マイクロブログにおける新しいキーワードの推薦

岩井　一晃† 　鈴木　 優†† 　石川　佳治††,†††

† 名古屋大学工学部電気電子情報工学科情報コース
†† 名古屋大学情報連携基盤センター
††† 国立情報学研究所
E-mail: †{iwai,suzuki}@db.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

あらまし　本稿では，マイクロブログにおける利用者によって入力されたキーワードで得られた投稿文から新しいキーワードを推薦する．現在，マイクロブログは多くの人に利用され、多種多様な投稿で溢れている．ところが，自分の閲覧したい特定の分野についての投稿文を全て網羅することは不可能である．これは，投稿文の多いことや，マイクロブログのリアルタイム性により，利用者の知識外の文字列が重要なキーワードとなる可能性が高いからである．そこで本研究では，利用者の入力したキーワードに関連性が高いと考えられる新しいキーワードを推薦することで検索支援を行う．利用者の入力したキーワードを利用し，取得した投稿文を n-gram を用いて，頻出文字列を取得する．解析した頻出文字列より新しいキーワードを発見する．新しいキーワードの精度を上げるために発言者の情報を解析し，関連語として正しいかどうかの判別を行う．本論文では，提案した手法の説明を行う．また，本手法と他の頻出文字列を取得する方法として考えられる形態素解析を用いた手法との比較実験を再現率精度曲線を用いて行い，本手法の考察を行った．

キーワード　マイクロブログ,Twitter,n-gram, 関連語

Keywords Recommendation for Microblogs

Kazuaki IWAI†, Yu SUZUKI††, and Yoshiharu ISHIKAWA††,†††

† Undergraduate School of Information Science, Nagoya University
†† Information Technology Center, Nagoya University
††† National Institute on Information
E-mail: †{iwai,suzuki}@db.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

1. はじめに

現在、Twitter (注1)に代表されるマイクロブログは多くの利用者によって利用されている．Twitter に関しては 2011 年 1 月現在，おおよそ 1800 万人の利用者(注2)が日本国内に存在している．マイクロブログの利点として，身近な人とのコミュニケーションツールとなる点，日常の生活では知り合えない人の発見，交流ができるという点が挙げられる．大量の利用者の中から実世界上の知り合いでなく，利用者が興味がある人物を探す方法として，利用者が投稿文を検索する方法がある．公式サイトの検索フォームに自らの興味がある物事のキーワードを入力すると入力したキーワードを含む投稿文が取得できる．利用者がスポーツに興味があり，TV での中継放送中にそれらに関するキーワードを公式サイトを用いて検索したとする．その場合，中継に関する大量の投稿文が得られる．また投稿文中に利用者の知識外，予測できない文字列が出現する．その文字列が，利用者の検索する分野における重要なキーワードであった場合，利用者はその文字列に関する話題を閲覧することができない．このように，利用者が閲覧した分野についての投稿文を網羅することは難しい．

この問題を解決するために，本研究では利用者の入力したキーワードにより投稿文を取得し，利用者の入力したキーワードに関連性の高いと考えられる文字列を新しいキーワードとして推薦するシステムを提案する．新しいキーワードを推薦する方法において，利用者が入力したキーワードにより取得した投稿文における頻出文字列を新しいキーワードとする方法が考えられる．この方法において頻出文字列を抽出する際，方法の一つとして形態素解析が利用が考えられる．しかしマイクロブログに投稿される文はくだけた表現になる傾向がある．投稿文のくだけた表現は形態素解析で利用される辞書に登録されていない単語である可能性が高い．よってマイクロブログ特有のくだけた文は形態素解析では正しく解析できない．これより，形態

(注1)：http://twitter.com/
(注2)：Google 提供の DoubleClick Ad Plannner の調査による (http://www.google.com/adplanner/)

素解析を用いた手法によって推薦される新しいキーワードは良い精度が得られないと考えられる．

そこで，本研究ではキーワード抽出の支援を n-gram と発言者の情報を用いた関連語判別を用いて行う．n-gram を用いて頻出文字列を取得することにより，形態素解析で判別することのできなかった文字列を取得することができる．n-gram は文字列を文字の並びだけで認識している．そのため n-gram によって出現回数を算出した場合，形態素解析を利用して出現頻度を算出する場合よりも誤判断が発生する可能性が低いと考えられる．また n-gram によって投稿文より，頻出文字列を取得した場合，頻出文字列の部分文字列も取得してしまうため，これらを削除する．また発言者の情報を用いて，関連語の候補が関連語かどうかを判別する．発言者の情報を用いることで，少数の発言者が何度も投稿している文字列を取り除くことができる．これによって，高い精度で新しいキーワードを推薦できると考えられる．

2. 関連語の抽出

新しいキーワードを抽出する際，利用者が入力したキーワードにより検索されたパブリックストリーム上の投稿文を利用し，頻出文字列を発見する．取得される投稿文は利用者の入力したキーワードが含まれた投稿文である．利用者の入力したキーワードに共起して出てくる単語は出現頻度が高ければ高いほど，利用者の入力したキーワードと関連していると考えられる．しかし出現頻度だけで考えると文字列の文字数が少ないものが多く出現する傾向があると考え，関連語の候補として利用する出現頻度は単純な出現回数でなく，単純な出現頻度をその文字列の長さで検出された全ての文字列の出現回数の総計で割ったものを用いる．以下ではこれを出現割合と呼ぶこととする．

本研究では n-gram を用いて文字列の出現回数を算出する．マイクロブログの投稿文は通常の文章と異なり，くだけた文になりがちである．このため，文字列を予め用意した辞書を用いて判断する形態素解析よりも，同じ文字の並びをするものを同じ文字列として考える n-gram を用いた手法のほうが向いていると考えたからである．n-gram を利用して頻出文字列を抽出しようとする際，ある文字列の部分文字列が出現する．ある文字列の部分文字列であり，ある文字列の部分文字列以外に出現していなければそれは不必要なものと考えられるため，そのような部分文字列の削除を行う．また，単語の頻出頻度だけでなく，発言者の情報を用いて単語が関連語かどうかの判別を行う．発言者情報を用いるのは，例えば少数の人数がキーワードと関連しない分野であるが、そのキーワードを含む投稿文を多くしたと仮定する．その場合関連語ではないが，頻出文字列として見た場合，関連しない分野の頻出文字列も関連語となる可能性がある．これを防ぐために，少人数が多く発言している頻出文字列を取り除くことしなければいけない．そこで発言者情報を用いて，そのような単語を関連語としないということが必要となってくるからである．

本研究のシステムの流れは図 1 のようになっており，本章ではそのシステム中の以下のそれぞれのシステムについて述べる．

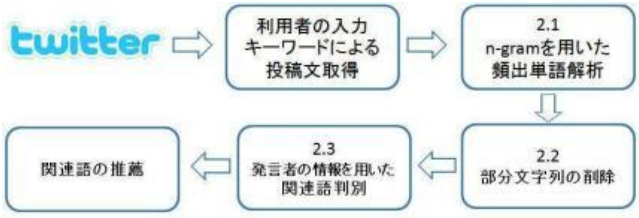

図 1　本研究のシステムの流れ

- **n-gram を用いた頻出文字列解析について**

利用者によって入力されたキーワードから得た投稿文に n-gram を利用し，頻出文字列を探し出す．2.1 節においてこの機能について述べる．

- **部分文字列の削除について**

頻出文字列の中から部分文字列となっているものを探し，関連語の候補から削除する．2.2 節においてこの機能について述べる．

- **発言者の情報を用いた関連語判別について**

発言者の情報を用いて関連語かどうかを判別し，関連語の精度をあげる 2.3 節においてこの機能について述べる．

2.1 n-gram を用いた頻出文字列解析

n-gram を用いた頻出文字列解析を行う．n-gram は対象の文書集合中に n の長さのある文字列が何度出現したかを調べる手法である．この n-gram を用いることで文字列の出現回数を数え，出現回数の多いものを頻出文字列であると判別する．

n-gram を行う対象文章群は利用者の入力したキーワードを含む投稿文である．この投稿文に n-gram を用いた頻出文字列解析を行い，頻出文字列と判別される文字列は利用者の入力したキーワードに共起する文字列であるといえる．利用者の入力したキーワードに多くの頻度で共起する単語は，利用者の入力したキーワードに関連する文字列であるともいえる．これより，頻出文字列は関連語の候補となる．

形態素解析でも n-gram を用いた手法と同様に頻出文字列の解析を行うことが可能である．しかし，形態素解析を行う際，事前に辞書を登録する必要がある．この場合，マイクロブログ特有のハッシュタグやくだけた表現といった辞書に登録されていない文字列に対応することができない．ところが形態素解析で対応できない文字列が新しいキーワードとなる可能性がある．このように辞書に登録されていない単語が関連語として検出されない問題を解決するために本研究では n-gram を用いた頻出文字列解析を行う．

またこの時，2.3 節の発言者の情報を用いた関連語判別で発言者の情報も利用するため，投稿文は発言者毎に処理を行う．前処理として同じ発言者の投稿文は一つの投稿文として扱えるようにする．n-gram を用いた頻出文字列解析とする際，正規表現を用い，不要と考えられる文字列は n-gram を適用しないこととする．具体的にはマイクロブログ特有の表現であり，意味をなさないと考えられる“ RT ”という文字列，数字と記号だけで成り立っている文字列などを文字列として扱わないこととする．これらを除いた文字列の数をそれぞれ n=2,3...12 の範囲

の頻出文字列解析を行い，文字列の出現頻度および発言者の情報を保存する．

2.2 部分文字列の削除

n-gram を用いた頻出文字列解析により，部分文字列が頻出文字列として検出される．ところがそれだけで文字列として意味をなさない部分文字列は検出する必要の無い頻出文字列である．それはある文字列の部分文字列であった場合，部分文字列もある文字列と同数出現するからである．例えば，n-gram を n=6 で行った場合に“ アップデート ”という頻出文字列が検出されたとする．この場合，n=5 の n-gram の頻出文字列解析で“ ップデート ”,“ アップデー ”が“ アップデート ”の出現回数と少なくとも同じ数だけ検出される．“ ップデート ”,“ アップデー ”は“ アップデート ”の部分文字列である場合，不要な文字列である．よって本研究では頻出文字列として出てきた文字列の部分文字列を関連語とみなさないよう削除することとする．この際，文字列，その出現回数を配列に保存する．これを n=12 の時以外，n が現在の n より 1 大きい場合の n-gram で解析を行った時に保存された配列を参照することにより部分文字列を削除することができる．具体的な方法として，n=i で部分文字列の削除を行う場合，n=i+1 のときに保存された配列を参照し，文字列の前後どちらかから一文字削ったものに一致する文字列を検索する．一致した文字列が発見された場合，出現回数を参照する．出現回数が同じであった場合，削除を行う．部分文字列と判断された文字列の方が出現回数が多い場合，部分文字列以外に出現していると考えられるので，n=i+1 の時に出現した文字列の出現回数を部分文字列として出現した回数として，n=i の文字列から n=i+1 の出現回数を引いたものが，その文字列が部分文字列以外で出現した回数となる．よってその回数をその文字列の出現回数とする．

2.3 発言者の情報を用いた関連語判別

2.1 節 で述べた n-gram を用いた頻出文字列解析と 2.2 節で述べた部分文字列の削除で述べた手法で頻出文字列をとることはできるが，それらは必ずしも関連語として扱うことはできない．例えば，複数の人が選手名を発言し，少数の発言者が宣伝のため商品を宣伝する商品名を含んだ同じ投稿文を複数回行った場合である．この場合，頻出文字列だけで判別すれば，選手名も商品名も出現頻度が高い文字列である．しかし商品名は頻出文字列であるが，関係のある分野とは言い難い文字列であるが，関連語として抽出されることとなる．実況中継中，宣伝のためにスポーツに関係する商品や書籍の情報を投稿するボットが存在する可能性は高い．この場合，頻出文字列として扱うことができると上記の二つで判別され，その頻出文字列は関連語でないにも関わらず関連語として扱われてしまう．これを防ぐために，発言者の情報を用いて関連語であるかどうかの判別を行う．そのために主要な発言者群の頻出文字列を関連語として扱い，それ以外は関連語として扱わない．発言者群とは発言者間の類似度が高いと認められる発言者の集合である．

利用者が入力したキーワードにより投稿文を取得する際, その時間多くの人がその分野についての発言を多く行っていると仮定すると, 主要な発言者群は利用者の入力したキーワードに関連する分野の発言を行っていることになる．よって主要な発言者群とは利用者の入力したキーワードに関連する分野を発言した発言者群となり，主要でない発言者群は利用者の入力したキーワードと関連のない分野を発言した発言者群となる．よって発言者群の特定を行うことによって，その発言者が利用者の入力したキーワードに関することを述べているかどうかを判別することができる．

発言者群の特定のため，発言者の評価点を算出する．評価点とは，発言者の情報を用いてどれだけ多くの人と同じ文字列を発言したか，どれだけ多くの回数をほかの人と同じ文字列を発言したかを総合的に評価するものである．これが高ければ高いほど，より主要な発言者であるといえる．評価点の算出は以下の四つの式に従う．

$$A_{i,j}(M_i, x_j) = \begin{cases} \mathrm{x}_j & (x_j < M_i) \\ M_i & (x_j \geqq M_i) \end{cases} \tag{1}$$

$$D = \sum_{i=0}^{m} \sum_{j=0}^{n} A_i, j\ (M_i, x_j) \tag{2}$$

$$N = \sum_{i=0}^{m} F_i \tag{3}$$

$$S = N \times (1 - \alpha) + D \times \alpha \tag{4}$$

(1) 式により，文字列毎の発言者間の類似度の算出を行う．発言者間の類似度は，文字列毎の発言回数が重なった回数の総和である．i は文字列の出現回数によって上位から順につけられた文字列の番号，j は発言者の名前が出現した順につけられた発言者の番号とする．$A_{i,j}$ は i 番目の文字列における j 番目の発言者との発言者間類似度である．スコアリングを行う発言者が i 番目の文字列を発言した数を M_i とし，x_j は i 番目の文字列を発言している j 番目の発言者が i 番目の文字列を発言した数とする．この時，M_i より x_j の値が大きい場合 M を，そうでない場合は x_j を $A_{i,\ j}$，つまり発言者間の類似度の値とする．

（2）式により，発言者間の類似度の算出を行う．m は文字列の異なり数，n は発言者の人数である．(1) 式により求められた A_i, j は i 番目の文字列におけるスコアリングを行う発言者と j 番目の発言者の類似度である．各文字列，各発言者間の A_i をすべて足し合わせたものを D とし，これを評価点を算出される発言者の類似度とする．

（3）式により，発言者数の算出を行う．N は評価点を算出される発言者と同じ文字列を発言した発言者数であり，各文字列の F_i の合計である．F_i は各文字列における，発言者と同じ文字列を発言した発言者の人数である．各文字列の F_i を足し合わせたものを N とする．

（4）式により, 発言者毎のスコアを求める．スコアが高ければ高いほど, より多くの人と同じ文字列を投稿していることとなる．S がスコアとなり，これは（2）式で求めた発言者間の類似度である D, (3) 式で求めた発言者数 N を用いて行う．α はパラメータである. 発言者数をより重要であると評価するため，試験的に α =0.2 とすることで，発言者数の値が高い発言

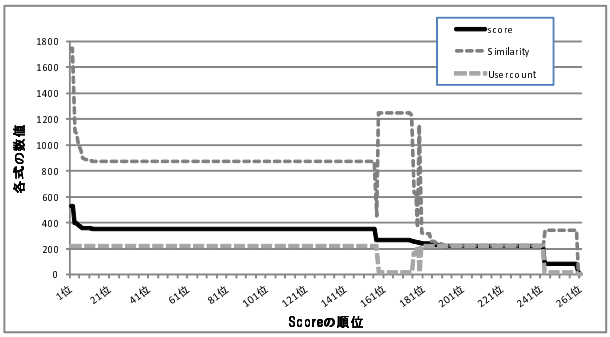

図 2　評価点算出の結果例

者のスコアが高くなるようにする．

図 2 は実際に上の式を適用した結果である．縦軸はそれぞれの数値，横軸は S の数値が高い順位になっている．この結果に限らず，スコアは上位の人ほど変動が大きくなる傾向にあった．よってスコアが高いものの上位半分は主要な発言者群の一員とみなす．スコアが下位半分の人の中から主要で無い発言者群を発見する．スコアが下位半分の人うち，自分よりスコア順が一つ上の人との差が大きい人を主要な発言者群と主要で無い発言者群の境目とする．図 2 においては順位が 160 位付近で観測される，発言者の類似度が高くなり，発言者数が減少する部分が境目になる．この場合，境目と判断された人のスコア以下である人は全て主要で無い発言者群の一員とする．結果からわかるように，少数の発言者で多く同じことを発言している発言者群を主要で無い発言者群として識別できている．

次に関連語の候補の発言者の参照する．関連語の候補の中に主要な発言者群で無い発言者が，その関連語の発言者全体の閾値以上の人数が存在した場合，関連語の候補からその文字列を外す．これにより，発言者の情報を用いない関連語判別より高い精度で関連語を抽出することができると考えられる．

最後に関連語の候補から，利用者が入力したキーワード，またそのキーワードを含む文字列を除くことにより，利用者の入力したキーワードの関連語が抽出することができる．

3. 実　　験

本手法によって得られた新しいキーワードと形態素解析を用いて得られた新しいキーワードの再現率・適合率を調べ，再現率・精度曲線を作成する．再現率・精度曲線を用いることで，本手法と形態素解析を用いた手法のどちらの推薦が精度が高いかを調べる．またどのような文字列が推薦されたかを見ることにより，何が原因で精度が上がったのか，下がったのかを考察し，手法の改善を行う．

3.1　利用する投稿文

今回の実験で利用する投稿文は以下の通りである．

- **投稿文** 1

2010 年 12 月 31 日に放映された格闘技の大会に関する投稿文を取得することを目的とし, 利用者がキーワードとして“ K-1,Dynamite ”と入力したと仮定して行う．実際に取得された投稿文は英数字のみのものを除き 6246 ツイートであった. 投稿文は 12 月 31 日 20:53〜23:27 に投稿されたものである．

- **投稿文** 2

2011 年 1 月 21 日に中継された AFC アジアカップ 2011，日本対カタール戦に関する投稿文を取得することを目的とし，利用者がキーワードとして ”daihyo ”と入力したと仮定して行う．実際に取得された投稿文は英数字のみのものを除き，8518 ツイートであった．投稿文は 1 月 21 日 23:17〜23:56 に投稿されたものである．

- **投稿文** 3

2011 年 1 月 25 日に中継された AFC アジアカップ 2011，日本対韓国戦に関する投稿文を取得することを目的とし，利用者がキーワードとして ”daihyo ”と入力したと仮定して行う．実際に取得された投稿文は英数字のみのものを除き，11389 ツイートであった．投稿文は 1 月 21 日 22:21〜22:54 に投稿されたものである．

このデータを利用する意義として，関連語として選手名，対戦国名がとれる可能性が高い点，また利用者が予想できない事が起こりうる点があげられる．提案した利用者の入力したキーワードによって得られる投稿文からの関連語の抽出手法を，Twitter のツイートを利用し評価する．

3.2　実 験 方 法

上記で述べたデータを利用し，実際に本研究の手法で関連語を抽出する．また比較対象として形態素解析を用いた関連語抽出も同じデータを用いて実験を行う．この形態素解析を用いた関連語抽出では，MeCab [1] を形態素解析器として利用する．今回利用する辞書は辞書は IPA コーパスにより CRF でパラメータ推定された IPA 辞書である．

3.1 節で述べた投稿文をそれぞれ，本研究の手法と MeCab を用いた形態素解析の手法で解析し，関連語の候補を取得する．今回，性能評価のために再現率・精度曲線を用いる．推薦された文字列のうち，適合する文字列の数を数える．本手法と形態素解析でそれぞれ新しいキーワードとして推薦された文字列のうち適合する文字列数の合計が，全文書中の適合文書の数となる．本手法，形態素解析で推薦された文字列が重複していた場合，適合文書より，重複していた文字列の数を引く．

推薦された関連語についてそれぞれ人手で適合するか否かを調べ，推薦された関連語中の適合率を調べる．適合するか否かの判断基準として，投稿文が取得された時間に推薦された関連語をキーワードとして入力した場合，利用者が入力したキーワードと関連する分野が取得できると考えられるものを適合するとして評価した．適合するかどうかの判断基準として，利用者は入力したキーワードの番組についてのみ興味を持ち，その番組に関係の無いものは全て不適合とする．また本手法，形態素解析どちらかに適合すると判断された文字列の部分文字列であった場合，それは適合しないとする．例としては投稿文 1 では本手法において“ 長島☆自演乙☆雄一郎 ”が新しいキーワードとして推薦された．この文字列の部分文字列である“ 自演乙 ”,“ 自演 ”,“ 雄一郎 ”なども同じように新しいキーワードとして推薦されているが，これらは適合しないと判断する．

前述の基準を用いて，本手法と形態素解析の精度を調べるこ

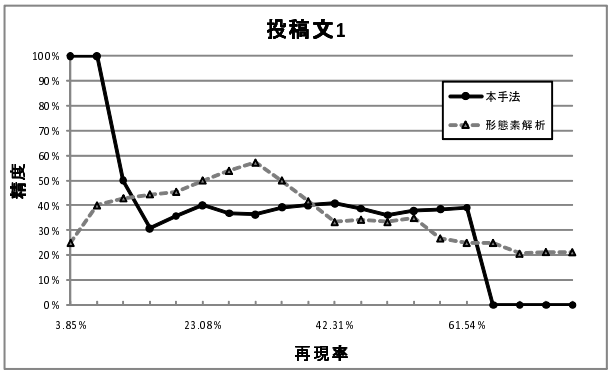

図 3　投稿文 1 の再現率・精度曲線

とにより，高い精度を持つシステムがより高い性能を持つということができる．

また投稿文から引用文を除いて，文字列の出現回数によってキーワードの推薦を行った場合と投稿文から引用された文を除かず，文字列の出現割合によってキーワードの推薦を行った場合の推薦されたキーワード，上位 10 個の結果をそれぞれ投稿文 1 のデータを解析した場合のみ載せる．これにより，出現回数を用いた場合と出現割合を用いた場合の比較，引用された文を除かない場合と引用された文を除いた場合の比較を行う．

3.3 実験結果

取得した投稿文それぞれの実験結果について述べる．

- **投稿文** 1

投稿文 1 のデータから推薦された新しいキーワードの結果はそれぞれ，n-gram の解析で出現回数順を用いたものの結果が表 1，n-gram の解析で出現割合順を用い，引用された文を除かなかったものの結果が表 2，n-gram の解析で引用された文章を除き，出現割合順を用いたものの結果が表 3 である．また形態素解析を用いたも手法結果が表 4 となっている．これらは推薦される新しいキーワード候補のうち，それぞれの基準で上位 10 位までのものを示している. また本手法と形態素解析を用いた手法の再現率・精度曲線は図 3 となっている．

本手法で出現回数だけで取った場合，推薦されたキーワードの上位 10 位までの結果は表 1 のようになった．二文字の文字列である“ った ”などの文末表現,“ ない ”,“ から ”などのある特定の分野に出現するわけでない文字列が多く推薦された．これは多くの人が発言した文字列中に含まれているが，その文字列が複数の種類があるためにそれらが関連語の候補とならないため，部分文字列としても扱えないからであると考えられる．これより単純に出現回数だけで行った場合，不必要な短い文字列が上位を占め，長い文字列があまり推薦することができなくなる．よって本研究では文字列の出現回数をその文字列のサイズで出現した文字列全ての出現回数の総計で割ったものである出現割合を用いて新しいキーワードの推薦を行う．

また本手法と同様の方法で引用文を除かない場合，多くの人に引用される文の一部である文字列が新しいキーワードとして推薦された．推薦されたキーワードの上位 10 位までの結果は表 2 のようになった．今回，多く引用されていた文は以下の文である．

> 自演乙ミルキィホームズまとめ：
> http://twitpic.com/3ljw8d
> http://twitpic.com/3ljvwl
> http://twitpic.com/3ljw28
> # milkyholmes # dynamite # jienotsu

本手法では URL を解析対象としないためこの引用文から取得されるのは“ 自演乙ミルキィホームズまとめ：”,“ # milkyholmes ” # dynamite ” # jienotsu ”の文字列，もしくはこれらの部分文字列となる．出現割合を利用しているため，長い文字列がより上位に出現しやすい．引用された投稿文が多い場合，n-gram のもっとも大きいサイズのものが多く取れ，検索のノイズになりやすい．よって本研究では引用された投稿文は除いて，n-gram の解析を行う．

本手法を用いた新しいキーワード推薦を行うシステムによって，推薦されたキーワードの上位 10 位までの結果は表 3 のようになった．出現割合を用いたことにより，出現回数を用いた方法よりも長い文字列を多く取得することができた．しかし，出現割合を用いた場合でも，出現回数が多いために“ った ”,“ って ”,“ ない ”のような特定の分野のみに出現する文字列ではない文字列も多く推薦されてしまった．本手法で推薦された新しいキーワードのうち適合していた文字列は 17 個であった．

形態素解析を用いた新しいキーワード推薦を行うシステムによって，表 4 のような文字列が新しいキーワードとして推薦された．形態素解析で推薦される新しいキーワードは全体的に短いものが多かった．その理由として，本手法で一つの文字列として推薦されているものが，形態素解析で利用される辞書に登録されていないため，辞書に登録されている文字列に分解して判別するためである．例として本手法では選手名である“ アリ

表 1　投稿文 1 出現回数

順位	出現回数	文字列
1	776	った
2	776	って
3	614	ない
4	407	から
5	386	して
6	375	てる
7	341	# TBS
8	312	紅白
9	275	自演乙
10	274	さん

表 2　投稿文 1 本手法を用いて引用文を除かない場合

順位	出現割合	出現回数	文字列
1	0.0179	280	# MILKYHOLMES
2	0.0101	158	自演乙ミルキィホームズま
3	0.0101	158	乙ミルキィホームズまとめ
4	0.0101	158	演乙ミルキィホームズまと
5	0.0101	158	ミルキィホームズまとめ：
6	0.0094	205	# JIENOTSU
7	0.0068	965	って
8	0.0064	912	った
9	0.0061	432	# TBS
10	0.0051	726	ない

表 3　投稿文 1 本手法

順位	出現割合	出現回数	文字列
1	0.0364	36	# MILKYHOLMES
2	0.0075	341	# TBS
3	0.0069	776	って
4	0.0069	776	った
5	0.0054	614	ない
6	0.0041	23	長島☆自演乙☆雄一郎
7	0.0038	14	MILKYHOLMES
8	0.0038	14	# MILKYHOLME
9	0.0035	407	から
10	0.0034	275	自演乙

表 4　投稿文 1 形態素解析

順位	出現回数	文字列
1	401	TBS
2	362	紅白
3	325	自演
4	305	試合
5	272	石井
6	235	ガキ
7	233	青木
8	223	さん
9	213	渡辺
10	182	今年

スター ”という文字列が推薦されている．形態素解析を用いたシステムでは“ アリスター ”という文字列を一つの文字列として認識することができなく,“ アリ ”,“ スター ”と認識できる形まで分解して 2 つの文字列として認識し，それぞれを推薦している.“ アリ ”,“ スター ”は“ アリスター ”の部分文字列であるため，今回適合しないと判別される．形態素解析を行い，名詞だけを取得するとしたため，本手法で推薦されてしまった“ って ”,“ った ”といった，特定の分野のみに出現する文字列でない文字列の多くは推薦されなかった．形態素解析を用いた手法で推薦された新しいキーワードのうち適合していた文字列は 20 個であった．本手法，形態素解析を用いた手法で推薦された新しいキーワードのうち適合していた文字列であり，重複していたものは 9 個存在した．

投稿文 1 の再現率・精度曲線は図 3 のようになった．再現率が 15 %から 40 %，65 %以降では形態素解析を用いた手法が，本手法よりも高い性能を示している．その原因として，2 文字の適合しない文字列は出現回数が他の文字列に比べ多いため，再現率 3 %から 40 %までの文字列に多く含まれていたことが考えられる．また選手名が名字の 2 文字で投稿されることが多いことが挙げられる．出現回数が一緒であっても，出現割合は文字列の長さによって大きく変動する．このため 2 文字の選手名である文字列は 3 文字以上の選手名よりも多い出現回数が必要となる．形態素解析においては出現回数によって判別される．よって本手法では推薦される文字列ではない選手名が，形態素解析を用いた手法では推薦される文字列となった．これにより，推薦された新しいキーワードのうち適合していた文字列の数が増えるため，65 %以上の再現率で本手法よりも高い性能を示したと考えられる．

- **投稿文** 2

投稿文 2 のデータから推薦された新しいキーワードの結果はそれぞれ，本手法を用いたものの結果が表 5 である．また形態素解析を用いた手法の結果が表 6 となっている．これらはそれぞれ出現回数が上位 10 位のものを示している. また再現率・精度曲線は図 4 となっている．

本手法を用いた新しいキーワード推薦を行うシステムによって推薦されたキーワードの上位 10 位までの結果は表 5 のようになった．今回は利用者の入力したキーワードはハッシュタグとして利用されていた“ # daihyo ”の“ daihyo ”を用いた．サッカー関連のハッシュタグは多く存在し，上位 10 位まではハッシュタグ，もしくはハッシュタグの部分文字列となっている．10 位以下の結果は“ った ”,“ って ”,“ ない ”など特定の分

表 5　投稿文 2 本手法

順位	出現割合	出現回数	文字列
1	0.0525	1106	# TVASAHI
2	0.0368	307	ASIANCUP2011
3	0.0368	307	# ASIANCUP201
4	0.0326	687	# ASIACUP
5	0.0254	212	# SAMURAIBLUE
6	0.0244	356	# ASIANCUP
7	0.0204	431	# JFA2011
8	0.0184	268	# ZACJAPAN
9	0.0169	517	# AC2011
10	0.0169	516	# SOCCER

表 6　投稿文 2 形態素解析

順位	出現回数	文字列
1	1625	TVASAHI
2	1508	ASIACUP
3	1343	JFA
4	1007	ASIANCUP
5	974	SOCCER
6	906	PK
7	753	韓国
8	748	日本
9	577	AFC
10	550	JPN

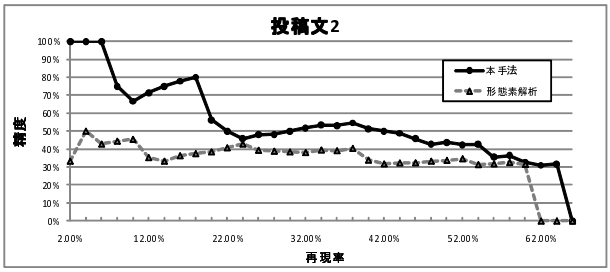

図 4　投稿文 2 の再現率・精度曲線

野に関わらず出現する文字列が投稿文 1 の時と同様に推薦されていた．しかし投稿文 1 の時と異なり，2 文字の文字列でかつ，選手名であるものも多く推薦されていた．本手法を用いて推薦されたキーワードのうち適合すると考えられる文字列は 32 個存在した．

形態素解析を用いた新しいキーワードの推薦を行うシステムによって推薦されたキーワードの上位 10 位までの結果は表 6 のようになっている．本手法で得られた結果と同じように“ TVASAHI ”,“ JFA ”,“ ASIACUP ”,“ ASIANCUP ”,“ SOCCER ”といった，ハッシュタグに関する文字列が多く取得された．また形態素解析では閾値以下の英字が連続した文字列を名詞と判断するため，本手法で推薦されたハッシュタグを正しく認識出来なかったと考えられる．例えば本手法で推薦された文字列の“ # JFA2011 ”の場合，形態素解析を行った場合“ # " JFA " 2011 ”の 3 つの文字列と識別されてしまう．これにより形態素解析を用いた手法では“ JFA ”が推薦される．これは本手法で推薦された文字列の部分文字列であるため，不適合と判断される．形態素解析を用いた手法で推薦された新しいキーワードのうち適合していた文字列は 30 個であった．本手法，形態素解析を用いた手法で推薦された新しいキーワードのうち適合していた文字列であり，重複していたものは 12 個存在した．

投稿文 2 の再現率・精度曲線は図 4 のようになった．全体を通して本手法が高い性能を示している．形態素解析ではハッシュタグを識別することができなく，本手法の上位にハッシュタグが多く存在していたため，形態素解析を用いた手法で推薦された文字列の上位は不適合と判別された．その結果，再現率が低い時点での精度が低くなったと考えられる．また投稿文 1 の時の結果とことなり，本手法でも 2 文字の選手名が形態素解析を用いた手法と同程度，新しいキーワードとして推薦されたため，再現率が高くなっても精度が形態素解析を用いた手法を下回ることがなかったと考えられる．

- **投稿文** 3

投稿文 3 のデータから推薦された新しいキーワードの結果はそれぞれ，本手法を用いたものの結果が表 7 である．また形態素解析を用いた手法の結果が表 8 となっている．これらはそれぞれ出現回数が上位 10 位のものを示している. また再現率・精度曲線は図 5 となっている．

本手法を用いた新しいキーワード推薦を行うシステムによって推薦されたキーワードの上位 10 位までの結果は表 7 のよう

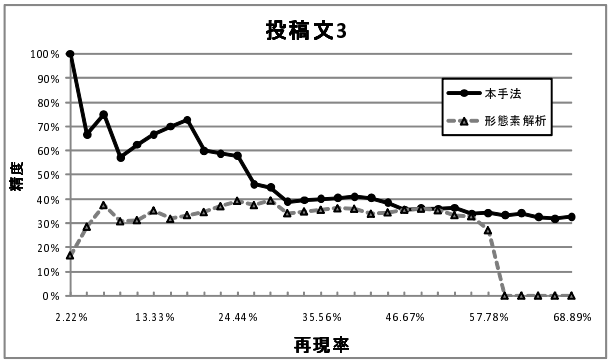

図 5　投稿文 3 の再現率・精度曲線

になった．今回も投稿文 2 の時と同様に，利用者の入力したキーワードはハッシュタグとして利用されていた“ # daihyo ”の“ daihyo ”を用いた．投稿文 2 の時よりも発言者の数が多く，またハッシュタグも多くつけられていた．今回も投稿文 2 の時と同様に上位 10 位まではハッシュタグ，もしくはハッシュタグの部分文字列となっている．10 位以下の結果は，投稿文 2 の時は“ った ”,“ って ”など特定の分野に関わらず出現する文字列が不適合な出現確率の上位に来ていたが，今回は“ おしいいいいいいいいいい ”,“ わあああああああああ ”などとテレビの中継に対しての感嘆などのみの投稿が投稿文 2 に比べ多く見られた．このような文字列が複数人に発言され，また文字列自体も長いため，出現割合で上位に出現し，新しいキーワードとして推薦された．本手法を用いて推薦されたキーワードのうち適合すると考えられる文字列は 31 個存在した．

形態素解析を用いた新しいキーワードの推薦を行うシステムによって推薦されたキーワードの上位 10 位までの結果は表 8 のようになっている．投稿文 2 のときと同様にハッシュタグに関する文字列が多く取得された．本手法を用いた場合,“ おしいいいいいいいいいい ”などといった感嘆のみの投稿文は形態素解析を用いた手法では“ おしい ”という形容詞と“ い ”という文字列の連続したものと識別される．同様に“ わあああああああああ ”も“ わあ ”という感動詞と“ あ ”という文字列の連続したものと識別されている．形態素解析を用いた手法では名詞のみを推薦される新しいキーワードとするため，これらの文字列は推薦されない．形態素解析を用いた手法で推薦された新しいキーワードのうち適合していた文字列は 26 個であった．本手法，形態素解析を用いた手法で推薦された新しいキーワードのうち適合していた文字列であり，重複していたものは 12 個存在した．

表 7　投稿文 3 本手法

順位	出現割合	出現回数	文字列
1	0.0694	427	ASIANCUP2011
2	0.0692	426	# ASIANCUP201
3	0.0646	1613	# TVASAHI
4	0.0569	350	# SAMURAIBLUE
5	0.0375	1383	# ASIACUP
6	0.0298	566	# ASIANCUP
7	0.0248	450	# ZACJAPAN
8	0.0248	974	# SOCCER
9	0.0221	136	# ASIACUP2011
10	0.0210	525	# JFA2011

表 8　投稿文 3 形態素解析

順位	出現回数	文字列
1	1625	TVASAHI
2	1508	ASIACUP
3	1343	JFA
4	1007	ASIANCUP
5	974	SOCCER
6	906	PK
7	753	韓国
8	748	日本
9	577	AFC
10	550	JPN

投稿文 3 の再現率・精度曲線は図 5 のようになった．全体を通して本手法が高い性能を示している．しかし 50 %付近では形態素解析を用いた手法と大きく変わらない精度を示した．これは投稿文 2 に比べ，感嘆のみの投稿文が多く存在し，それが精度を下げたためであると考えられる．“ あ ”などが連続した文字列は n-gram で解析する際に予め削除しているが，それは n-gram の解析で，その文字が全て“ あ ”など削除すべき文字列であった場合のみを削除しているため，今回のような文字列は削除することができなかった．

3.4 考　　察

全体的に本手法が形態素解析を用いた手法よりも良い性能を示した．その原因として，形態素解析では取得することのできない文字列を本手法は取得することが出来，それらを新しいキーワードとして推薦できたことが挙げられる．しかし，形態素解析では取得することのできない文字列が適合していない場合は，精度を下げてしまうノイズを形態素解析よりも多く取得することになる．どの投稿文にも存在した，ある特定の分野だけ限らず，どのような分野の投稿文にも出現しうる文字列や，感嘆文などが主に今回，精度を大きく下げる文字列となった．今後の課題として本システムの高精度化が必要となる．高精度化するために，今回取得してしまったノイズを関連語と識別しない新しい制約が必要と考えられる．具体的にはすべての投稿文に共通して“ った ”,“ って ”,“ ない ”という文字列が多く存在したため，複数の投稿文に共通して出現する適合しないが新しいキーワードとして推薦される文字列を予め削除を行う．または形態素解析と併用し，本手法によって新しいキーワードとして推薦される文字列を形態素解析によって評価し，明らかに名詞でないものを削除するなどといった方法によりノイズの削除を行うことが必要だと考える．

4. 関 連 研 究

近年，Twitter を筆頭としたマイクロブログに関する研究が多く存在する．Twitter 上でのトレンド語を検索する Web サービスとして buzzter (注3)がある．このサービスではパブリックストリームから利用者全体の投稿文を取得し，キーワード毎の評価値を発言者全体のうち，キーワードを発言した割合を用いて算出する．算出した評価値をランク付けし，トレンド語を発見する．本研究では利用者の入力したキーワードを利用し投稿文を取得しているため，入力されたキーワードに関する新しいキーワードを取得するため，利用者が入力したキーワードに関する分野に特化した新しいキーワードを取得できると考えられる．

ある語句を自動抽出する手法として加藤らの方法 [2] がある．この研究はパブリックストリームより得られたすべての投稿文を利用し，前日との文字列の出現頻度が大きく変動するものを単語として考え，形態素解析の辞書に登録すると同時に，情報の発生源，それがどのように伝播したか，どのような議論が行われたかを観察する社会分析を行う研究である．本研究も利用

(注3)：http://buzztter.com/ja

者の入力したキーワードについてのツイートを見渡すという目的は類似しており，また既存の辞書を用いた形態素解析を利用した Twitter の解析は困難であると考える点は類似している．しかし，既存の辞書の形態素解析の問題点についての解決手法は大きく異なっており，頻出文字列の取得方法も異なる．

また利用者の入力したキーワードで検索されたツイートの中から有用な記事の発見を行う岩木らの方法 [3] がある．この研究では本研究と同じように，単語の出現頻度のみでなく，発言者の情報を用いてマイニングを行っている．本研究でも発言者の情報を用いている．しかし本研究では，人のネットワークの関係を加味せず，取得したデータ上で，主要な発言者群か否かを調べ，そのデータを利用する点で異なっている．

また例えば大石らの方法 [4] といった検索クエリの拡張に関連研究も多く存在する．本研究の利用者の入力したキーワードは検索クエリとして見ることができ, 関連語は検索クエリに追加される新しいクエリとしてみることができる．しかし多くの検索クエリの拡張の研究の目的は複数のクエリを入力し, その複数のクエリで AND 検索を行うことで情報を絞り込むことである．本研究では複数の検索クエリを入力し, その複数のクエリで OR 検索を行うことで, 利用者の知識外のツイートを取得することを目的としている点で大きく異なる．

以上より，本研究の特徴を考えると，形態素解析による頻出文字列の解析は Twitter に向いていないとして n-gram を利用した解析を行った点，発言者の情報を用いた頻出文字列が関連語か否かの判断を行った点，そしてそれらを用いた新しい検索キーワードを加えた検索クエリの拡張により，さらに多くの利用者の入力したキーワードの分野のツイートを得ることを可能にしたという点が挙げられる．

5. ま と め

現在，Twitter 上から利用者が自分の興味のある人物を探し出すために，自らの検索を行う場合，公式サイトの検索フォームよりキーワードを入力して投稿文を検索する必要がある．この時，多くの投稿文が存在し，検索中にも大量の投稿文が投稿される．この中から新しい話題を発見し，新しい話題のキーワードを含め，複数のキーワードを検索フォームに入力し，投稿文を見ると仮定する．この時，利用者は大量の投稿文を見る必要があり，利用者の負担が大きいと考えられる．よって本研究では利用者の入力したキーワードを利用し，投稿文を取得し，利用者の入力したキーワードに関連性の高いと考えられる新しいキーワード自動抽出する手法を提案した．

本研究ではこのようなキーワード抽出の支援を n-gram と発言者の情報を用いた関連語判別を用いて行う．Twitter の投稿文は通常の文章と異なり，くだけた文章になりがちであるため，形態素解析によって頻出文字列の解析を行うには不向きと考え，本研究では n-gram を用いた．n-gram で頻出文字列を取得する際，部分文字列が出現するため，部分文字列の棚ぞを行った．発言者の情報を用いてノイズの除去を行った．少数の人達が同じような，利用者の入力したキーワードが含まれてかつ関係のない文字列を取り除いた．実験では，本手法と Twitter での頻出文字列の解析が不向きと考えた形態素解析を用いた手法の比較実験を行った．その結果，大部分で本手法が優れていることが示すことができた．しかし，本手法で取得された関連語の候補は多くのノイズを含んでおり，それらによって本手法の精度は低くなった．

今後の課題として，大部分で本手法は形態素解析よりも優れていたが，まだ不必要なノイズを多く取得している．これにより，本手法の精度が低くなった．さらに高精度なキーワード抽出を目指すために，それらのノイズの除去を行うシステムの考案が考えられる．具体的には，複数の投稿文に共通して出現する適合しないが，新しいキーワードとして推薦される文字列群を予め削除を行う．または形態素解析と併用し，本手法によって新しいキーワードとして推薦される文字列を形態素解析し，明らかに名詞でないものを削除するといったことが考えられる．また本手法を利用した発展的なシステムの開発が考えられる．現在，本手法は投稿文を取得して，新しいキーワードを推薦している．よって本手法を用いて，リアルタイムにキーワードの推薦を行うシステム，投稿文の種類分けなどを行うシステムなどが考えられる．

謝　　辞

本研究の一部は，科研費（22300034，20300036，20500104）による．ここに記し，謝意を示す．

文　　献

[1] 形態素解析器“ Mecab ”http://mecab.sourceforge.net/
[2] 加藤 慶一, 秋岡 明香, 村岡 洋一, 山名 早人“ ミニブログにおける注目語抽出手法の提案と注目語を用いたメディア間での話題追跡 ”WebDB Forum2010
[3] 岩木 祐輔, アダム ヤトフト, 田中 克己“ マイクロブログにおける有用な記事の発見 ”DEIM Forum 2009
[4] 大石 哲也, 峯 恒憲, 長谷川 降三, 藤田 博, 越村 三幸“ 関連単語抽出アルゴリズムを用いたクエリ拡張 ”DEIM Forum 2009