5N-2

つぶやきに基づいたユーザ推薦システムの提案

原　克彬† 中山　泰一†
†電気通信大学情報工学科

1 はじめに

ここ数年のソーシャルネットワーキングサービス（以下 SNS）の成長は目覚しいものがあり，注目を集めている．中でも，twitter[1] やモバゲー等の成長は著しく，twitter は 2011 年 9 月にアクティブユーザが 1 億人を越え，モバゲーでも 2011 年度にユーザ数が 3000 万人を越えたとの報告があった．

近年の SNS にはユーザ検索機能とは別に，ほかのユーザを友人に推薦するシステムが存在する．既存のシステムの従来の方法では SNS 内におけるユーザの友人の中で共通している友人などを推薦しており，現実の社会での友人などを推薦するのが主な目的であった．しかし，SNS 内での友人との接触はユーザ検索機能を利用すれば十分であり，ユーザ推薦システムは別の役割を担うことができると考えられる．そこで本研究では新たな手法の提案として，趣味や嗜好から新たなユーザを推薦するシステムを提案した．twitter を題材として，ユーザのつぶやいた内容から割り出した趣味嗜好を基に，推薦するユーザ名を出力することを目標とする．

2 提案手法

2.1 利用するアルゴリズム

2.1.1 TFIDF 法

TFIDF 法 [2] とは文書中の単語に関する重みづけの一種であり，主に情報検索や文章要約などの分野で利用される．ひとつの文書内にて，あるキーワードが出現した回数である単語の出現頻度 tf と，そのキーワードが出現した文書数の逆数である逆文書頻度 idf という二つの要素から計算される．本研究では TFIDF 法を用いて抽出した名詞の重み付けを行った．あるキーワードの TFIDF 値を求める計算式は全文書数を M，キーワードが出現した文書数を X とした時以下のとおりである．

$$tf \log \frac{M}{X} \tag{1}$$

2.1.2 ベクトル空間法

本研究では文書を多次元空間上のベクトルとして表現し，ベクトル空間法 [3] 用いて二つのベクトルを比較することにより類似度を調べる．つまり，ベクトルの方向は文書の特徴であるので，二つのベクトルのなす角が小さいほど似ているということである．本実験においては文書を形態素解析し，出現した名詞を文書の要素として各ベクトルを求めた．ここでユーザ A のベクトルを $\overrightarrow{A}$，ユーザ B のベクトルを $\overrightarrow{B}$ とすると，ユーザ A とユーザ B の類似度を求める計算式は以下のようになる．

$$\frac{(\overrightarrow{A} \cdot \overrightarrow{B})}{(|\overrightarrow{A}||\overrightarrow{B}|)} \tag{2}$$

2.2 提案システムの流れ

本システムの流れとしては，あるユーザの最新のつぶやきを twitterAPI を通して 200 件取得する．取得する際に，リツイートやつぶやきに対して返信を行ったときに発生する，RT 表示やユーザ名の除外を行う．そして，そのつぶやきデータを形態素解析システム IGO[4] を用いて解析を行い，結果より名詞のみ抽出する．形態素解析は登録されている辞書に基づいて行われるため，辞書に登録されていない未知語や語尾の表現によっては解析の邪魔になりうる語が発生するので，一文字のひらがなや記号を結果より除外する．残った単語を文書の要素として TFIDF 法を用いてベクトルを算出する．そのベクトルがユーザの特徴であり，ベクトルの類似度を求めることで，ユーザ の類似度を求める．ベクトルの計算にはベクトル空間法を使用する．ここで出力される類似度は値が高ければ高いほど，ユーザ間で同様の内容のつぶやきを行っている事となる．

2.3 特徴

本システムの手法は従来のシステムのように知人・友人を推薦することが主な目的の手法とは異なり，趣味や嗜好の対象が似ているユーザを推薦することがで

Suggestion of Tweets-based User Recommending System
Katsuaki Hara† Yasuichi Nakayama†
†Department of Computer Science, The University of Electro-Communications
182-8585, Chofu, Japan
hara-k@igo.cs.uec.ac.jp

きる．また，従来は趣味や嗜好の似通ったユーザを探す際にはプロフィールを参照して対象のユーザの情報を得るのが一般的であったが，編集時期が不明であるプロフィールより，直近の情報であるつぶやきを機械的に解析したほうがユーザの解析情報として精度の高いものが得られると考えられる．本システムでは，つぶやきの機械的な解析を実現することにより，精度の高い解析とプロフィールを閲覧するなどのユーザの手間の軽減ができると考えている．

3 実験

3.1 実験方法

twitter のユーザを無作為に 21 人選択した．被験者と 21 人の類似度を計算し，最高値・中央値・最低値を持つ三人のユーザを抽出した．被験者には結果を知らせずに，その 3 人のユーザのつぶやきのデータを見てもらい，内容重視で 3 人との親近感の感じ方に順位をつけてもらった．被験者は 5 名である．

3.2 実験結果

表 1 は各被験者が最高値・中央値・最低値の類似度を持つのユーザをどのように順位づけしたかをまとめたものである．被験者に対して最高の類似度を持つ各ユーザに関しては 5 人中 4 人が一番親近感を感じると解答したものの，中央値に関しては解答の一致が 0，最低値も 5 人中 1 人が一致という結果になった．

表 1: 各類似度のユーザへの被験者の評価

類似度 / 被験者	最高値	中央値	最低値
被験者 A	1	3	2
被験者 B	1	3	2
被験者 C	1	3	2
被験者 D	2	1	3
被験者 E	1	3	2

3.3 考察

今回の実験は，ユーザの特徴をつぶやきの名詞を抽出することで求めた．類似度が最高値のユーザはどの被験者も親近感を感じたという結果を得たが，中央値と最低値においては一致率が極めて低くなった．その背景としてユーザの特徴が名詞だけに現れるものではないことが主な原因として考えられる．

名詞だけを抽出した場合，つぶやいている内容の題材については一定の精度が得られると考えられるが，その題材をどう思っているかを考慮することができないため，つぶやきの内容に差異が生じてくる．たとえば，ユーザ A が「私は本が好きだ」とつぶやき，ユーザ B が「私は本が嫌いだ」とつぶやいたとする．このとき形容詞である「好き」と「嫌い」の 2 語を無視し，名詞だけを抽出すると同一の結果を得る．文書の内容が正反対であるのに対し本システムの解析結果は内容が一致していると出力するため，結果に齟齬が生じてしまう．

中央値や最低値については精度が高いとは言えない結果になったが，最高値については 8 割一致という結果となった．推薦システムはその性質上，上位の結果しか必要としていないため，極論を言えば最高値の一致率が高ければ中央値と最低値については無視してもよいと考えられる．また，今回の実験では無作為に選択したユーザ 21 人を対象としていたため類似度の最高値もそこまで高くならなかった．これらの事を踏まえると，twitter の全ユーザを対象にした場合，推薦システムに必要な上位の類似度は本システムにて高い精度で求められると考えられる．

4 まとめと今後の課題

本稿では twitter のつぶやきを TFIDF 法を用いてベクトル化し，ベクトル空間法にて類似度を計算することによりユーザの趣味嗜好を加味したユーザ推薦システムの手法の提案を行った．本システムを利用することにより，趣味や嗜好の似通ったユーザを探す際に生じるプロフィールの閲覧等の手間を省くとともに精度の高い解析情報を得られると考えている．しかし，考察でも述べたとおり，名詞だけを用いてユーザの特徴を算出するのには限界があるため，以降は形容詞や動詞なども視野に入れた解析システムを考えていきたい．

参考文献

[1] twitter
http://twitter.com/

[2] G. Salton and M. J. McGill: “ Introduction to Modern Information Retrieval ”, McGraw-Hill Book Co. (1983).

[3] 北研二・津田和彦・獅々堀正幹:情報検索アルゴリズム，共立出版 (2002).

[4] Igo - Java 形態素解析器
http://igo.sourceforge.jp/