マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案

向井　友宏　　黒澤　義明　　目良　和也　　竹澤　寿幸
広島市立大学大学院　情報科学研究科
{mukai, kurosawa, mera, takezawa}@ls.info.hiroshima-cu.ac.jp

1. はじめに

近年，Twitter [1]をはじめとするマイクロブログサービスが急速に普及してきており，日々多数の投稿がなされている．マイクロブログは従来のブログサービスと比べて非常に高速な情報伝達スピードを持つ情報発信ツールである．

本研究はマイクロブログ特有のリアルタイムな投稿を活用し，ユーザに対して効果的な情報推薦を行う手法を提案する．過去の投稿を分析することでユーザの嗜好を推測し，実際の商品データを用いて効果的なタイミングで情報推薦を行うことは実用的であり利用価値は高いと言える．

本研究ではユーザの嗜好情報獲得のためにTwitter特有の機能である「リツイート」に，また最適なタイミングの発見のために時系列における投稿数の急激な変化「バースト」にそれぞれ着目する．

2. マイクロブログについて

マイクロブログはごく短い文章をつぶやくように書き込み公開する簡易型ブログである．他のユーザをフォローすることによりソーシャルネットワーキングサービスとしての機能も有する．手軽に利用でき，リアルタイム性が高いことが大きな特徴として挙げられる．本研究では比較的利用者の多いTwitterに着目する．

2.1. Twitterについて

Twitterは代表的なマイクロブログサービスであり，多数のユーザに利用されている．Twitterでは投稿することを一般に「ツイート」と呼び，ツイートの文字数は140字に制限されている．「返信」や「リスト」，「お気に入り」機能等を持つ．また「リツイート」と呼ばれる機能も持ち，これは自分が興味を持ったツイートを引用し，自分をフォローしているユーザに手軽に流すことができる機能である．

3. 関連研究

3.1. Twitterを用いたプロファイリングに関する研究

プロファイリングの関連研究としてTwitterの持つ「リスト」や「お気に入り」機能に着目してユーザのプロファイリングを行う研究が挙げられる[2][3]．しかしリストはユーザが様々な尺度で自由に他ユーザをまとめたものであり，その分類方法によっては必ずしも嗜好情報を表しているとは言えない．またお気に入り機能に関してはユーザ毎に使用頻度に大きなばらつきがあるため，あまり機能を活用していないユーザのデータは得られにくい．またお気に入りを付けた日付等が保存されない仕様であるため時間軸を考慮したリアルタイムな推薦には不向きであると言える．

3.2. 情報推薦に関する研究

情報推薦の代表的な手法として「内容ベースフィルタリング」と「協調フィルタリング[4]」が挙げられる．

内容ベースフィルタリングはユーザの商品購入履歴データ等からユーザの嗜好情報を獲得し，嗜好とマッチするコンテンツを推薦する手法である．他のユーザに影響されないためcold-start状況にも対応できるが，セレンディピティのある推薦が少ないという問題点も持つ．

協調フィルタリングは，「嗜好が類似したユーザは同じ情報を欲する」という仮説に基づき，類似ユーザの嗜好情報を用いて推薦コンテンツを導出する手法である．セレンディピティのある推薦が可能であるが，cold-start状況では適切な推薦が出来ない問題点を持つ．

近年これらを組み合わせてお互いの短所を補ったハイブリッドフィルタリングの研究が進んでおり[5]，より良い推薦が可能となってきている．

3.3. 本研究

本研究ではユーザのプロファイリングを行うためにTwitterの持つ「リツイート」機能に着目する．あるユーザがリツイートしたツイートはそのユーザの嗜好を反映していると考えられ，利用価値は高いと考える．

また情報推薦手法としてハイブリッドフィルタリングの実現を目指す．本研究では各ユーザのプロファイリングを行った後，ユーザのクラスタリングを行う．これにより協調フィルタリングと同様の効果が得られ，セレンディピティのある推薦も可能になると考える．

4. 提案手法

本研究ではマイクロブログ特有のリアルタイム性の高い投稿を活用して効果的な情報推薦を行う手法を提案する．提案手法実現のためにはユーザの過去一定期間のつぶやきを分析しその嗜好情報を獲得する第1ステップ，時系列におけるユーザのつぶやき具合から最適なタイミングを発見する第2ステップが必要となる．以下この2ステップについて述べる．

4.1. 第1ステップ（ユーザの嗜好情報の獲得）

本研究ではTwitter特有の表現である「リツイート」

に着目する．リツイートは一般に自分が興味を持ったツイートを，自分のフォロワーへ流すために用いられる．このことからあるユーザがリツイートしたツイートはそのユーザの嗜好を反映していると考えられる．

本研究では各ユーザの過去一定期間のリツイートを収集し，リツイート中の名詞を抽出して tf-idf 値を算出し，各ユーザのプロファイルを作成する．またセレンディピティのある推薦を可能にするため，プロファイルの類似するユーザのクラスタリングを行う．

しかし同様の意味を持つ単語でも表記が異なると全く違う単語として扱われてしまうため，類似した嗜好を持つユーザをうまくクラスタリングすることが出来ない可能性が考えられる．そこでオンライン百科事典 Wikipedia の「カテゴリ」に着目し，その単語の属するカテゴリもプロファイルとして利用することを考える．これにより，より意味的に嗜好の類似したユーザのクラスタリングが可能になると考える．

4.2. 第 2 ステップ（最適なタイミングの発見）

本研究では最適なタイミングを発見する指標として，時系列におけるつぶやき数の急激な増加（バースト）に着目する．バーストが発生したとき，そのタイミングで何らかのイベントが起こったと考えられる．そしてその影響でユーザの興奮度合いが上昇し寛容な気分になっている可能性が高いと考えられる．この流れに乗じて商品情報の推薦を行うことで，よりユーザに受け入れられやすい情報推薦が可能になると考える．

本研究ではバーストを検出するために八村らの示した「文書数に基づくバースト度」の評価式 [6]を改良した式(1)を用いる．判定値 B が閾値 α を超えた時，バーストが発生したと考える．

$$B = \frac{N}{\sqrt{\bar{A}}} \cdot \frac{N-A}{N+A} \quad \cdots(1)$$

N　：その区間におけるつぶやき数
A　：直前 X 区間のつぶやき数の平均
$\bar{A}$　：直前 Y 区間のつぶやき数の平均

しかし，ネガティブなイベントが発生した際にバーストが起こる可能性もある．そのような場合ユーザが寛容な気分になっている可能性は低く，ユーザに受け入れられやすい推薦ができるタイミングとは考えにくい．そこでバーストが発生した際のつぶやき集合に対して極性評価を行い，ポジティブなイベントが発生したと評価された場合にのみ情報推薦を行うことにする．極性の評価には手掛かり語を用いる．つぶやき集合中にポジティブな手掛かり語が多く見られる場合，発生したイベントはポジティブであると判断できる．

表 1．クラスタのプロファイル（抜粋）

クラスタ2		クラスタ9	
キーワード	ベクトル	キーワード	ベクトル
大喜利	0.4751	日本	0.2593
野球	0.2908	野球	0.1855
演芸	0.1925	ロッテ	0.1748
スターズ	0.1925	監督	0.1495
落語	0.1894	試合	0.1411
競馬	0.1839	野球用語	0.1168
セリフ	0.1481	選手	0.1071
目標	0.1418	スポーツ用語	0.1056
プロ	0.1333	放送	0.1047
部長	0.0980	スポーツ関連のスタブ項目	0.0995

5. 評価実験

5.1. データ収集

本研究では頻繁にポジティブ・ネガティブ双方のバーストが発生する「プロ野球の試合」に着目した．そこで 2010 年度の日本シリーズが開催された期間中の，千葉ロッテマリーンズファンと思われる 524 ユーザのつぶやき，計 215,229 ツイートを収集した．試合の行われなかった日のつぶやきは収集対象から除外した．収集したつぶやきのうち 11 月 7 日の試合中のつぶやきを評価用データとし，それ以外を分析用データとした．

5.2　各ユーザのプロファイル作成

524 ユーザのうちリツイートを多用している 71 ユーザに対して，4.2 で述べた手法を用いて事前に引用 URL とハッシュタグ，ユーザ名，「RT」「QT」の文字列を取り除いた全リツイート中から名詞とそのカテゴリを抽出し，tf-idf を用いてプロファイルを作成した．ただし試合中のリツイートはそのほとんどが試合内容の実況ツイートに関するものであったため，ユーザの嗜好を反映しているとは考えにくく分析の対象から除外した．また記号のみで構成された名詞やひらがな・カタカナ 1 文字のみで表記された名詞，Wikipedia にエントリが存在しない名詞等も対象から除外した．名詞抽出の際には形態素解析器 MeCab[7]，解析辞書 UniDic[8]を用いた．

5.3. ユーザのクラスタリング

セレンディピティのある推薦を可能にするために，各ユーザのプロファイルを素性としてユーザのクラスタリングを行った．クラスタリングツールとして，Repeated-Bisection 法を採用している「bayon[9]」を用いた．クラスタリングにより 71 ユーザが 9 クラスタに分割された．各クラスタの中心ベクトルがそれぞれのクラスタの特徴語を表していると言えるため，中心ベクトルの値を各クラスタのプロファイルとした．得られたプロファイルの一部を表 1 に示す．比較的各クラスタのプロファイルに差異が現れており，例えばクラスタ 2 は落語等に興味を持つ噺家クラスタ，クラスタ 9 は試合中以外にも千葉ロッテに関するつぶやきに高い関心を示している千葉ロッテマニアクラスタであると言える．

表２．収集した極性評価用の手掛かり語（抜粋）

ポジティブな極性		ネガティブな極性	
ありがと	ワッショイ	あかん	ヤケ酒
いいぞ	勝ちフラグ	いらいら	不安
うれしい	勝つ	おわた	不調
おめ！	奇跡	つまらん	何やってるんだ
すげぇ	打った	はあ？	勘弁
カッコイイ	最高	イライラ	勝てない
テンション上	決まった	エラー	完敗
ナイス	神	ショック	悪夢
ヤッター	ｲｲﾈ	チンタラ	馬鹿
ヨッシャ	ｷﾀ	ミス	ショボーン

表３．マッチング結果

クラスタ2		
商品タイトル	一致キーワード	ベクトル積
最高峰ハイカット　革底スパイク　ミズノプロHI （コサイン類似度 0.0815）	野球	0.0161
	プロ	0.0145
	職業	0.0024
	団体競技	0.0015
	球技	0.0008
WBC日本代表　オーセンティックユニフォーム （コサイン類似度 0.0683）	野球	0.0144
	日本	0.0071
	楽天	0.0023
	団体競技	0.0013
	球技	0.0007
テレフォンアームスタンド （コサイン類似度 0.0566）	電話	0.0229
	和製漢語	0.0036
	デザイン	0.0012
	職業	0.0007
	なし	0.0000

クラスタ9		
商品タイトル	一致キーワード	ベクトル積
WBC日本代表　オーセンティックユニフォーム （コサイン類似度 0.3002）	日本	0.0243
	選手	0.0204
	野球	0.0092
	ロッテ	0.0082
	監督	0.0034
靖国 YASUKUNI(DVD) （コサイン類似度 0.2229）	日本	0.0312
	監督	0.0042
	東アジア	0.0040
	アジアの国	0.0039
	君主国	0.0034
請福泡盛千葉ロッテマリーンズ公認石垣島キャンプ記念ボトル4合瓶 （コサイン類似度 0.2065）	ロッテ	0.0494
	千葉	0.0096
	ロッテグループ	0.0057
	新宿区の企業	0.0057
	ホテル運営会社	0.0057

5.4. 最適なタイミングの決定

試合中に投稿された 524 ユーザの全つぶやきを用いて推薦に最適なタイミングを決定することを試みた．

5.4.1. バーストの検出

分析用データを用いて時系列におけるつぶやき数の変化度合からバーストを検出することを試みた．バーストの検出には式(1)を用いた．式(1)中の X，Y については予備実験の結果 X=3，Y=30 とした場合に比較的良い結果が得られたため，この値を用いた．

実験結果から閾値 α を 0.2 程度とすることでバーストをうまく検出できることが分かった．そこで本研究では閾値を 0.2 とし，判定値 B が一旦閾値 0.2 を超え，その後再び 0 .2 未満になるまでの区間をバーストとすることにした．この手法を用いることでリアルタイムにバーストの検出が可能になる．

5.4.2. 極性評価に用いる手掛かり語の収集

分析用データ中の試合中のつぶやきをチェックし，バーストの極性評価に用いる手掛かり語を人手で収集した．その結果ポジティブな極性を表す 131 語，ネガティブな極性を表す 176 語を収集した．収集した極性評価用キーワード（手掛かり語）の一部を表 2 に示す．

5.4.3. 極性評価

収集した手掛かり語を用いて実際にバーストの極性評価を行った． P をポジティブな極性を持つ単語数，N をネガティブな極性を持つ単語数，A を P と N の和としたとき P/A が閾値 0.7 を超えた場合そのバーストはポジティブ， N/A が閾値 0.7 を超えた場合そのバーストはネガティブの属性を持っているとして評価実験をした結果，極性評価はほぼ成功した．そこで本研究では以後閾値を 0.7 とした．

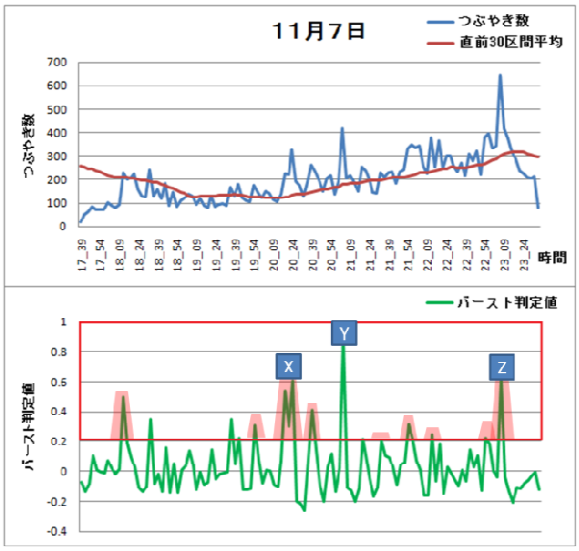

図 1．バースト検出と極性評価の結果

5.4.4. 評価用データを用いての実験

最後に分析用データから求めた手掛かり語，バースト判定式，極性評価式を基に，評価用データを用いてのバースト検出，極性評価の実験を行った．結果を図 1 に示す．図中の赤く塗られた部分がポジティブと評価されたバーストである．図 1 においては実際につぶやき数が急増したタイミングで同様にバースト判定値も高い数値を示しており，バーストの検出は成功していると言える．またバースト区間中の実際のつぶやきデータを分析した結果，極性評価もほぼ成功していると言えた．極性評価結果についての詳しい考察については第 6 章で述べる．

5.5. 商品データとのマッチング

最後に一つの事例として 5.3 で得られた各クラスタのプロファイルと楽天商品データ [10]とのマッチングを行い，推薦対象となる商品情報を導出することを試みた．

まず楽天商品データから 1,000 商品を選択し，各商品についてそのタイトルと商品説明文を用いて 5.2 と同様に名詞とそのカテゴリ分類を抽出し，tf-idf を用いてプロファイルを作成した．

作成した各商品のプロファイルと 5.3 で得られた各クラスタのプロファイルとのマッチングを行い，推薦商品を導出した．評価尺度としてコサイン類似度を用いた．結果の一部を表 3 に示す．

クラスタ9は表1より千葉ロッテマニアクラスタであると言え，表3において千葉ロッテ関連の商品が推薦商品として上位に挙がっているという結果を見るとマッチングは成功していると言える．

クラスタ2に関しては5.3で噺家クラスタと定義したが表3においては「野球」や「プロ」といったプロファイルが重視され推薦商品が導出されている．これは「落語」などにマッチする商品データが存在しなかったためであり，もしこのような商品データが存在すれば推薦商品の上位にランクインする可能性はあったと考える．

6. 考察

5.4.4で行ったバーストの極性評価について考察する．以下，図1中に示したバーストが検出された時間帯であるX，Y，Z中に投稿されたつぶやきを基に分析を行う．

Xは千葉ロッテ3-6のビハインドから今江選手のタイムリーで1点返し，さらにその後満塁になって里崎選手の同点タイムリーが飛び出した 20:18～20:27 の時間帯に発生したバーストである．歓喜と興奮のつぶやきが数多く見られ，ポジティブなイベントが起こったと考えられるため極性評価は成功していると言える．

Yは同点の場面において相手チームのエラーからチャンスが広がり，その後キム・テギュン選手がタイムリーを打ち千葉ロッテが勝ち越しした場面である 21:03～21:06 の時間帯に発生したバーストである．全体的にポジティブなつぶやきが多かったが，「エラー」等のネガティブな極性を持つ手掛かり語が頻出したため全体としてはどちらの極性か判別できない結果となってしまっている．

Z は千葉ロッテが日本シリーズ優勝を決めた 23:06～23:09 の時間帯に発生したバーストである．千葉ロッテ優勝の歓喜に沸くポジティブなツイートが数多く見られ，極性評価は成功していると言える．

全体的に見ると極性評価はうまくいっている印象である．しかしYのような例も見受けられ，改善の余地はあると考える．例えばYのバースト中において「エラー」や「悪送球」等のネガティブな極性を持つ手掛かり語が現れるつぶやき中には「和田」というキーワードが頻出している．「和田」とは相手チームの和田選手のことである．実際にエラーをしたのは和田選手であり，千葉ロッテファンにとってこれはネガティブなイベントではないと言える．これに着目することでYのバースト中の「エラー」や「悪送球」などの手掛かり語は相手チームに対して発せられている語であり，このような場合「和田」等のキーワードと共起するネガティブな極性を持つ手掛かり語は極性評価のための対象として用いる必要が無いことが分かる．その結果Yのバーストはポジティブな極性を持つと評価される．このように，極性を表す手掛かり語がどの対象に対して発せられているのかを考慮することで精度の向上が見込まれると考える．

6．おわりに

本研究はマイクロブログ特有のリアルタイムな投稿を活用し，効果的な情報推薦を行うための手法を提案した．「リツイート」を基にユーザのプロファイリングを行い，セレンディピティのある推薦を実現するためにユーザのクラスタリングを行った．また最適なタイミングとして「バースト」に着目し，それを検出し極性評価を行う手法を示した．最後に実際の商品データと各クラスタのプロファイルとのマッチングを行って推薦商品を導出する例を示し，提案手法の有効性を示した．

しかしリツイートの投稿数は通常のつぶやきと比較するとかなり少なく，分析に十分な量が得られにくいという問題がある．今後ライトユーザに対して手法を適用する場合や収集期間をさらに短くしてプロファイリングを行う場合，他の指標にも着目していく必要があると考える．

参考文献

1) Twitter : Twitter. http://twitter.com/
2) 榊 剛史, 松尾 豊 : ソーシャルブックマークとしての Twitter リスト機能の応用, The 24th Annual Conference of the Japanese Society for Artificial Intelligence (2010).
3) 眞野裕也, 青山俊弘 : ミニブログユーザの記事嗜好を用いたクラスタ発見, Journal of JACT, Vol.15, No.3, pp.43-46 (2010).
4) Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.: Item-based Collaborative Filtering Recommendation Algorithms, Proc. of WWW'01, pp. 285-295, (2001).
5) 吉井和佳, 後藤真孝, 駒谷和範, 尾形哲也, 奥乃博 : ユーザの評価と音響的特徴との確率的統合に基づくハイブリッド型楽曲推薦システム, 情報処理学会研究報告, 音楽情報科学 2006(90), pp45-52 (2006).
6) 八村太輔, 湯本高行, 赤星祐平, 小山 聡, 田中克己 : Web 検索結果のクラスタリングと観点抽出に基づく閲覧インタフェース, DEWS2005, 4B-i11 (2005).
7) 京都大学情報学研究科 : 形態素解析エンジン MeCab (2007). http://mecab.sourceforge.net/
8) 伝 康晴, 山田 篤, 小椋秀樹, 小磯花絵, 小木曽智信 : 形態素解析辞書 UniDic (2010). http://www.tokuteicorpus.jp/dist/
9) 藤澤瑞樹 : クラスタリングツール bayon (2009). http://code.google.com/p/bayon/
10) 楽天技術研究所 : 楽天商品データ. http://rit.rakuten.co.jp/