経時的な関連語句の変化を考慮したクエリ拡張による Twitter からの情報抽出手法

藤木 紫乃[†1]　上田 高徳[†1]　山名 早人[†2†3]

†1 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
†2 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1
†3 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †{fujiki, ueda, yamana}@yama.info.waseda.ac.jp

あらまし 近年，SNS を通じてリアルタイムな情報共有が盛んに行われている背景を受け， 新鮮さと関連性の両方を考慮した情報検索が研究されている．しかし，Twitter は 1 つの投稿が 140 文字までと限られているため，検索語句と関連が深いものの検索語句を正確に含まない投稿も多く行われている. また, 検索語句に関連する語句は, 新しい事件や話題の発生により経時的に変化していく．従来の研究では，時間を問わず大量に投稿されるスパムツイートの影響により，検索語句を適切に選択できず，投稿を適切に検索することができない場合があった．本稿では，経時的な関連語句の変化を考慮した時間的なクエリ拡張を行い，検索語句に関連した新鮮な情報を抽出する手法を提案する． 実験の結果，提案手法は従来手法に比較し精度を落とすことなく，スパムツイートを排除したタイムリーな情報検索を行うことができた.

キーワード マイクロブログ，Twitter，情報検索

1. はじめに

全世界で 5 億人[1]のユーザを持つ Twitter[2]は，140 文字以内の文章（以下，ツイートとする）を投稿することで交流を行うソーシャルネットワークサービスである．Twitter では，時事的な話題や TV 番組の内容などについてリアルタイムな情報共有が行われている．しかし，多くのユーザはリアルタイム性とその文字数の少なさのために，文法的に不正確で端的なツイートを行いがちである．そのため，Twitter 上で検索を行う場合，検索クエリを正確に含まない場合であってもクエリと意味的に関連の深いツイートは検索できない可能性が高い．このため，現在，Twitter の投稿のような短文に対して効果的な検索システムが求められている.

従来の検索システムでは，文書自体の重要度と，検索クエリと文書との関連度を組み合わせてランキングを行い，適した文書を検索結果として返している．しかし，近年の Twitter の隆盛を受け，短文でリアルタイム性の高い文書に適した検索手法が研究されている.

このような情報検索手法には，Twitter 公式の検索機能[2]や，ツイートの質的評価と時間的特徴を用いてランキングを行う研究[3][4], ランク学習によってソーシャルな特徴や時間的な特徴，文章的な特徴などを組み合わせてランキングを行う研究[5]がある.

Twitter は 1 つの投稿が 140 文字までと限られているため，検索語句と関連が深いものの検索語句を正確に含まない投稿も多く行われている．本稿では，こうした検索語句を直接含まないものの，検索語句と関連の深いツイートの検索を可能とする．これを実現するため, Twitter データを用いたクエリ拡張を行う. この時, 検索語句に関連する語句は，新しい事件や話題の発生により経時的に変化していくことに着目する.

例えば，歌手名を検索した場合，新曲を発売した時期における関連語句は，曲名や歌詞，PV の URL や奏者名等であると考えられる．しかし，その歌手が結婚すると報じられた時に必要とされる関連語句は，結婚相手や結婚の経緯に関する語句であると考えられる.

Twitter 検索と一般的な Web 検索の比較調査[6]によると，Twitter で検索を行うユーザの約半数はタイムリーな情報を求めていることが分かった．タイムリーなツイートを取得するという目的に対しては，ツイート投稿時間の範囲をしていた検索方法も考えられるが，検索したい情報に対して適切な時間範囲を設定するのは難しいと考えられる．よって，関連語句の経時的な変化を捉えることは，ユーザの検索効率向上に役立つと考えられる．また，Twitter で検索されやすい人気の語句は，しばしばスパム的なツイートに含まれることがある．このようなツイートを検索結果から排除する検索手法が必要である.

本稿では，経時的な関連語句の変化を考慮した時間的なクエリ拡張を行い，検索語句に関連した新鮮な情報を抽出する手法を提案する．ツイート中の語句，ハッシュタグ，URL の共起関係から関連語句を見つける. 長期間のデータにおける関連語句と，検索タイミングにおける短期間のデータにおける関連語句の比較から, 検索タイミングにおいて適切なクエリ拡張を行う.

本稿は以下の構成をとる．まず 2 節で関連研究をま

表 1　関連研究のまとめ

手法名	コーパス	特徴	スパム対応	正解セット
Twitter 公式検索，2006[2]	Twitter		×	不要
Massoudi ら，2011[3]	Twitter	クエリ拡張	×	不要
Efron ら，2012[4]	Twitter	ツイート拡張	×	不要
Zhang ら，2012[5]	WEB	ランク学習	○	必要
提案手法	Twitter	クエリ拡張	○	不要

とめ，3 節で提案手法について説明する．4 節で実験と評価を行い，最後に 5 節でまとめを述べる．

2. Twitter 検索の関連研究

2.1. Twitter における情報検索の調査

Teevan ら[6]は，一般的な WEB 検索と Twitter 検索の違いを調査しまとめている．調査の結果，ユーザは時間的に関連する情報，すなわちインターネット上の流行り言葉や，Twitter のユーザ，著名人を探すために Twitter 検索を利用すると分かった．また，一般的な WEB 検索が検索語句に関する情報を得るために行われ，検索結果が主に基本的な事実を返すのに対し，Twitter 検索は話題を観察するのに用いられ，検索結果はよりソーシャルな内容やイベントの情報であると述べている．更に，英語による WEB 検索と Twitter 検索のクエリの違いに着目すると，WEB 検索クエリの平均語数は 3.08 であるのに対して，Twitter 検索クエリの平均語数 1.64 であった．

以上より，Twitter 検索では，少数のクエリを与えて関係する最新情報を得るという利用法が多いことが報告されている．

2.2. Twitter の公式検索機能

Twitter[2]では，公式に検索機能を備えている．ユーザは単語，フレーズ，ハッシュタグ，言語，ユーザ名，メンション，地名，感情表現など，様々な条件を指定して検索を行うことができる．検索結果には，「トップ」，「すべて」，「あなたがフォローしているユーザ」の 3 種類がある．正確なアルゴリズムが公開されていないため推測になるが，「すべて」は検索語句を含むツイートを時系列降順に，「あなたがフォローしているユーザ」は検索ユーザのフォロイーの投稿の中から検索語句を含むツイートを時系列降順に表示している．「トップ」は，「すべて」のうち，リツイート回数やお気に入り回数が多いツイートをランキングの上位に置いていると考えられる．

しかし，Twitter 公式検索では検索語句を正確に含むツイートしか得られないため，検索結果の適合率は高いが，再現率は低いと予想される．2.1 項で述べたように Twitter において多くのユーザは少ないクエリしか用いないため，本来は検索語句と関係するが検索語句を正確に含まないツイートが取得できない．

2.3. 時間的性質を用いた情報検索

Massoudi ら[3]は，ツイート検索のためにクエリ拡張と質的評価を組み合わせている．クエリ拡張では，検索日時に近い共起語に高い重みを付与している．質的評価には，大きく分けて 2 種類の指標を用いている．1 つはテキストの品質によるもので，感情，文字長，URL の有無などに基づいて計算される．もう 1 つはマイクロブログの特徴によるもので，RT 回数，著者のフォロワー数，最新性によって計算される．この 2 つを組み合わせた質的評価とクエリ拡張によって，高い性能の検索を行なっている．

Efron ら[4]は Twitter やデジタルライブラリのように短い文書からなるコーパスにおける検索のため，語彙的，時間的な文書拡張手法を提案した．

語彙的な観点では，あらかじめ全ツイートに対して，ツイート D に類似したツイート集合からなる拡張表現 D′を作成する．クエリ Q に対する文書 D の適合確率 P(Q|D)と拡張表現 D’の適合確率 P(Q|D′)を組み合わせることで，文書の拡張を行う．時間的な観点では，Massoudi らの手法と同様に，投稿時間が検索タイミングに近いツイートほど上位に得られるようにランキングを計算する．

しかし，Massoudi らによる手法と Efron らによる手法ではスパムツイートが考慮されていないため，特定のクエリにおいて検索結果にノイズが含まれる可能性が高い．ここで述べるスパムツイートとは，アフィリエイト目的あるいはアカウントのフォロワーを増やしてマーケティングに利用する目的で，芸能人の名前や商品名を羅列して注目を集めようとするツイートである．スパムツイートの実例を以下に示す．ツイート内の URL は情報商材購読リンクである．

> ★マジで…！完全無料！ノーリスクで、５６０円が何度でも貰える方法です。（URL） #相互フォロー #RT #followback #autofollowjp #AKB48 #SMAP #野球 #ラーメン

このようなツイートは時間に関係なく大量に投稿されているため，クエリ拡張において時間的な話題とは関係ない語句が抽出されてしまう．また，スパムツイートを行うアカウントは，多くの一般ユーザよりもフォロワー数が多い傾向があるため，質的評価におい

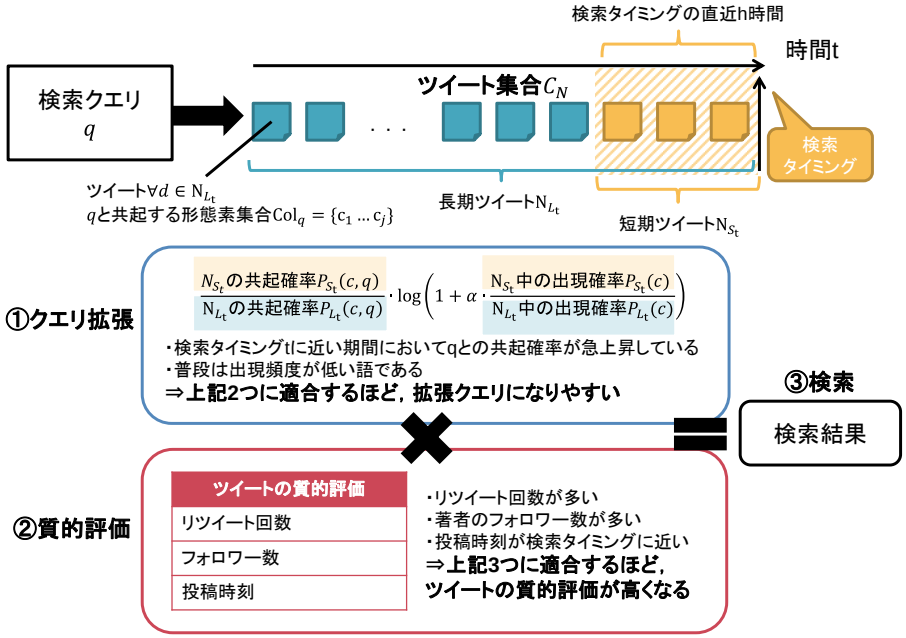

図 1　提案手法の流れ

ても一般的なツイートを上回り，検索結果として出力されることがある．

2.4. ランク学習により様々な特徴を組み合わせた情報検索

Zhang ら[5]は，様々な特徴量を用いてランキングを作成し，機械学習を用いて最適に組み合わせるランク学習手法による情報検索を提案した．特徴量には，クエリ固有の特徴,感情表現などを用いている.しかし，この手法では半教師あり学習を行なっているため，正解データを準備する必要がある．

2.5. Twitter 検索手法のまとめ

本節では大きく分けて 3 つのツイート検索手法を紹介した．1 つは Twitter の公式検索，もう 1 つは時間的な特徴と語彙的な特徴を組み合わせた手法，そして最後が，ランク学習によって様々な特徴量を組み合わせる手法である．各手法の特徴と，本稿で提案する手法の特徴を表 1 にまとめた．

本稿では，スパムツイートの影響を受けずにタイムリーな検索を可能にする手法を提案する．比較のため Massoudi らの手法を対抗手法として考察を進める．

3. 提案手法

本節では提案手法について述べる．提案手法の流れを図 1 に示す．

まずツイートを一定期間収集する．このツイート集合から，検索クエリと共起する語句を拡張クエリ候補として集める．検索タイミングに近い期間に出現頻度の上昇した共起語句を拡張クエリとする．拡張クエリはそれぞれ，検索クエリとの関連度を表すスコアを持つ．次に，ツイート検索を行う．元々のクエリと拡張クエリを用いて，検索タイミング以前に投稿されたツイートに対して OR 検索を行う．検索クエリとの関連度の高いクエリがより多く含まれているツイートを高く評価するよう，スコアを足し合わせる．また，検索結果として得られた各ツイートに対して，リツイート回数等を用いて質的な評価を行う．最後に，クエリ拡張による関連度評価と，ツイート自体の質的評価を組み合わせてツイート検索能力を向上させる．

各項で具体的な処理について述べる．

3.1. 事前準備と記号説明

まず事前準備として，コーパスとなる一定期間のツイートを収集する．合計 N ツイートからなるコーパス C_Nのうち，検索タイミング t 以前の全てのツイートを長期ツイートN_{L_t}と呼ぶ．また，長期ツイートのうち検索タイミング t の直近 h 時間に投稿されたツイートを短期ツイート$\mathrm{N}_{S_t} \subseteq \mathrm{N}_{L_t}$と呼ぶ．コーパス中の全てのツイート$\forall d \in N_{L_t}$に対して形態素解析を行い，単語，ハッシュタグ，URL を抽出し，共起関係を得る．

3.2. クエリ拡張

コーパスに対して検索語句としてクエリ qを与えたとき，長期ツイートN_{L_t}より，クエリ q と共起する形態素集合 $Col_q = \{c_1 \dots c_j\}$を抽出する．$Col_q$は最終的な拡張クエリ集合$\mathrm{EQ}_q$の候補となる．それぞれの$c \in Col_q$について，検索タイミング t に近い期間においてクエリqとの共起確率が急に上昇した語句を拡張クエリとする．このとき，普段の出現確率が低い語句であるほどスコアが高くなるようにすることで，スパムツイートによるクエリ拡張への悪影響を緩和させる．拡張クエリを

決定するためのスコアリング関数を（式 1）に示す．短期ツイートN_{S_t}内における単語wの出現確率を$P_{S_t}(w)$，短期ツイートN_{S_t}内における単語w_1, w_2の共起確率を$P_{S_t}(w_1, w_2)$とする．同様に，長期ツイートN_{L_t}内における単語tの出現確率を$P_{L_t}(w)$，長期ツイートN_{L_t}内における単語w_1, w_2の共起確率を$P_{L_t}(w_1, w_2)$とする．それぞれの具体的な式を（式 2），（式 3），（式 4），（式 5）に示す．

$$\mathrm{Score}_{expansion}(c, q) = \frac{P_{S_t}(\mathrm{c}, \mathrm{q})}{P_{L_t}(\mathrm{c}, \mathrm{q})} \cdot \log\left(1 + \alpha \cdot \frac{P_{S_t}(\mathrm{c})}{P_{L_t}(\mathrm{c})}\right) \quad \textbf{（式 1）}$$

$$P_{S_t}(\mathrm{c}, \mathrm{q}) = \frac{\left|\{d: c, q \in d, d \in N_{S_t}\}\right|}{\left|\{d: q \in d, d \in N_{S_t}\}\right|} \quad \textbf{（式 2）}$$

$$P_{L_t}(\mathrm{c}, \mathrm{q}) = \frac{\left|\{d: c, q \in d, d \in N_{L_t}\}\right|}{\left|\{d: q \in d, d \in N_{L_t}\}\right|} \quad \textbf{（式 3）}$$

$$P_{S_t}(\mathrm{c}) = \frac{\left|\{d: c \in d, d \in N_{S_t}\}\right|}{\left|\{d: d \in N_{S_t}\}\right|} \quad \textbf{（式 4）}$$

$$P_{L_t}(\mathrm{c}) = \frac{\left|\{d: c \in d, d \in N_{L_t}\}\right|}{\left|\{d: d \in N_{L_t}\}\right|} \quad \textbf{（式 5）}$$

qが与えられたとき，$\mathrm{Score}_{expansion}(c, q)$が高い上位 k 個の形態素 c を，最終的な拡張クエリ集合$\mathrm{EQ}_q = \{eq_1 \ldots eq_k\}$とする．この際，（式 6）のように拡張クエリ上位 k 個のスコアの和で正規化し，拡張クエリのスコアを 0 から 1 の間に設定する．

実際の検索においては，拡張クエリだけでなく本来の検索クエリqも利用するため，qにもスコアを設定する必要がある．正規化によって拡張クエリのスコアの範囲を設定することで，qのスコアを設定しやすくする．k=10 と設定した場合には拡張クエリのスコアは平均 0.1 となる．このとき，スコアが他の拡張クエリより著しく高い場合，検索結果がその拡張クエリに強く影響されてしまう．これを防ぐため，拡張クエリのスコアが$2/k$を超えた場合，$2/k$に切り捨てることとする．αは，短期ツイートと長期ツイートにおける出現頻度比の効果を調整するパラメータである．

$$nScore_{expansion}(eq, q) = \frac{Score_{expansion}(eq, q)}{\sum_{i=1}^{k} Score_{expansion}(eq_i, q)}$$

$$0 < nScore_{expansion}(eq, q) < 1 \quad \textbf{（式 6）}$$

$$\sum_{k} nScore_{expansion}(eq, q) = 1$$

なお，誤字などにより極めて出現頻度の少ない語句が拡張クエリとなることを防ぐため，長期ツイートにおいて少なくともλ個以上のツイートに出現した語句のみを用いる．また長期ツイートにおいてクエリ q と共起したツイートが少なくともμ個以上ある形態素 c を拡張クエリとする．これは長期ツイートにおいてクエリ q との共起回数が著しく少ない形態素は，クエリ q と関連性が少ない可能性があり，これを拡張クエリとすることは，クエリ拡張の能力を下げることにつながるからである．ただし，μ個以上のツイートで共起した拡張クエリの個数が k 個に満たない場合，μの値を 1 ずつ下げて，拡張クエリがλ個になるようにする．

3.3. ツイートの質的評価

Massoudi ら[3]の提案手法のうち，言語を問わず適用できるマイクロブログ固有の特徴を用いて質的評価を行う．具体的には，ツイートの RT 回数，著者のフォロワー数，投稿時刻の 3 種類を用いる．

まず，（式 7）より，ツイート d の RT 回数が多いほど$P_{repost}(d)$のスコアを高くする．次に，（式 8）より，ツイート d の著者のフォロワー数が多いほど$P_{followers}(d)$のスコアを高くする．最後に，（式 9）より，ツイート d の投稿時刻が検索タイミングに近いほどスコアを高くする．γは最新性に関するパラメータである．t_dはツイート d の投稿時刻を表す．

最終的なツイート d の質的評価として，（式 10）より，3 種類の質的評価の平均を計算し，次項のツイート検索を行う．芸能人のようにフォロワー数の多いユーザの発言などは質的評価が高いと見なされ，クエリとの意味的な関連度が少ないツイートであっても抽出されてしまうことがあるため，質的な評価の重みを和らげるよう対数を取る．

$$P_{repost}(d) = \log(1 + n_{reposte}(d)) \quad \textbf{（式 7）}$$

$$P_{followers}(d) = \log\left(1 + n_{followers}(d)\right) \quad \textbf{（式 8）}$$

$$P_{recency}(d) = e^{-\gamma \cdot (\mathrm{t} - t_d)} \quad \textbf{（式 9）}$$

$$Score_{quality}(d) = \log\left(1 + \frac{P_{reposts}(d) + P_{followers}(d) + P_{recency}(d)}{3}\right) \quad \textbf{（式 10）}$$

3.4. ツイート検索

3.2 項で拡張したクエリを用いてツイート検索を行う．ランキングには，ツイートに含まれる拡張クエリのスコアの総和とツイート自体の質評価を組み合わせる．

ツイート d とクエリ q の関連度を$Score_{\mathrm{query}}(q, d)$とする．$Score_{\mathrm{query}}(q, d)$はクエリ q の拡張クエリ$EQ_q$を用いて，（式 11），（式 12）のように表す．

$$\mathrm{Score}_{\mathrm{query}}(q, d) = \sum_{eq \in EQ_q} nScore_{expansion}(eq, q) \cdot \Phi(\mathrm{eq}|\mathrm{d}) \quad \textbf{（式 11）}$$

$$\Phi(\mathrm{eq}|\mathrm{d}) = \begin{cases} 1 & ,if\ eq \in d \\ 0 & ,otherwise \end{cases} \quad \textbf{（式 12）}$$

最後に，全てのツイートに対して，（式 11）より得られたクエリとの関連性評価と，（式 10）より得られた

ツイート自体の質的評価を組み合わせたスコアリングを行う．これを（式 13）で表す．

$$\mathrm{Score}_{total}(q,d) = \mathrm{Score}_{query}(q,d) \cdot Score_{quality}(d) \qquad \textbf{（式 13）}$$

4. 実験と評価

4.1. 使用データ･実験環境

本実験では形態素解析エンジンに lucene-gosen[7]を用いた．Twitter 中に多く用いられる新語や人名，インターネット用語に対応するため，IPA 辞書に加えてはてなキーワード[8]を利用し，文章中の名詞，動詞，形容詞の原型と，ハッシュタグ，URL を抽出した．表記揺れの対策のため，全ての文字の表記は事前に半角小文字，全角カタカナに統一した．

2012 年 12 月 26 日～31 日にかけて，Twitter Streaming API を用いてツイート収集し，そのうち平仮名か片仮名を 1 文字以上含む 336 万ツイートを長期ツイートとして実験を行った．また，検索タイミングを 2013 年 1 月 1 日 0 時 0 分とし，長期ツイートのうち最後 3 時間に投稿された 20 万ツイートを短期ツイートとした．長期ツイートに含まれた文書中の要素（名詞，動詞の原型，形容詞の原型，ハッシュタグ，URL）はおよそ 44 万種類であった．

まず，「NHK」と「AKB48」の 2 つを検索クエリとして実験を行った．各実験パラメータについては，既存研究[3]に習い，拡張クエリ数 k=10，拡張クエリの最低出現回数 λ=20，出現頻度パラメータ α=100 とする．また，長期ツイートにおける最低共起回数 μ=3 とした．検索結果の評価には，上位 10 ツイートを利用した．また，$nScore_{expansion}$には，正規化した拡張クエリのスコアの総和に加えて，元クエリのスコアを 1.0 として加算したものを使用した．

4.2. クエリ拡張結果

3.2 項の式により得られた拡張クエリを表 2 に示す．

表 2　検索クエリが「NHK」の場合の拡張クエリ

	Massoudi らの手法[3]		提案手法	
順位	拡張クエリ	スコア	拡張クエリ	スコア
1	nhk 紅白	0.17	紙吹雪	0.12
2	紅白	0.12	ヨイトマケ	0.12
3	（NHK 紅白歌合戦の画像 URL）	0.12	北島三郎	0.11
4	キャプチャ	0.12	サブちゃん	0.11
5	ゴールデンボンバー	0.09	ゆく年くる年	0.10
6	伝える	0.08	美輪	0.09
7	秒	0.08	由紀さおり	0.09
8	全て	0.07	プリプリ	0.09
9	見れる	0.07	斉藤和義	0.09
10	紅白歌合戦	0.07	天童よしみ	0.09

表 3　検索クエリが「AKB48」の場合の拡張クエリ

	Massoudi らの手法[3]		提案手法	
順位	拡張クエリ	スコア	拡張クエリ	スコア
1	autofollowjp	0.15	紅白歌合戦	0.20
2	smap	0.14	紅白	0.19
3	野球	0.14	ギンガムチェック	0.15
4	followback	0.12	今年	0.09
5	相互フォロー	0.12	ももクロ	0.06
6	ラーメン	0.09	島崎遥香	0.06
7	rt	0.09	素敵	0.05
8	初心者	0.06	幸せ	0.05
9	楽天	0.05	日付	0.05
10	方法	0.05	ステージ	0.04

4.3. 検索結果と評価

2 つのクエリによる検索結果を 表 4 に示す．評価として MAP と nDCG，適合ツイート数を示す．今回，検索結果のツイートがクエリに適合しているかは人手で確認した．

表 4　検索結果の評価

クエリ	NHK		AKB48	
	既存手法	提案手法	既存手法	提案手法
MAP@10	1.0	1.0	0.0	1.0
nDCG@10	1.0	1.0	0.0	1.0
適合ツイート数	10	10	0	10

フォロワー数稼ぎなどに悪用されにくいクエリ「NHK」については，検索精度は同じであった．しかし，頻繁にスパムツイートに利用される「AKB48」という語句をクエリにした場合，既存手法では適合ツイートを上位に検索出来なかったのに対し，提案手法では時間的に関係のある適合ツイートを抽出できた．

また，2012 年 12 月 31 日 22 時，23 時，24 時の時点で Twitter のトレンドに入っていた 30 個のトレンドから重複を除き，実験に利用可能である 18 個のトレンドを検索クエリとして実験を行った．18 のトレンドを表 6 に示す．拡張クエリのスコアの総和に，元クエリのスコアを 0.0，0.25，0.50，0.75，1.0 と変えて拡張クエリに加えた場合の 4 種類の実験を行った．この実験結

表 5　18 トレンドを対象として元クエリのスコアを変化させた場合の評価

元クエリのスコア	0.0		0.25		0.5		0.75		1.0	
	既存手法	提案手法	既存手法	提案手法	既存手法	提案手法	既存手法	提案手法	既存手法	提案手法
MAP@10	**0.80**	0.72	**0.89**	0.87	0.92	**0.92**	0.92	**0.93**	0.92	**0.95**
nDCG@10	**0.84**	0.77	**0.92**	0.90	0.94	**0.94**	0.94	**0.96**	0.94	**0.97**
平均適合ツイート数	**6.89**	6.56	**8.28**	8.11	8.89	**9.11**	8.89	**9.17**	8.89	**9.28**

果を表 5 に示す. 実験の結果, 元クエリのスコアを 0.5 以上とした場合に，最も時間的に適した検索を行えることが分かった.

表 6　実験を行った 18 トレンド

白組優勝	gakisp	ガキ使	猫物語	紙吹雪	nhk 紅白
歌合戦	年越し	misia	アニソン	nhk	石川さゆり
大晦日	フォロー	2012 年	SKE	スマホ	レストラン

既存手法では普遍的な関連語句も拡張クエリとするため，拡張クエリだけで検索した場合にも多くのツイートを見つけることができるが，同時にスパムツイートのような関連性の低いツイートも取得してしまう. 一方，提案手法では，時間的な関連語句を強く重視して拡張クエリとする. そのため拡張クエリのみで検索した場合，テレビ番組など時間的な要素の強い検索クエリに対しては高い精度で検索が可能になるものの，時間的な側面の少ない語句を検索クエリとした場合には，極端な拡張クエリが得られてしまうことから，本来の検索意図から外れた結果となる場合がある. しかし，拡張クエリのスコアだけでなく元クエリにもスコアを与えることにより，本来の検索意図を残しつつ，より時間的に関連度の高い話題を見つけることが可能になった.

5. まとめ

本稿では，経時的な関連語句の変化を考慮したクエリ拡張による Twitter 検索手法を提案し, 評価を行った. 既存手法では，スパムツイートに含まれやすい語句，例えば芸能人や商品名などを検索クエリとした場合，スパムツイート内にて共起する語句の影響を受け，適切なクエリ拡張が行えないという問題があった. 提案手法はクエリ拡張において，拡張クエリと元クエリとの共起頻度，拡張クエリ単体での出現頻度を，それぞれ検索タイミング付近の期間と全体の期間について求めることで，最近の話題を考慮しつつ，スパムに影響されないクエリ拡張を行った. 実験の結果，18 クエリの検索の平均適合ツイート数を, 既存手法の 8.89 から, 提案手法では 9.28 と増やしつつ, スパムツイートに含まれやすいクエリにおいて，既存手法では 1 つも適切なツイートを検索できなかったのに対し，提案手法では上位 10 件全て適切なツイートを検索することが可能になった. 以上より，クエリ拡張結果とツイートの品質評価を組み合わせることにより，より幅広く，元のクエリに関係するツイートを取得できた.

今後の課題として，検索タイミング付近の共起頻度比率や出現頻度比率を計算するにあたって，どの程度の期間に設定するのが適切なのか，また出現頻度パラメータの設定について検討したい.

参　考　文　献

[1] Semiocast — Twitter reaches half a billion accounts — More than 140 millions in thgae U.S.: http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. （2013 年 1 月 5 日アクセス）

[2] Twitter: https://twitter.com/ （2013 年 1 月 5 日アクセス）

[3] K. Massoudi, M. Tsagkias, M. D. Rijke and W. Weerkamp: “Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts,” In *Proc. of the 33rd European Conference on Advances in Information Retrieval (ECIR)*, 2011.

[4] M. Efron, P. Organisciak and K. Fenlon: “Improving Retrieval of Short Text Through Document Expansion,” In *Proc. of the 35th ACM International Conference on Special Interest Group on Information Retrieval (SIGIR)*, 2012.

[5] X. Zhang, B. He, T. Luo and B. Li: “Query-biased Learning to Rank for Real-time Twitter Search,” In *Proc. of the 21th ACM International Conference on Information and Knowledge Management (CIKM)*, 2012.

[6] J. Teevan, D. Ramage and M. R. Morris: “#TwitterSearch: A Comparison of Microblog Search and Web Search,” In *Proc. of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.

[7] lucene-gosen: http://code.google.com/p/lucene-gosen/ （2013 年 1 月 5 日アクセス）

[8] はてなキーワード一覧ファイル - Hatena Developer Center: http://developer.hatena.ne.jp/ja/documents/keyword/misc/catalog （2013 年 1 月 5 日アクセス）