主観表現と客観表現を用いたTwitterにおける有益なツイートの推定

小池 達也† 高木 友博†

†明治大学大学院理工学研究科基礎理工学専攻 〒214-8571 神奈川県川崎市多摩区東三田 1-1-1

E-mail: †{koike, takagi}@cs.meiji.ac.jp

あらまし 情報ネットワークにはリアルタイムな情報が膨大に溢れている．Twitterではユーザが呟き機能によって情報を発信することで情報の共有が行えるが，有益でない呟きの情報も多く共有される．そこで，有益な呟き情報以外を自動的に除去する必要があると考えられる．従来は機械学習を行う研究が多くされているが，学習はドメインに依存してしまう問題点がある．そこで我々は，呟きに含まれる主観表現と客観表現について解析することでドメインに依存せず即時に有益な情報度合いを計ることが可能であることを提示する．有益な呟きはリツイートされる前提で評価実験を行い，その有効性を示す．

キーワード 情報フィルタリング, Twitter

1. はじめに

ウェブ上には様々な種類の膨大なデータが溢れている．特にビッグデータに対して情報推薦また情報フィルタリングを行い情報提示するデータをカスタマイズする研究が盛んに行われている．近年はTwitter[1]のような情報ネットワークまたマイクロブログサービスやソーシャルネットワークの進歩によりリアルタイムな情報でかつ呟きのような短文なテキストのデータが増えるようになった．イベントなどの流行時では類似した情報で情報爆発またのバーストが起こり，バースト検出することも可能になった．一方で流行の意図を把握してないユーザに対して多量の類似した短文なテキストを提示してしまう問題点が挙げられる．またこのような短文なテキストである呟き情報(以後，ツイート)は他者に読んでもらう期待度が通常のウェブページやブログなどと比較して低い．このような呟き情報が共有された際に有益な情報度合いは一部の読者によるお気に入りなどのユーザ評価が行われた後，評価が高いツイートをまとめるサービスが多い．しかしながら多くの読者がいるユーザの呟き情報が多く評価されることにより有益な情報が偏っていると考えられる．また一部の読者による評価がなければ呟き情報の有益な情報度合いが計れない問題点が考えられる．

本稿では前述した問題点を解決させるために，主観表現と客観表現を用いたTwitterにおける有益なツイートの推定方法を提案する．まず2章で本研究の関連研究を本目的との位置付けを示しながら説明する．3章で本目的を達成するための提案手法について分析結果と共に示す．4章では提案手法に対する評価実験を示す．最後に5章で本論文を締めくくる．

2. 関連研究

本章では5つ関連研究を示す．有益な情報にするために，スパム情報やネガティブ情報をフィルタリングすることやツイートの文字数が多いほど価値が高いことを素性とするなどで機械学習を行う手法がある．他にも主観表現と客観表現を利用することでユーザの目的意識に沿った情報にすることができる．これらについて一つずつ説明する．

2.1 スパム情報フィルタリング

電子メールのスパム情報フィルタリングに関する研究が多く行われている．一方でTwitterにおけるスパムユーザ判定に関する研究[2]がある．機械学習の手法であるサポートベクターマシン(以後，SVM) を用いてスパムユーザ判定を行っている．問題点としてスパムユーザ判定精度の向上にはジャンルに依存した学習が必要なこと，また情報がスパムであるというスパムツイート判定はスパムユーザ判定より難しいと示されている．特にスパムツイート判定はツイートの文字数が少ない短文なテキストであるために，ノンスパムツイートに対して過度にスパムツイートとして分類してしまう問題点が挙げられている．

2.2 ネガティブ情報フィルタリング

感情極性により情報フィルタリングする方法がある．感情極性とはその語が一般に良い印象を持つ

か(ポジティブ), 悪い印象を持つか(ネガティブ)を表し，単語感情極性対応表[3]にはこの感情極性が示されている．この感情極性はレビュー分類などの研究などで活用されている．本研究ではネガティブな情報をフィルタリングすることで有益な情報を提示できると考えられるが，ネガティブではない情報が有益とは限らないという問題点が挙げられる．

2.3 ツイートの文字数によるリツイート傾向

Twitter はユーザが呟く文章を入力し配信することができる．これを呟きまたツイート(tweet)と呼び，このツイートを閲覧したいユーザは発信ユーザをフォロー(follow)つまり選択しておくことで文章が配信される．配信されるユーザのことをフォローワー(follower)と呼ぶ．ユーザはツイートだけでなく，自身がフォローしているユーザから配信されたツイートを自身のフォローワーに配信するリツイートという機能がある．リツイートは他ユーザが配信する文章をフォローワーに情報拡散させる価値があるツイートでされることが多い．

様々な観点からツイートを分析している研究[4]があるが，ここではツイートの文字数とリツイートの関係を分析[5]している研究を挙げる．この研究ではツイートの文字数が多いほどリツイートされると言及している．文字数を考慮することでリツイートされるツイートが推定できると考えられるため本研究ではツイートの文字数に着目して有益なツイートの推定をすることを考えられるが，ツイートの文字数が多いことは情報が有益であると断定できないという問題点がある．

2.4 機械学習によるリツイート推定

機械学習を用いてリツイート推定を行う研究[6]がある．機械学習では素性選択が重要であり，考えられる素性が一覧[7]で示されている．ここでは本稿で取り上げる主観表現と客観表現については言及していない．機械学習は正解に対して最適化が行われるためにデータ依存になることが多く大域的に性質を示せないことが多いとされる．様々な素性で機械学習を行ったがトピックが変わると精度が落ちる[8]という考察もされている研究がある．本研究では有益な情報であるという正解設定をリツイートとし機械学習によるリツイート推定を行うとツイートの有益度合いは明確化できない．またリツイート推定することは本研究の意義ではなく有益なツイートの推定とは異なる．

2.5 主観表現と客観表現の判定

主観表現に着目することで要約語を生成する際に，形容詞,形容動詞,動詞は主観表現であるとして分析に用いる研究[9]やニュース発信者を分析するために主観的記述と客観的記述であるかを辞書と構文解析を用いた研究[10]，また主観情報が主体のウェブページか客観情報が主体のウェブページかを分類するために文末表現を利用した SVM による機械学習を行う研究[11]がある．要約語の生成，ニュースの分析，主観表現また客観表現の主体判定いずれも本研究の目的とは異なる．

3. 提案手法

本章ではまず有益なツイートの推定に対するアプローチ法として主観表現と客観表現を用いることを定義と共に示す．次に主観表現と客観表現に対する分析を示したのちに，分析から考えられる提案手法について述べる．

3.1 有益の定義と判定方法

本論文では事実と主観を含むことを有益であるとする．本意は客観的事実情報と主観的評価情報のいずれも重要であると考えられる．例えば「iPhone6 発売だって。大きくて使いにくそう。」というツイートであれば「発売」が客観的事実情報であり「使いにくそう」が主観的評価情報である．この何れかが欠落していると情報の価値が下がり有益ではないと考えられる．なぜなら読者が欠落部分を推測することが必要で，難しいからである．

事実と主観を含むかを判定するためにまず *Sen* を用いて形態素に分割したのち基本形を抽出した．次に日本語評価極性辞書の名詞編[12]と用言編[13]のいずれも用いて単一形態素，連続する複数形態素が辞書に含まれるか判定した．事実は客観表現であり客観語の辞書への出現，主観は主観表現の出現つまり主観語の辞書への出現である．

3.2 主観表現と客観表現に対する分析

これより主観表現と客観表現が含まれることが有益な情報と関連があるか分析し定量評価を行う．2.4 節でも触れたようにツイートがリツイートされることを有益な情報であると仮定し以後の分析結果を示す．

対象としたデータは Twitter のツイートとした．twitter4j[14]ライブラリを用いてサンプリングツイートを取得し，以下 5 項目の何れかに該当するツイートを除外したデータセットを作成した．データセットの詳細について，ツイートのサンプリング日，ツイート数，リツイート数とその割合は表 1 の通りである．

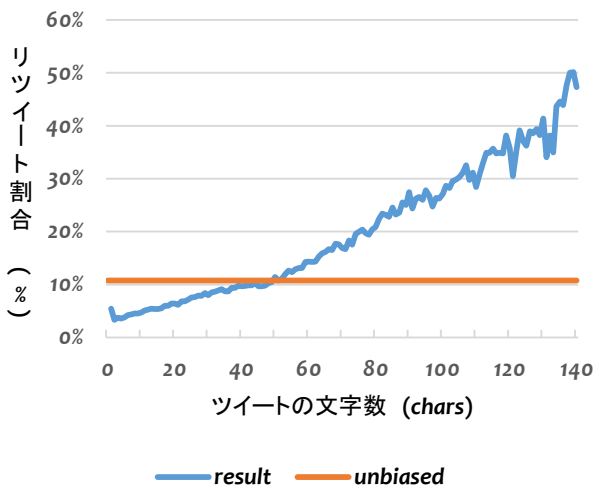
a) 日本語(平仮名，片仮名，漢字)を含まない
b) 言語設定が「*ja*」ではない
c) リンク情報や画像が含まれると考えられる文字列「*http*」を含む
d) 特定ユーザに返信を意味するリプライで使用する文字「@」を含む
e) リツイートされたツイートが重複しないよう最新のリツイート以外のツイート

まず初めに，ツイートの文字数とリツイートの関係を示す．2.3 節ではツイートの文字数が多いほどリツイートされると主張する関連研究を示した．ここでは日本語に言及した場合でもツイートの文字数が多いほどリツイートされるか確認した．さらに「リンクを含むツイートは，リツイートされる確率が 86%高い[15]」という結果からリンク情報を含むと文字数が長くなると懸念されるためにリンク情報が含まれない状態で分析を行った．図 1 が結果である．グラフにおける橙色はツイートの文字数に対するリツイートされたツイートの割合がデータセットに対して偏りない場合の値であり表 1 のとおり約 11%となることが予想できる．グラフにおける水色は結果であり，例えば約 120 文字のツイートに着目するとリツイートされたツイートは約 30%含まれることが確認できる．この結果から日本語に限定またリンク情報を含まないツイートでもツイートの文字数が多いほどリツイートされると言える．

表 1：対象とした Twitter ツイートデータの概要

sampling date	*2014/10/25-29*
tweet	*1,235,281 tweets*
retweet (rate)	*133,171 tweets (10.8%)*

図1：ツイートの文字数に対するリツイート割合



次に，主観語と客観語(以後，どちらかの語に言及しない場合は対象語と呼ぶ)の出現数を調査した．リツイートされていない通常のツイート(Standard)とリツイートされたツイート(Retweet)に分割し対象語の出現数をそれぞれ数えた．また対象語の合計出現語数を分割したツイート数で除算することにより 1 ツイートにおける平均出現対象語数を求め比較する．結果は表 2 に示す．主観語(Subjectivity)と客観語(Objectivity)いずれもリツイートされたツイートはリツイートされていない通常のツイートより語の出現語数が多くなる．

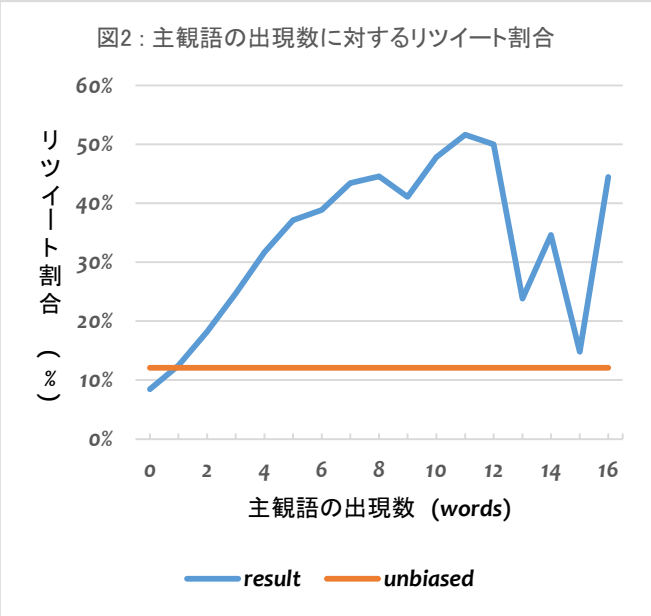
その後，主観語が 1 語以上出現すると主観が含まれている，また客観語が 1 語以上出現すると事実が含まれていることを確認するために，対象語が 1 語でも出現するツイート割合を調査した．通常のツイートとリツイートされたツイートに分割し対象語が 1 語でも出現するツイートを数えた．分割したツイート数で除算することにより，割合を求め比較する．結果を表 3 に示す．リツイートされたツイートはリツイートされていない通常のツイートより対象語を含むツイートが多くなる．

表 2：主観語と客観語の出現数と 1 ツイート平均

	Standard	Retweet
Subjectivity	499,890 words (0.45 words/tweet)	118,691 words (0.89 words/tweet)
Objectivity	655,470 words (0.59 words/tweet)	174,393 words (1.31 words/tweet)

表 3：主観語と客観語の出現ツイート数と割合

	Standard	Retweet
Subjectivity	337,974 tweets (37%)	62,552 tweets (47%)
Objectivity	410,425 tweets (37%)	76,281 tweets (57%)



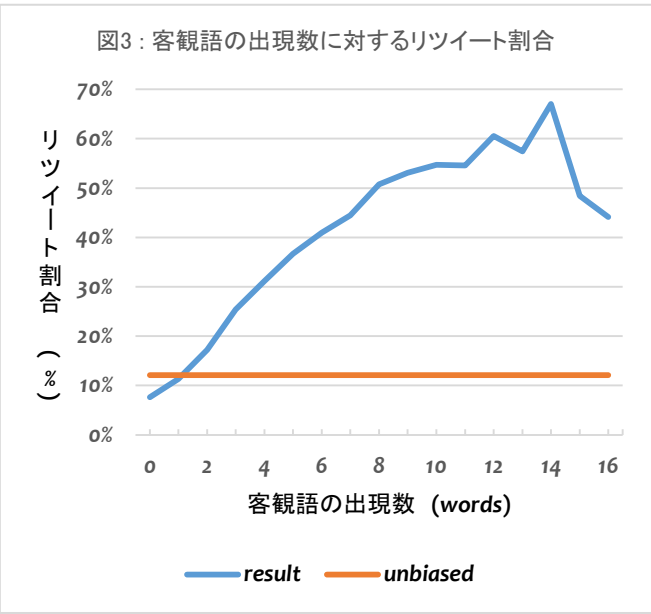


表 4：提案手法のパラメータ設定値

property	value	property	value
subRate	0.25	subA	5.6e-03
		subB	1.2e-02
obRate	0.32	obA	4.5e-03
		obB	1.2e-02
lenRate	0.43	lenA	1.4e-04
		lenB	-1.0e-02

リツイートされたツイートはリツイートされていない通常のツイートに比べて主観語や客観語が出現しやすいことが判明した．これより対象語の出現数におけるリツイート割合を調査した．対象語が何語出現する場合においてリツイートされているツイートが占める割合を確認する．主観語の結果を図 2 に，客観語の結果を図 3 に示す．何れの結果も対象語の出現数が増えるにつれリツイートされているツイートの割合が多くなっていることが，ほぼ確認できる．何れも対象語が 15 語出現するツイートにおけるリツイート割合は，対象語が 14 語出現するツイートにおけるリツイート割合より低いが多くの対象語が出現するツイート数の少なさによる分析の影響であると考察した．

3.3 提案手法

主観語と客観語を多く含んでいるツイートが有益である可能性が高いことが分析結果より判明したため，対象語の出現数を用いた対数近似式で有益度スコア算出をすることが可能であることを提案する．またツイートの文字数が多いほどリツイートされる確率が上がることからツイートの文字数に対する線形近似式を利用することでリツイートされる確率の向上を試みた．主観語の出現数を $subCnt$，客観語の出現数を $obCnt$，ツイートの文字数を $length$ としたとき提案するスコア付け方法は式 1~4 である．パラメータ設定は表 4 のとおりである．各ツイートのスコアを算出し，スコア値が高いツイートをランキング上位から提示する．

$$subScore(subCnt) = subRate \cdot \{subA \cdot ln(subCnt) + subB\} \quad \dots(1)$$

$$obScore(obCnt) = obRate \cdot \{obA \cdot ln(obCnt) + obB\} \quad \dots(2)$$

$$lenScore(length) = lenRate \cdot (lenA \cdot length + lenB) \quad \dots(3)$$

$$Score(subScore, obScore, lenScore) = subScore + obScore + lenScore \quad \dots(4)$$

4. 評価実験

本章では提案手法による有益な情報度合いのスコア付けが妥当であるか確認するために評価実験を行った．スコアが高いランキング上位に有益なツイートが出現するかについて 3.2 節でも示したようにツイートがリツイートされることを有益とし同じデータセットを用いた．

4.1 評価指標

採用した評価指標は次に示す 2 つであり情報検索タスクにおける一般的な評価尺度とした．

a) ***Precision at k*** (以後，***P@k***)
b) ***normalized Discounted Cumulated Gain*** (以後，***nDCG***) [16]

a) はランキング ***k*** 番目までの精度(リツイートされたツイートである)である．ランキング上位にリツイートされたツイートが多く出現しているかを確認する．

b) は推定閲覧数に対するリツイート数を利用し評価を行う．評価 a) において多くのフォローワーが存在するユーザのツイートはリツイートされる確率が高いことからフォローワーの数を推定閲覧数としリツイート数から除算することでリツイート確率を算出する．リツイート確率が高いツイートが有益であるとしランキング上位 ***k*** 番目 (***threshold k***)までの評価(以後，***nDCG@k***)を行う．評価式は式 5,6 である．関連度(rel)の設定については$retweetCnt$をツイートのリツイート数，$followerCnt$ をツイートが配信したユーザのフォローワー数としたとき式 7 である．リツイート数をフォローワー数で除算することにより疑似的なリツイート割合が算出できる．フォローワー数に 1 を加算する意図はフォローワー数が 0 であってもツイートを検索することにより閲覧が可能であるためリツイートされた際のスコア計算処理におけるゼロ除算をさけるためである．$inverseRank(v)$ はv を昇順ソートしたときのランクである．

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(k+1)} \quad \dots(5)$$

$$nDCG@k = \frac{DCG_k}{argmax(DCG_k)} \quad \dots(6)$$

$$rel = \begin{cases} 0 & (retweetCnt = 0) \\ 1.0e-06 \cdot inverseRank\left(\frac{retweetCnt}{followerCnt+1}\right) & else \end{cases} \quad \dots(7)$$

4.2 比較手法

提案手法が有効であるか示すために，比較手法を設定する．

a) ***word***
主観語と客観語の出現数のみを用いた手法である．提案手法におけるツイートの文字数を考慮した計算を無視する．これより対象語の出現数という単純な情報のみでの性能を確かめる．3.3 節の式 4 における$lengthScore$ を 0 としてスコア付けを行う．

b) ***length***
ツイートの文字数のみを用いた手法．関連研究から考えられるベースライン手法である．ツイートの文字数[1,140]をスコアとする．

c) ***follower***
フォローワー数のみを用いた手法．フォローワー数が多いユーザのツイートほどリツイートされる確率が高くなると考えられるために比較手法として設定した．フォローワー数をスコアとする．

d) ***random***
一様分布の疑似乱数[0,1)を用いた手法．

4.3 評価実験結果

表 5 に各手法の ***P@k*** を示す．上位 ***k*** 件の閾値を{1, 10, 100, 1000, 10000, 1%}とした．***random*** 手法を除き，上位 ***k*** 件における結果値が最高な値が得られた手法を橙色，その次を黄緑色，最低な値が得られた手法を水色で表に彩色した．まず提案手法はベース

ラインの *length* 手法より優れていることが *P@1%* における精度を比較することで分かる(表における赤字). しかしながら *P@k* においては *follower* 手法が最も優れる. 4.1 節また 4.2 節でも示したようにフォローワー数が多ければツイートの閲覧数が増えるため, リツイート数 1 件は獲得しやすく結果にも顕著に表れている.

次に各手法の *nDCG@k* を表 6 に示す. 表 5 と同様の尺度で表に彩色した. *P@k* で問題とされた *follower* 手法が最も結果が悪いことが示されている. フォローワー数が多いとツイートの拡散力は大きいがフォローワー数に対するリツイートユーザの割合は多くないことが分かる. 提案手法はフォローワー数が多いことが保証されてないのにも関わらず有益であると考えられるリツイートがフォローワーに対して多く行われている. また単純に対象語だけを用いた *word* 手法に加え, ツイートの文字数を考慮した提案手法が有効であることを *nDCG@10000*, *nDCG@1%* からも言える.

よって *P@k* による評価結果から提案手法は他手法に比べリツイートされたツイートを上位 1%にランキングすることができた. *P@k* における *follower* 手法の最良である問題点を解決させるために *nDCG@k* による評価を行い, 結果として提案手法がフォローワー数の多くないユーザによるツイートに対してもフォローワーの多くが有益であるとリツイートしているツイートを上位 1%にランキングすることが可能であると考えられる.

表 5: 各手法の *P@k* 比較

top k	proposal	word	length	follower	random
1	0	0	1	1	0
10	0.9	0.9	0.5	1.0	0.1
100	0.66	0.68	0.49	1.00	0.10
1000	0.596	0.576	0.469	0.988	0.116
10000	0.5055	0.4473	0.4750	0.9137	0.1123
1%	0.4985	0.4339	0.4747	0.9040	0.1114

表 6: 各手法の *nDCG@k* 比較

top k	proposal	word	length	follower	random
1	0.0000	0.0000	0.0326	0.1474	0.0000
10	0.2781	0.2761	0.2292	0.0803	0.0000
100	0.3251	0.3304	0.2354	0.1097	0.0309
1000	0.3195	0.3113	0.2586	0.1023	0.0474
10000	0.2820	0.2509	0.2582	0.1774	0.0533
1%	0.2834	0.2465	0.2619	0.1985	0.0537

5. おわりに

本稿では事実と主観を含むことを有益と定義し主観表現と客観表現を用いた Twitter における有益なツイートの推定を主観語と客観語を用いて行う手法を提案した. 定量評価をフォローワーによるリツイートによって評価実験を行った. 主観語と客観語を用いることで有益であろうツイートを上位にランキングさせることができた. またツイートの文字数に対するリツイート割合の分析を行ったところ. 日本語に限定した上, リンク情報が存在しない場合でもツイートの文字数が多いツイートほどリツイートされることが確認できた.

今後の課題として文末表現を利用したスコア付けによりさらなる精度向上またツイート内容の定性評価などを行い類似ツイートの除去をし, 多様な情報へカスタマイズする手法の検討, また個々のユーザを考慮した有益な情報の提示手法について検討することでより良い有益な情報の提示を行いたいと考えている.

参 考 文 献

[1] Twitter, https://twitter.com/ .

[2] F.Benevenuto, G.Magno, T.Rodrigues and V.Almeida, “Detecting Spammers on Twitter”, Proc. of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), 2010.

[3] 高村大也, 乾考司, 奥村学, “スピンモデルによる単語の感情極性抽出”, 情報処理学会論文誌ジャーナル, Vol.47 No.02 pp.627-637, 2006. http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html

[4] creativi.tea - ツイッター分析のシリーズの目次, http://teapipin.blog10.fc2.com/blog-entry-298.html

[5] 山名哲也, “Twitter における情報拡散の特徴分析”, http://www.nadasemi.ii.konan-u.ac.jp/publication/research/2012/final/yamana.pdf, 2012.

[6] M.Jenders, G.Kasneci, and F.Naumann, “Analyzing and Predicting Viral Tweets”, International World Wide Web Conference (WWW), 2013.

[7] C.Castillo, M.Mendoza, B.Poblete, “Information Credibility on Twitter”, International World Wide Web Conference (WWW), 2011.

[8] A.Finn, N.Kushmerick, B.Smyth, “Genre Classification and Domain Transfer for Information Filtering”, Proc. of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval, pp.353-362, 2002.

[9] 池田和史, 小松恭子, 服部元, 松本一則, 滝嶋康弘, “キュレーションサービスのための主観的コメントの要約手法”, データ工学と情報マネジメントに関するフォーラム(DEIM), 2014.

[10] 石田晋, 馬強, 吉川正俊, “記事の主観性を考慮したニュース発信者の特徴分析とその応用”, データ工学と情報マネジメントに関するフォーラム(DEIM), 2010.

[11] 松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，“文末表現を利用したウェブページの主観・客観度の判定”，データ工学と情報マネジメントに関するフォーラム(DEIM)，2009.

[12] 東山昌彦，乾健太郎，松本裕治，“述語の選択選好性に着目した名詞評価極性の獲得”，言語処理学会第 14 回年次大会論文集，pp.584-587，2008.

[13] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一，“意見抽出のための評価表現の収集”，自然言語処理，Vol.12 No.3 pp.203-222，2005.

[14] twitter4j，http://twitter4j.org/ja/

[15] SEO Japan - Twitter に関する意外過ぎる 10 大統計，http://www.seojapan.com/blog/twitter-10-stats, 2013.

[16] K.Jarvelin and J.Kekalainen, “Cumulated gain-based evaluation of IR techniques”, ACM Transactions on Information Systems (TOIS), Vol.20 pp.422-446, 2002.