年度卒業研究概要

日記へのコメントに基づくマッチメイキング

大谷　紀子研究室
0332123　塩澤　健太

１．研究の目的

ウェブログやソーシャルネットワーキングサービス(以下 SNS)に代表されるように，今日ではウェブサイト上で日記を公開する人が増えている．ウェブサイト上の日記では，コメントを付加できる機能が特徴といえる．コメントとは，日記の本文を読んだ閲覧者(以下 ユーザ)が気になった箇所へ対して書き込んだ意見や感想である．一つの話に対して感じ方が人によって異なることと同様に，ユーザ全員が本文の同じ部分に対してコメントを返すわけではない．同じ部分に反応したユーザ同士ならば，嗜好の類似や共通点があると考えられることから，コメントに基づいてユーザを分類することが可能といえる．本研究では，自動的に趣味や嗜好の類似したユーザを見つけ出すことを目的として，日記に対するコメントからユーザを分類するための手法を提案する．

２．ユーザ分類の手法

ユーザを分類するためには，コメントからユーザが反応を示した箇所を抽出する必要がある．ユーザごとに反応箇所を抽出し，他のユーザとの類似度を判定する．ユーザの反応箇所を抽出するためには，コメント中から本文と同じ言葉を見つければよい．また文章を特徴付ける単語は名詞であると考え，本文とコメントから名詞を抽出する．本研究で抽出する名詞の定義を以下に示す．

１．　2 文字以上の漢字列
２．　2 文字以上のカタカナ列
３．　2 文字以上のアルファベット列

ユーザ間の類似度の判定には，ベクトル空間モデルを用いる．ベクトル空間モデルとは単語の出現頻度等を軸として, 文書を 1 つのベクトルで表現し, ベクトルの向きによって内容を判断する技法である．

本文，コメントから単語を抽出し，各々重複したものを成分として，コメントをつけたユーザをベクトルで表す．例えば本文とユーザ A のコメントに「WORD」という語句が 5 回出現した場合，ユーザ A の持つベクトルは，WORD という軸の成分が 5 となる．別の日記の本文に「SAMPLE」という語句が 3 回出現した場合，ユーザ A の持つベクトルは WORD 軸の成分が 5 となり，SAMPLE 軸の成分が 3 となる．また別の日記にも「WORD」が出てきた場合,「WORD」は前述の「WORD」とは異なり，別の本文に反応した単語と考え，ユーザ A のベクトルは WORD 軸の成分が 5 となり，SAMPLE 軸の成分が 3 となり，WORD 軸の成分が 2 となる．ユーザごとのベクトルを生成後，それぞれベクトルの内積による類似度[1]を算出する．

３．結果

2006 年 9 月から 12 月分の 1 アカウント分の日記データを分類した. 本文とコメントにおける重複語

句が検出されたユーザ 19 名において, 類似度が 0 を超えた 2 ユーザの組み合わせ 29 組の類似の再現性を調査したところ, 閾値を 0.3 としたときの再現率は 11.59%, 適合率は 66.67%となった. 検証には「同じ高校の出身者」や「趣味がバイクのユーザ」など，所属や趣味に共通点のあるユーザ同士を正解とした．各閾値に対する再現率と適合率を図 1 に示す.

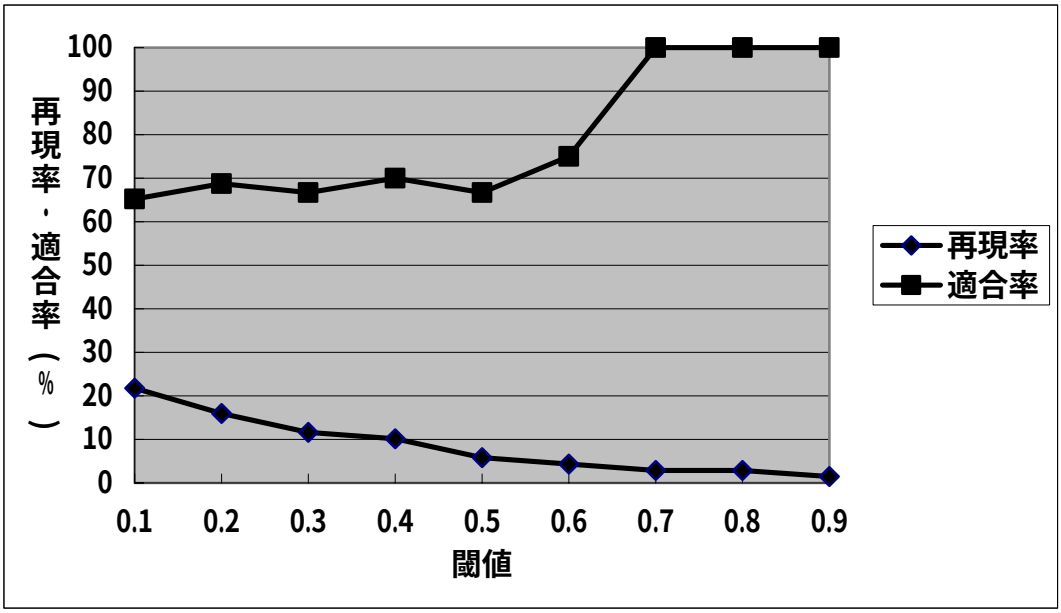

図 1：再現率と適合率

４．まとめ

19 名中 2 名の全組み合わせは 171 通りであるのに対し，同じ単語に反応を示したユーザの組み合わせは 29 通りだった．要因として提案手法では重複した単語の検出が難しかったということが考えられる．しかし，重複語句の数が最も多い 2 名を調べたところ，類似度は 0.231 となり嗜好が似ているとはいえない．したがって重複語句数の多さが，必ずしも類似度を上げるとは限らないといえる．類似度が 0.976 の組み合わせもあったが, 重複単語が 2 語で他に重複語句がなかったことが上記の結果を導いたと推測できる．また適合率に比べて再現率が著しく低いという結果が出た．したがって本研究の手法では，ユーザの有効な分類ができなかった．理由として，特徴付ける語句や，重要でない語句との区別をしなかったことが挙げられる．また上記の「別の日記から抽出された単語」を同じ単語であると見なせば，結果の違いが考えられる．さらに反省点を組み合わせれば，より精度の高い分類が可能であると考える.

精度の向上が課題となるが，本研究の発展形として類似度で分類したユーザを自動で，ウェブ上のコミュニティへ参加できるようにしたり，ユーザ同士を自動で紹介したりするなど，ウェブサイト上の日記に新たな機能を付加することができると考える.

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, 第五号, pp.99-109，2004.