D-024

マイクロブログからの口コミ情報抽出に関する研究

Extraction of word-of-mouth information from microblogs

友利明央† 青江順一† 森田和宏† 泓田正雄†
Akio Tomotoshi Jun-ichi Aoe Kazuhiro Morita Masao Fuketa

1．はじめに

インターネットや携帯電話は, 社会に広く普及し, 国民の生活, 文化, 経済など社会的インフラ基盤として必要不可欠なものとなってきている．また，スマートフォンが普及したことにより，マイクロブログの利用者は増加傾向にある．本研究では，多数あるマイクロブログの中でも Twitter を用いる．

現在においても Twitter は，情報を気軽に発信できるツールであると考えられる．そのため，情報量は膨大となり必要な時に必要な情報をスムーズに得ることが難しくなってきていると考えられる．また，この情報量を人手で分類をおこなうには，膨大な時間と労力が必要となる．そのため，近年ではツイートにおける自動分類の研究が盛んにおこなわれている[1][2]．しかし，ツイートを見ただけでは, そのツイートが何を目的として発信しているかを自動で判断するのは非常に困難である．例えば，ある事柄・物についての評価なのか，ある商品の宣伝・広告なのか，あるいは出来事をツイートとしただけなのか，このようにツイートする目的は多様に存在する．また，Twitter を用いると比較簡単に自分の意見をツイートできると考えられるので, 利用者の「生の声」が多く存在しているのではないかと考えられる．さらに，Twitter はリアルタイム性や伝播力が優れている[3]．そのため，すぐに「生の声」の情報を得ることができると考えられる．

そこで，本研究では膨大な情報源である Twitter を用いて利用者の「生の声」である口コミを自動で抽出する手法の提案をおこなう．また，自動分類ができれば，人手における労力負担の改善に繋がると考えられる．

2．関連研究

Twitter 上の bot 判別による情報伝達の効率化という研究がある[1]．この研究では，Twitter 上の人間アカウントと bot アカウントを判別することを目的としている．本研究では，口コミつまり人間アカウントによるツイートの抽出が目的なので対象が異なる．しかし，人間アカウントによるツイートと bot アカウントによるツイートを区別している点で少し類似している．

3．提案手法の流れ

Twitter から口コミ情報を抽出するための流れを以下に示す.

Step1．ツイートを取得

TwitterAPI を用いてキーワードを入力し, そのキーワードが含まれるツイートを 100 件取得する.

Step2．ツイートの重複を削除

同じツイートやリツイートによる重複があると処理に時間がかかるため, それらを削除する.

Step3．口コミ情報の抽出

形態素解析，多属性照合，分野連想，感性理解を用いてツイートから口コミ情報を抽出する．抽出方法は 4 節で詳細を述べる．口コミ情報抽出の流れを図 1 に示す.

4．提案手法の流れ

口コミ情報の抽出方法は，5 つの項目から成り立つ．各項目でそれぞれ得点を付与しておき，その合計値によって口コミかどうかの判定をおこなう．合計値が 0 以上のツイートを口コミ，0 未満のツイートをその他と分類した．以下の各項目における得点付けは，事前準備の結果を用いている.

① 形態素数 22 以下：+70

② 動詞(有り)：+30，動詞(無し)：-30

③ 分野(有り)：+20，分野(無し)：-20

④ 感性(有り)：+10，感性(無し)：-10

⑤ パターン(表 1 を参照)

①～④までの項目は上記に示している通りに得点付けをおこなう．⑤に関しては，ある単語や記号の出現パターンをルールに登録しておき，ツイートに含まれている場合は，そのパターンに付与している点数に応じて加点，減点をおこなう．パターンの例を表 1 に示している.

表 1．パターン毎の得点例

パターン	得点
笑、w	+5
♪	+10
ニュース	-70
http(2 回以上)	-65
NAVER まとめ	-50

5．実験

口コミ情報の抽出精度を確認するために実験をおこなった.

†徳島大学，The University of Tokushima

5.1 実験設定

実験設定として，収集したツイートを用いてクローズデータ 1（100 件），クローズデータ 2（100 件）を用意した．また，新たに収集したツイート 100 件をオープンデータとした．評価方法として，適合率，再現率を用いた．以下に式の詳細を記述する．

$$適合率(\%)=\frac{正しくロコミ情報として抽出した数}{ロコミ情報として抽出した数}\times 100$$

$$再現率(\%)=\frac{正しくロコミ情報として抽出した数}{ロコミ情報として抽出すべき数}\times 100$$

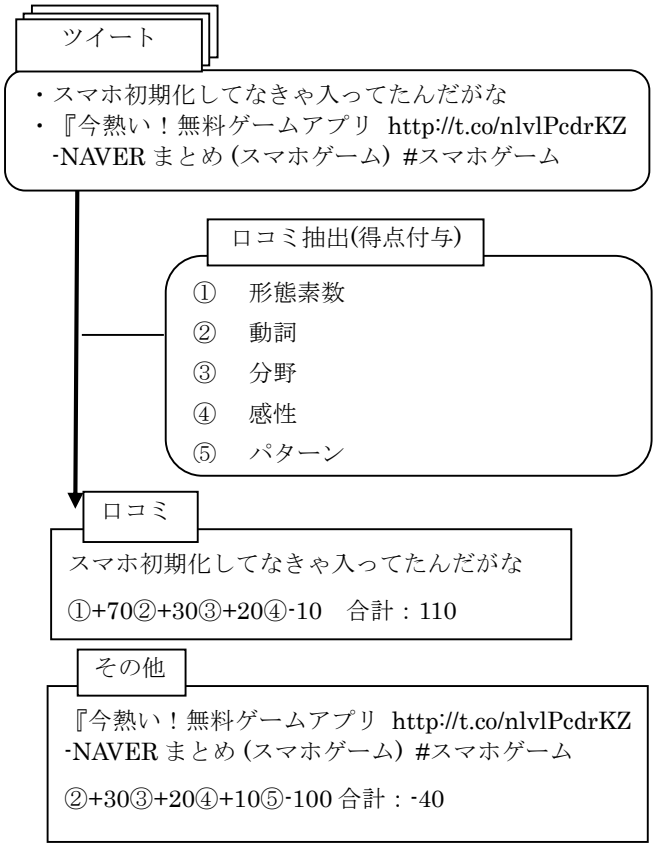

図 1 ロコミ情報抽出の流れ

5.2 実験結果

ロコミ情報抽出のクローズデータとオープンデータの実験結果を表 2 に示す．

ロコミ情報の抽出結果より,クローズデータ 1，クローズデータ 2 はどちらもこのデータをもとにルール，パターンの作成をおこなった．そのため，どちらも適合率，再現率 90%を超えたので良好な結果が得られた．オープンデータに関しては，再現率が 90%を超えて，適合率も 80%を超えているため，良好な結果が得られたといえる．また、類似研究として，対象は異なるが、新聞記事を対象とした研究[4]では 85%，Web 文書を対象とした研究[5]では 87%の精度を得られている．そのため，マイクロブログを使用した本研究も同等な結果を得られているので，良好な結果といえる．しかし，抽出漏れや誤抽出が存在しているので，次節で各項目における考察を述べる．

表 2．実験結果

	適合率	再現率
クローズデータ 1	95%	97%
クローズデータ 2	93%	92%
オープンデータ	85%	92%

5.3 考察

ロコミ情報の抽出漏れの例としては，ツイートに URL が含まれてしまったために，文章が短い(形態素数 22 以下)と判断されなかった．URL が写真のものか，そうでなかで判断することで改善できると考えられる．また，ロコミツイートにおいて,⑤のパターンによる得点付与でマイナス値が大きくなってしまったため抽出漏れがおきた．この例に関しては，ルールの見直し，パターンの得点調整をおこなうことで改善できると考えられる．

その他の情報をロコミ情報として誤って抽出した例も存在した．その例のほとんどが，広告・宣伝をしているツイートであった．これらのツイートは，Twitter の文字数制限である 140 字におさまるように，固くない表現を使用し，わかりやすく書かれている．また，語尾に”!”や“♪”などの記号を使用しているツイートもいくつか存在した．この問題点を改善する方法としては，ルールの拡充もしくは新たな手法の考案が必要である．

6. まとめと今後の課題

本稿では,Twitterから得られたツイートにおいて①～⑤までの項目における得点付けを用いて,ロコミ情報を抽出するための手法について説明を述べた．そして,ロコミ情報の抽出精度を確認するために実験をおこなった．その結果,抽出漏れや誤抽出の改善としてルールの拡充をおこなう必要がある．

今後は,考察で述べた問題点の改善をおこない，精度の向上を計る．そして，誤抽出をなくすような新たな手法の考案をおこない，これらの問題点に取り組んでいきたい．

参考文献

[1] 湯田雅，矢吹太郎，佐久田博：
Twitter 上の bot 判別による情報伝達の効率化
[2]渡邉研斗，鍋島啓太，岡崎直観，乾健太郎：
Twitter 上での誤情報と訂正情報の自動分類
[3] 今さら聞けない Twitter、Facebook、Line の 3 大ソーシャルの使い分け
http://news.livedoor.com/article/detail/8061216
[4]立石健二，石黒義英，福島俊一
インターネットからの評判情報検索
[5]立石健二，森永聡，山西健司，福島俊一
Web 上の自動意見分析-情報抽出とテキストマイニングの融合-