情報処理学会論文誌 Vol.54 No.4 1655–1666 (Apr. 2013)

マイクロブログ検索のための時間情報と非時間情報を統合したクエリ拡張

宮西 大樹[1,a)] 関 和広[1,b)] 上原 邦昭[1,c)]

受付日 2012年6月21日, 採録日 2013年1月11日

概要：本研究では，順位学習によって時間情報と非時間情報を統合したクエリ拡張をマイクロブログ検索に適用する．1つ目の時間情報を用いたクエリ拡張手法には，時間の新近性と話題の時間変化を取り入れる．マイクロブログ検索では最近起こった出来事について調べることが多いため，新鮮な文書やそれに出現する単語ほどユーザの意図に適合しやすくなる新近性の特徴を持つ．また，話題に関連する文書や単語が話題の盛り上がった時期に出現しやすくなる．2つ目のクエリ拡張手法では，クエリと単語の検索エンジンのヒット数を基にした統計情報から，時間情報に依存せずに話題の関連語を同定する．そして，両手法の利点を活かすため，順位学習の枠組みを用いて複数のクエリ拡張手法を統合し，学習器が予測した話題の関連語でクエリ拡張を行う．Tweets2011 コーパスを用いた実験により，時間情報を用いたクエリ拡張がマイクロブログ検索に有効であり，時間情報を用いないクエリ拡張手法と統合することで，検索性能をさらに向上させることができることを示す．

キーワード：リアルタイム検索，マイクロブログ，Twitter，クエリ拡張

Synthesizing Temporal and Atemporal Query Expansion for Microblog Search

Taiki Miyanishi[1,a)] Kazuhiro Seki[1,b)] Kuniaki Uehara[1,c)]

Received: June 21, 2012, Accepted: January 11, 2013

Abstract: Topics on recent news and events (e.g., the venue for 2022 FIFA World Cup) are actively mentioned on microblog service, where their related terms (e.g., soccer and Qatar) are likely to be mentioned together during the same period. This intuition suggests that the temporal variation of a topic of interest can be utilized to discover terms related to the topic. Based on the idea, we propose a query expansion method that takes advantage of temporal properties of terms to measure the association between a term and a given seed query. To represent those temporal properties, our proposed method uses a temporal profile constructed from the time-stamps of documents (tweets in this study) retrieved by a search engine. We carefully design and examine different temporal profiles resulting from different query formulations and demonstrate that the combination of a seed query and a topic-related term tends to be similar to the temporal profile of a given topic. In addition, we present atemporal query expansion methods based on modified web similarity to deal with time insensitive topics. The temporal and atemporal query expansion methods are combined using a learning to rank model to compensate for the limitations of individual methods. Experiments on the Tweets2011 corpus show that the use of temporal properties is effective in query expansion and the fusion of the temporal and atemporal properties significantly improves retrieval performance.

Keywords: real-time search, microblog, Twitter, query expansion

1. はじめに

マイクロブログは，今，世界で何が起きているかを簡単に知るために広く使用されているサービスである．その数あるマイクロブログサービスの中でも，Twitter は最

[1] 神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University, Kobe 657–8501, Japan
[a)] miyanishi@ai.cs.kobe-u.ac.jp
[b)] seki@cs.kobe-u.ac.jp
[c)] uehara@kobe-u.ac.jp

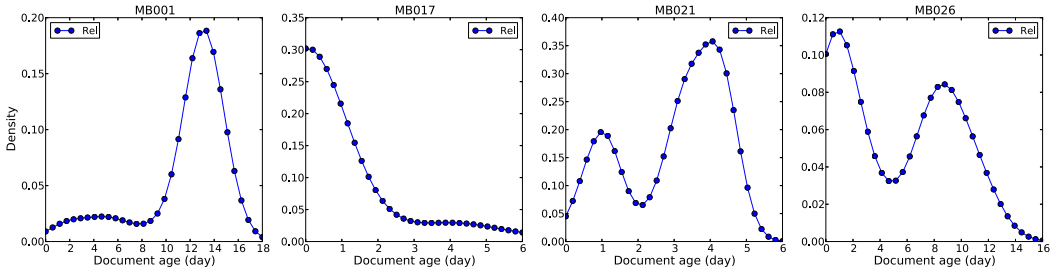


図 2 4 つのトピック（MB001，17，21，26）に関連する tweet の時間変化のパターン
Fig. 2 The temporal variations of four topics (MB001, MB017, MB021, and MB026) from TREC 2011 Microblog Track based on relevant tweets.

<num>	MB001
<title>	BBC World Service staff cuts
<querytime>	Tue Feb 08 12:30:27 +0000 2011

図 1 TREC 2011 マイクロブログトラックで用いられたトピックの例
Fig. 1 Example topic from the TREC 2011 Microblog Track.

も有名なソーシャルネットワークサービスの 1 つである．Twitter 上では，ユーザ同士がコミュニケーションを行ったり，ある話題について自身の意見を述べたりするために，1 日で約 3 億 4 千万[*1]の tweet（Twitter のユーザが作成する文書）が作成されている [9], [11]．Twitter の興味深い特性として，実社会で人々が注目するような出来事が起これば，その時間に大量の tweet が多くの人によって作成される傾向がある．その結果，時間ごとの tweet の集合が，特定の時間に何の話題が注目されているかを知るための重要な手がかりとなる．たとえば，BBC World Service が 5 つの言語サービスを止める予定であるというニュースが 2011 年 1 月 25 日～27 日に報道されると，同時期にその話題に適合すると見なされた tweet が多数投稿された．このマイクロブログ上での性質を，TREC 2011 のマイクロブログトラック [19] で使用された 4 つの情報要求（トピック），“BBC World Service staff cuts（MB001）”，“White Stripes breakup（MB017）”，“Emanuel residency court rulings（MB021）”，“US unemployment（MB026）”とトピックに対してユーザの意図に適合する一連の tweet を話題と見なして説明する．ここで，トピックの例を**図 1** に示す．図中の $\langle num \rangle$ がトピックの番号，$\langle title \rangle$ がユーザクエリ，$\langle querytime \rangle$ がクエリの発行された時間（クエリ時間）を示している．**図 2** は各トピックのクエリ時間からトピックに適合する tweet のタイムスタンプを差し引いた値（文書年齢）の分布をカーネル密度推定した結果である．横軸が文書年齢（日数），縦軸が文書年齢のカーネル密度の推定値を表す．横軸の 0 がクエリ時間であり，縦軸の推定値が高いほど文書がある時間に密集して（話題が盛り上がって）いることを表す．図より，ある特定の時間帯においてトピックに適合する tweet[*2]が多く発行されており，トピックごとに話題の時間変化のパターンが異なっていることが確認できる．さらに，トピックに適合する tweet の中には話題に関連する単語が含まれやすい．たとえば，MB001 の場合ではクエリ語の「BBC」，「cuts」，「staff」とクエリに意味的に関連する単語「axe」，「jobs」は同時期に複数の tweet に現れやすくなると考えられる．このことから，いつ話題が活発に言及されるかという話題の時間変化を予測できれば，話題に適合する文書や関連語を容易に同定できると考えられる．

このアイデアに基づいて，Efron [6] が提案した tweet のタイムスタンプを利用した擬似関連フィードバックの発想を独自に拡張し，マイクロブログ検索のための時間情報を考慮したクエリ拡張手法を提案する．クエリ拡張とは，ユーザの検索意図を補完するため，検索エンジンに入力したキーワード（ユーザクエリ）と意味的に関連する単語やフレーズ（以下，関連語もしくは意味的に関連する語と表記する）を元のクエリに加えて新しいクエリとして再検索を行う検索技術のことである [15]．通常のキーワード検索では，ユーザクエリ中の単語が文書中に出現する文書しか検索できない．そして，マイクロブログでは，1 つの文書（本研究では tweet）の長さがウェブページなど通常の文書と比べてきわめて短いため，ユーザの入力したクエリ語が文書中に出現しにくい．そこで，クエリに対して意味的に関連する語をユーザクエリに加えて再検索を行うクエリ拡張を用いれば，ユーザの意図に適合する文書（適合文書）をより多く検索できると考えられる．そして，機械学習の枠組みを用いて時間情報に依存しないクエリ拡張手法と上記の手法を組み合わせることで，さらなる検索精度の向上を目指す．これまで，機械学習をマイクロブログ検索に適用した方法として，tweet の特徴を基にして tweet 自身の再順位付けを行う研究がある [4], [17], [18]．このほかにも，所与のクエリに適合し，かつ新鮮な文書を検索するため，文書の新鮮さを表す新近性を取り入れた検索モデルが提案されている [6], [16]．新近性を考慮した方法は，図 2 の MB017 や MB026 のように適合文書がクエリ時間の近くに

[*1] http://blog.twitter.com/2012/03/twitter-turns-six.html

[*2] 図 2 の場合と同様に，手動でトピックに適合する見なされた文書（tweet）のこと．

存在するトピックに対して有効だと考えられる．また，新近性を言語モデルに取り入れることで検索精度を向上させる試みもある [3], [5], [13]．これらの手法は検索精度を向上させることに成功しているものの，話題の時間変化を順位学習の素性やモデルに組み込んでいないため，話題が盛り上がった時期に応じて，話題の適合文書や関連語を同定することは困難である．さらに，新近性のみを考慮したマイクロブログの検索方法では，図 2 の MB001 のようにクエリ時間から離れたところに適合文書が集中して（話題が盛り上がって）いたり，MB021 のように複数盛り上がりのあったりするトピックに対して有効に機能しない．

本研究では，新近性や多種多様な時間変化を持つ話題の関連語を予測し，これをクエリ拡張手法に応用する．また，時間情報を考慮したクエリ拡張手法を Web 類似度 [1], [2] を用いた非時間情報による手法と組み合わせることで，各手法の欠点を克服する．さらに，本手法を用いれば，いつ話題の関連語が注目されたか，といった時間的な解析も可能となる．最後に，約 1,600 万件の tweet からなる Tweets2011 コーパス[*3]を用いた実験により，本手法を用いることで検索精度の向上が可能であることを示す．

2. 時間情報を取り入れたマイクロブログ検索の従来手法

マイクロブログ上では，ユーザは最近起こった話題に関する文書を検索することが多い．そのため，図 2 のトピック MB017 や MB026 のように，最近の話題に関連する文書はクエリ時間付近に存在することが多くなる．Massoudi ら [16] はこのクエリと文書間に存在する新近性の特徴を生かして，次式に示す新近性を考慮したクエリ拡張手法を提案している．

$$S_{IRQE}(w,Q)=\log\frac{|\mathbb{N}_c|}{\{|d:w\in d|\}}\sum_{\{d:q\in Q\ ,\ w,q\in d\}}e^{-\beta(c-c_d)} \quad (1)$$

ここで，c はクエリ時間，c_d は文書 $d\in\mathbb{N}_c$ の作成された時間，$\mathbb{N}_c$ はクエリ時間以前に発行された文書集合である．また，β はクエリ拡張の候補の単語（候補単語）w を含む文書の新近性をどの程度考慮するかを決定するパラメータである．クエリ拡張には所与のクエリに対して，上記スコアの上位 K 個の単語を加えることで行う．このクエリ拡張手法では，クエリ Q に関連する単語はクエリ時間 c 付近に発行された文書に出現しやすいという仮定に基づいている．Twitter のデータを用いた実験により，このクエリ拡張手法を用いることで，既存の時間情報を考慮しないクエリ拡張手法 [12] よりも検索精度を向上させることができることが分かっている．

同様に，Efron [6] は時間情報，特に新近性と平滑化した話題の時間変化を取り入れたマイクロブログの検索手法を

[*3] http://trec.nist.gov/data/tweets/

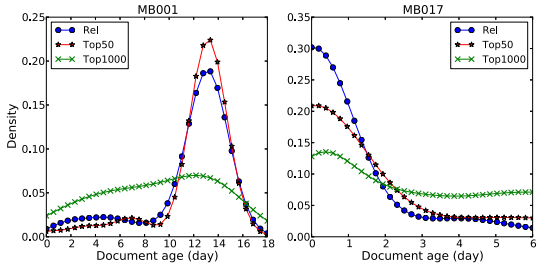


図 3 トピック MB001 と MB017 に対応する tweet の文書年齢を基にしたカーネル密度の推定値

Fig. 3 Two kernel density estimates corresponding to the topics MB001 and MB017.

提案している．この手法では，時間プロファイル [10] と呼ばれる検索エンジンで取得した文書の時間ごとの文書頻度の分布を作成する．この分布を見ることで，クエリ語を含む文書が多く集中している時間帯を知ることができるので，時間プロファイルを擬似的な話題の時間変化と見なすことができる．そして，適合文書の時間プロファイル（適合プロファイル）と所与のクエリで検索した上位の文書からなる時間プロファイル（クエリプロファイル）のカルバック・ライブラ情報量（KL 距離）は小さく，クエリプロファイルと適合しない文書の時間プロファイル（非適合プロファイル）の KL 距離は大きいという時間的な特性を用いて文書拡張を行っている．Efron の発想を分かりやすく説明するため，3 種類の tweet（図 2 と同じトピックの適合文書，検索エンジンを用いて検索した上位 50 件と上位 1,000 件の tweet）を用いて作成した 3 つの時間プロファイル（Rel, Top50, Top1000）の文書年齢に関するカーネル密度の推定値を**図 3** に示す．図中の青線（Rel）は適合 tweet に基づく推定値，赤線（Top50）と緑線（Top1000）は，それぞれ検索エンジンで検索した上位 50 件と 1,000 件の tweet に基づく推定値に対応する．MB001 と MB017 の上位 50 件の Precision はそれぞれ 0.74 と 0.36 であり，上位 1,000 件の Precision は 0.061 と 0.064 なので，Rel，Top50，Top1000 はそれぞれ適合プロファイル，クエリプロファイル，非適合プロファイルに相当する．図より，適合プロファイル Rel は非適合プロファイル Top1000 よりクエリプロファイル Top50 に似ていることが確認でき，クエリプロファイルは関連プロファイルの近似と見なすことができる．Efron はこの観察を基に，下記の式に従うスコアの降順に文書を再順位付けする擬似関連フィードバックを提案している．

$$s(D,Q)=\log Pr(Q|D)+\phi(T_Q,T_D) \quad (2)$$

ここで，$\phi(T_Q,T_D)=\log(\frac{m_{T_Q}}{m_{T_D}})$ であり，m_{T_Q} はクエリ Q で検索した文書の文書年齢の平均，m_{T_D} は Q で検索した上位の文書 D を擬似的なクエリとして再検索した文書の文書年齢の平均である．サンプル平均 m_{T_D} が小さい文書，つまり新しい文書を検索する擬似クエリ D ほどスコアが高くなり，D が古い文書を検索するほど D のスコアは小さ

くなる．ただし，クエリ Q が古い文書を多く検索するようであれば，文書 D による m_{T_D} の新近性の効果は薄れる．

この手法は，文書を擬似クエリとして取得した時間プロファイル（文書プロファイル）がクエリプロファイルと類似していれば，その文書は話題と関連していると仮定している．しかし，Efron のモデルでは与えられたトピックに関連する単語を同定することはできない．また，各時間プロファイル中のタイムスタンプの分布が正規分布に従うことを仮定しており，図 2 のトピック MB021 や MB026 のような多峰性の時間プロファイルを適切に処理できない．そこで，3.3 節において，多峰性の時間プロファイルを処理でき，時間的に変化する話題の関連語を同定できるクエリ拡張手法を提案する．さらに，3.4 節において時間の新近性を考慮したクエリ拡張手法について述べ，3.5 節で検索エンジンを用いた時間情報に依存しないクエリ拡張手法について紹介する．そして，3.6 節でこれら時間情報と非時間情報を利用したクエリ拡張手法を機械学習の枠組みを用いて統合する方法について述べる．

3. 提案手法

本章では，時間情報と非時間情報を利用したクエリ拡張方法について議論する．本研究で提案するマイクロブログ検索のクエリ拡張手法の手順を以下に示す．また，提案手法の各手順に対応する流れ図を**図 4** に示す．

(1) 所与のクエリ（初期クエリ）で検索した tweet からタイムスタンプと単語の頻度を取得して，時間プロファイルを作成する．
(2) 検索スコアの高い上位 M 件の tweet からクエリ拡張の候補となる単語を選ぶ．
(3) 初期クエリに拡張単語を加えて新たなクエリ（拡張クエリ）として再検索する．検索して得られた tweet を基に時間プロファイルを作成する．
(4) 取得した時間プロファイルと単語の頻度を，Web 類似度に基づく手法，新近性に基づく手法，時間変化の類似度に基づく手法の 3 種類のクエリ拡張手法に用いる．
(5) 3 種類のクエリ拡張手法を順位学習を用いて結合し，初期クエリの関連単語を予測する．
(6) 順位学習によって予測したクエリとの意味的関連度が高い上位 K 件の単語を初期クエリに加えて拡張クエリとし，tweet を再検索する（最終的な検索結果からは retweet [*4] を除く）．

ここで，TREC 2011 Microblog track の公式ルールに従い，retweet は非適合文書であると見なす．しかし，retweet の中には話題と適合するものもあり，クエリ拡張に有効な単語を含む場合がある．そこで，上記の最後のステップ以外，retweet を用いることにする．なお，本クエリ拡張手法は

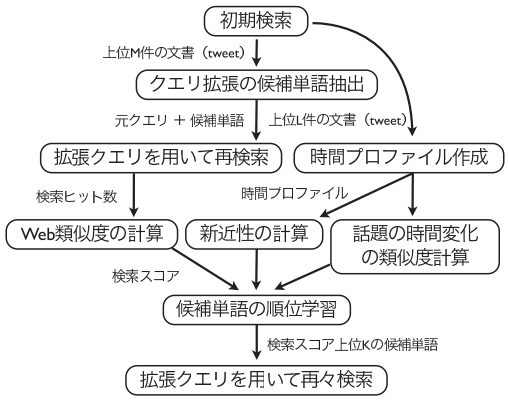


図 4 提案するクエリ拡張の流れ図
Fig. 4 Flowchart of our proposed query expansion method.

一種の擬似関連フィードバック [12] と見なすことができる．本稿で提案する擬似関連フィードバックでは，初期検索の結果の上位にある文書のタイムスタンプ，検索スコア（クエリ尤度）とその文書中の単語しか用いない．よって，Twitter などの実システムで巨大な tweet コレクションに対して本手法を利用する場合も，初期検索に要する時間が変わらない限り，実時間でクエリ拡張を行うことができる．

3.1 クエリ拡張のための時間プロファイル

本節では，話題の関連語を初期クエリに加えてクエリ拡張を行う方法を提案する．図 3 より，クエリプロファイル（Top50）は話題の時間変化を表す適合プロファイル（Rel）をうまく近似できることが分かった．そこで，著者らは初期クエリのクエリプロファイルと初期クエリに候補単語を加えた拡張クエリの時間プロファイル（拡張クエリプロファイル）とを比較することで，話題の関連語を同定できると考えた．拡張クエリの作成方法は複数考えられるので，ここでは，以下の 3 種類の検索方法を用いて拡張クエリプロファイルを作成して比較する．

- **Method 1** 候補単語 1 つを拡張クエリとして検索．
- **Method 2** クエリ語を少なくとも 1 つか候補単語を含む拡張クエリを用いて検索．
- **Method 3** クエリ語を少なくとも 1 つと候補単語の双方を含む拡張クエリを用いて検索．

これらの検索方法から，関連語と非関連語を識別するための適切な時間プロファイルの作成方法を選択するため，トピック “Taco Bell filling lawsuit（MB020）” を初期クエリとして，このトピックのと 3 つの候補単語「beef」,「meat」,「rt」に関する拡張クエリプロファイルを用いて予備実験を行う．このトピックは 2011 年 1 月末日に報道されたタコベルの人工牛肉（Taco Meat Filling）の訴訟[*5] に関する tweet を検索せよというものであり，候補単語「beef」と「meat」が話題の関連語と見なせる．一方，候補単語「rt」

[*4] 情報共有のため他のユーザの tweet を引用して自分の tweet とし再発行するための機能．
[*5] http://gizmodo.com/5742413/

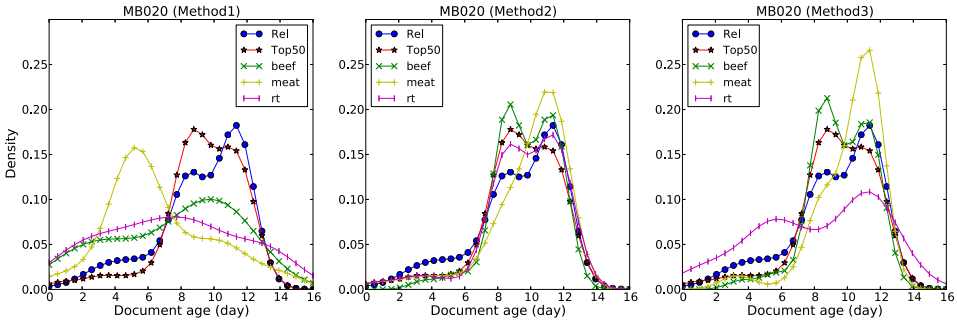


図 5 トピック MB020（Taco Bell filling lawsuit）に関して，3 種類の検索方法（Method 1, 2, 3）で作成した時間プロファイルの文書年齢をカーネル密度推定した推定値

Fig. 5 Three types of kernel density estimates using topic MB020 (Taco Bell filling lawsuit) based on the tweets retrieved by three methods (Method 1, 2, 3).

（retweet を表す文字列）は Twitter 上で広く用いられる一般的な単語であり，特定の話題に関連する単語ではない．各手法によって得られた時間プロファイルを図 5 に示す．図中の緑，黄，紫線はそれぞれ候補単語「beef」，「meat」，「rt」に対応する拡張プロファイルの文書年齢に関する推定値であり，Top50 と Rel は検索した上位 50 件の tweet と適合 tweet に関する推定値である．前述のトピックと同様に，このトピックにおいても適合プロファイル（Rel）とクエリプロファイル（Top50）が類似していることが確認できる．**Method 1** に対応する左のプロットでは，「beef」の拡張クエリプロファイルがクエリプロファイルよりも「rt」の拡張クエリプロファイルに似てしまっている．さらに，「meat」という単語が他の非適合文書にも多く出現するため，meat の拡張クエリプロファイルがクエリプロファイルと離れており，関連語の同定が困難であることが分かる．一方，**Method 2** に対応する中央のプロットでは，すべての時間プロファイルがクエリプロファイルに類似しており，拡張クエリプロファイルを用いて，話題に対して関連する語と関連しない語を識別することができない．最後に，**Method 3** に対応する右のプロットでは，「beef」と「meat」の拡張クエリプロファイルがクエリプロファイルと類似しており，非関連語「rt」の拡張クエリプロファイルはクエリプロファイルと類似していないことが分かる．この結果から，**Method 3** を用いることで，少なくともこのトピックにおいては，関連語と非関連語を識別可能な拡張クエリプロファイルを作成できることが分かる．そこで，拡張クエリプロファイルの作成には以後 **Method 3** を用いる．

3.2 時間分布のモデル化

初期クエリと候補単語の時間的性質をモデル化するため，Jones ら [10] のモデルを修正して用いる．最初に，クエリ Q を用いて取得できる特定の日 t の文書の頻度を表す分布 $P(t|Q)$ を以下に定義する．

$$P'(t|Q) = \sum_{D \in R} P(t|D) \frac{P(Q|D)}{\sum_{D' \in R} P(Q|D')} \tag{3}$$

ここで，R はクエリ Q を用いて検索エンジンから取得した上位 M 件の文書であり，$P(t|D) = 1$ は t と文書 D の作成日が同じであれば，$P(t|D) = 0$ となる．$P(Q|D)$ は文書 D のクエリ Q に対する関連度である．また，コーパス中でクエリに適合する文書の各日における文書頻度の不規則性に対処するため，以下のように平滑化を行う．

$$P(t|Q) = \lambda P'(t|Q) + (1 - \lambda) P(t|C) \tag{4}$$

ここで，コーパス C 中の文書を D で表し，コレクションの時間モデルを $P(t|C) = \frac{1}{|C|} \sum_{D \in C} P(t|D)$ として定義する．パラメータは λ は従来手法 [10] に従い 0.9 と定め，$P(t|Q)$ をクエリの時間モデルとする．従来手法では，各隣接する日どうしの確率でクエリの時間モデルを平滑化しているのに対し，マイクロブログ検索では日ごとの単語の使用頻度が重要になるため，元のモデルを修正して平滑化しないモデルを用いる．この影響により参考文献中 [11] のクエリの時間モデル $\tilde{P}(t|Q)$ ではなく，$P(t|Q)$ を本文中では使用する．また，このクエリの時間モデルは各日の単語頻度の時間的変化を表しているため，多峰性の時間プロファイルを表現できる．

3.3 時間変化の類似度

クエリの時間モデルとコレクションの時間モデルとの離れ具合いをとらえるため，Jones らは下記のように両者の KL 距離を計算した．

$$D_{KL}(P(t|Q), P(t|C)) = \sum_{t=1}^{T} P(t|Q) \log \frac{P(t|Q)}{P(t|C)}. \tag{5}$$

著者らはこのモデルの修正を行い，クエリプロファイルと拡張クエリプロファイルの離れ具合いを計算することで話題の関連語を同定する方法を提案する．この考えは 3.1 節の図 5（右プロット）で示した関連語を含む拡張クエリの時間プロファイルはクエリプロファイルに似ており，

非関連語を含む拡張クエリの時間プロファイルはクエリプロファイルに似ていないという性質から着想を得ている．つまり，クエリ語を含む文書と同時期に使用される単語を元のクエリと意味的に関連する語と見なすことができる．クエリの時間モデルを用いた候補単語の選定は以下の式に従うスコアを用いて行う．

$$\begin{aligned} S_{TSQE}(w,Q) &= -D_{KL}(P(t|w \cap^{+} Q), P(t|Q)) \\ &= -\sum_{t=1}^{T} P(t|w \cap^{+} Q) \log \frac{P(t|w \cap^{+} Q)}{P(t|Q)} \end{aligned}$$

ここで，3.1 節の **Method 3** に従い，拡張クエリプロファイルの作成のために所与のクエリ語を少なくとも 1 つと候補単語の双方を含む “$w \cap^{+} Q$” を拡張クエリの表現として用いる．初期クエリの時間モデルと拡張クエリの時間モデルの KL 距離が小さければ，拡張クエリに用いた候補単語を話題の関連語とする．さらに，このモデルはクエリ語を含む日ごとの文書数の分布を見ているので，多峰性の時間プロファイルを表現でき，話題の任意の時間変化をモデル化することができる．ただし，新近性は考慮できない．

3.4 時間の新近性

次に，式 (2) に示す Efron のモデルに着想を得て，新近性を取り入れたクエリ拡張手法を提案する．クエリ拡張に用いる各候補単語のスコアを以下に示す．

$$S_{TRQE}(w,Q) = \phi(T_Q, T_{Q'}) = \log\left(\frac{m_{T_Q}}{m_{T_{Q'}}}\right) \tag{6}$$

ここで，$m_{TQ'}$ は上位 L 件の検索結果を基にした拡張クエリプロファイルの文書年齢の平均である．2 章で紹介した Efron のモデルは文書自体にスコア付けを行うため，話題の関連語は同定できない．しかし，このモデルは単語にスコア付けを行うため，クエリ時間付近の文書に出現する話題の関連語をクエリ拡張の候補単語として同定できる．

3.5 Web 類似度

本節では，話題の時間変化を予測しにくいトピックや時間的な性質と関わりのないトピックに対処するため，時間情報を用いない話題の関連語の同定方法を紹介する．クエリと候補単語との関連度を計算する方法として，語彙統語パターンの抽出に用いられた Web 類似度 [1] を使用する．Web 類似度は，検索ヒット数を基にした類似度であり，クエリ表現 “$w \cap Q$” を使った検索ヒット数を Web 上での Q と w の共起数の近似として利用する．このように算出した Web 類似度は，時間情報とは無関係にクエリと単語・フレーズ間の意味的類似度を計算できるため，時間に依存するクエリ拡張手法とは異なる意味的関連語を予測できると考えられる．しかし，Twitter 上では，この共起数の近似を直接関連語の同定に用いることはふさわしくない．なぜなら，1 つの tweet に書き込める内容は 140 文字と制限されており，初期クエリと候補の単語が tweet 内で同時に出現することは考えにくいためである．たとえば，Twitter のユーザは “FIFA qatar” や “2022 soccer qatar” といった短く不完全な表現を用いる傾向にあるため，初期クエリ “2022 FIFA soccer” を拡張したクエリ “2022 FIFA soccer qatar” 中の単語すべてが tweet 中に出現することは考えにくい．この問題に対処するため，従来のクエリ Q 中の単語すべてと候補単語 w のすべてを含むクエリ表現 “$w \cap Q$” の代わりに，クエリ Q 中の単語を少なくとも 1 つと候補単語 w の双方を含む “$w \cap^{+} Q$” を新たなクエリ表現として使用する．このクエリ表現は，3.1 節，3.3 節，3.4 節での拡張クエリプロファイル作成のときも使用しており，これを既存の Web 類似度にも適用する．

修正した類似度の定義に使う表記 $H(Q)$ はクエリ Q 中の単語を少なくとも 1 つ含む文書数を表す．改変した単語 w とクエリ Q の Web 類似度の WebJaccard^{+} 係数を以下に示す．

$$\text{WebJaccard}^{+}(w,Q) = \frac{H(w \cap^{+} Q)}{H(w) + H(Q) - H(w \cap^{+} Q)}. \tag{7}$$

ここで，候補単語 w とクエリ Q 中の単語が偶発的にいくつかの tweet に出現することがあるため，すべての Web 類似度はヒット件数 $H(w \cap^{+} Q)$ がある閾値 c 以下であれば 0 に設定する（後の実験では $c = 4$ とした）．同様に，WebOverlap^{+}，WebDice^{+}，WebPMI^{+} を以下のように定義する．

$$\text{WebOverlap}^{+}(w,Q) = \frac{H(w \cap^{+} Q)}{\min(H(w), H(Q))} \tag{8}$$

$$\text{WebDice}^{+}(w,Q) = \frac{2H(w \cap^{+} Q)}{H(w) + H(Q)} \tag{9}$$

$$\text{WebPMI}^{+}(w,Q) = \log_2 \frac{H(w \cap^{+} Q)/N}{H(w)/N \cdot H(Q)/N} \tag{10}$$

ここで，N は検索エンジンによって索引付けされた文書数であり，本研究ではコーパス中の tweet 数に従い，$N = 16,141,812$ と定める．

クエリに関連する単語を同定するため，他の Web 類似度として Normalized Google Distance（NGD）[2] も用いる．他の Web 類似度と同様に，元の NGD の $H(w \cap Q)$ の部分を $H(w \cap^{+} Q)$ に置き換え，クエリと候補単語の関連度を以下の式を用いて計算する．

$$\begin{aligned} &\text{NGD}^{+}(w,Q) \\ &= \frac{\max\{\log H(w), \log H(Q)\} - \log H(w \cap^{+} Q)}{\log N - \min\{\log H(w), \log H(Q)\}} \end{aligned} \tag{11}$$

なお，この指標は類似度ではなく 0 から ∞ の値をとる．もしクエリ Q と候補単語 w が同じであれば，$\text{NGD}(w,Q) = 0$ になる．

これらの特徴は時間情報や初期検索の結果に依存しないため，話題の時間変化や初期検索の検索精度に頑健である．

表 1 各クエリ拡張手法の省略表記
Table 1 Abbreviation of each query expansion method.

略記	詳細
WJ	WebJaccard$^+$（Web 類似度）
WO	WebOverlap$^+$（Web 類似度）
WD	WebDice$^+$（Web 類似度）
WP	WebPMI$^+$（Web 類似度）
NGD	NGD$^+$（Web 類似度）
IRQE	新近性を考慮したクエリ拡張手法．式 (1) 参照
TRQE	新近性と話題の時間変化を考慮した手法．式 (2) 参照
TSQE	話題の時間変化の類似度を基にした手法．式 (6) 参照

以上のように定義した Web 類似度を時間情報を使わない特徴として，クエリの関連語を予測するための順位学習器の特徴として利用する．

3.6 クエリ拡張のための順位学習

本研究では，8 つのクエリ拡張手法を用いる．各手法と省略表記の対応関係を**表 1** にまとめる．前節で述べたように，すべてのクエリ拡張手法には得手不得手がある．たとえば，Web 類似度（WJ，WO，WD，WP，NGD）は検索エンジンから取得した検索ヒット数だけを用いてクエリと単語の関連度を計算しており，初期検索の結果や話題の時間的な変化，新近性といった時間情報に左右されずに関連語を見つけることができる．しかし，時間情報を利用していないため，ある時間に話題と関連する単語を見つけることができない．一方，時間情報，特に新近性を取り入れた IRQE（Incorporating Recency into Query Expansion）と TRQE（Temporal Recency for Query Expansion）は 図 2 のトピック MB017 や MB026 のように適合文書がクエリ時間付近に存在するトピックに対して有効である．しかし，これらのモデルはいつ話題が活発に議論されたか，といった話題の時間変化を考慮できていない．逆に，TSQE（Temporal Similarity for Query Expansion）はクエリプロファイルを話題の時間変化に見立てることで，話題の時間変化と同じ時間帯に使用される意味的な関連する語を見つけることができる．だが，TSQE は新近性を考慮していないため，新鮮な文書を優先的に検索できるわけではない．マイクロブログ検索で有効なクエリ拡張手法は自明ではなく，単体でクエリ拡張を行うよりも，複数の手法を組み合わせることで検索精度を向上させることができる可能性がある．しかし，どのクエリ拡張手法を重視するかは検索対象に依存する．そこで，人手で適合度の評価を行ったクエリと文書の組から新近性，話題の時間変化，Web 類似度に基づくクエリ拡張が推定した元のクエリと候補単語の意味的な関連度を入力（特徴）とし，候補語を元のクエリに加えて再検索した検索結果と元のクエリの検索結果の検索精度の差を出力とした訓練データを作成する．そして，この訓練データを用いて順位学習を行うことで，どのクエリ拡張手法がマイクロブログ検索の検索精度向上に有効な手法であるかを自動的に重み付けする．順位学習器としては Gradient Boosted Decision Trees（GBDT）[7] を用い，初期クエリに関連する候補単語を順位付けする．その後，学習器が予測した関連度をもとに上位 K 個の関連語を初期クエリに加えて新たなクエリとし，tweet の再検索を行う．

4. 実験による評価

提案手法の有効性を評価するため，TREC 2011 マイクロブログトラックのテストコレクション（Tweets2011 コーパス）の全 tweet を用いて評価実験を行った．このコレクションは 2011 年 1 月 23 日から同年 2 月 8 日までに収集された約 1,600 万件の tweet から構成され，50 個のトピック（図 1 の例と同形式のトピック）を持つ [21]．さらに，任意の情報検索システムの評価を可能にするため，それぞれのトピックについて，適合・不適合の tweet が明示されている．適合性の評価は，不適合（ラベル 0），適合（1），非常に適合（2）という 3 つのカテゴリに基づいて行われている．各 tweet は，所与のトピックに内容が関連かつ新鮮なものであれば，適合すると判定されている．

4.1 Tweet データ

Tweet データの索引の作成方法について説明する．取得した tweet は Indri [22] を用いて，各トピックのクエリ時間以前の tweet を索引付けした．索引付けの際は禁止語（stopword）を除去せず，大・小文字の区別は行わず，接辞の除去（stemming）として Krovetz stemmer を適用した．索引をトピックごとに作る方法は，実際のマイクロブログ検索と同様の条件であり，クエリが発行された時点での未来の情報を使用しないためである．実験では，トピック 1〜50（MB001〜MB050）の $\langle title \rangle$ をテストクエリとして用いた．また，テストクエリの配布前にマイクロブログトラックの参加者に配られた 12 個のトピックの $\langle title \rangle$ をサンプルクエリとして用いた．

4.2 訓練データの作成

順位学習のモデルを作成するため，12 個のサンプルクエリで検索した文書に人手で適合度の評価を行った．訓練データの作成のために用いた索引には，サンプルクエリのクエリ時間以前に作成された tweet を用いた．そして，Indri 検索エンジンを用いてデフォルトの設定で索引付けを行い，各クエリに対して上位 300 件の tweet を検索し，適合度の評価を行った．適合度は，0，1，2 の 3 段階で，それぞれ，適合しない，適合する，非常に適合するに対応している．適合度の定義としては，クエリ中の単語や同義語が tweet に含まれているか，tweet の内容が適切にクエリの内容を表しているかどうかをアノテータが主観的に判断した．さらに，この tweet とクエリの適合度を示したリ

ストを候補単語とクエリの関連度を学習・予測するための順位学習の訓練データへと変換した．まず，各トピックごとに適合度を評価した tweet に対して適合度の評価のゆれを緩和することを目的として，多段階の関連度を考慮できる Normalized Discount Cumulative Gain（NDCG）[8] を用いて評価した．次に，初期検索の上位 M 件の tweet から K 個の候補単語を選び，初期クエリに加えて拡張クエリとし，tweet を再検索した．そして，再検索して得られた tweet のリストを NDCG で再び評価した．拡張したクエリが良ければ NDCG の値を初期クエリより向上させることができると考えられるので，クエリ拡張を使う前と後の NDCG の値の差をクエリ拡張用の順位学習の訓練データの出力とした．訓練データの出力の値は 0 から 1 の値に正規化した．ここで，クエリ語は所望の話題との適合度が高いと考えられるので，クエリ中の単語に対応する出力は 1 に設定した．

4.3 順位学習に用いた特徴

順位学習の訓練データの特徴として，表 1 に示す各クエリ拡張手法によって予測したクエリと候補単語の適合度を用いた．そのため，特徴の数は使用するクエリ拡張手法の数と同じである．候補単語としては初期クエリで検索した上位 100 件（$M = 100$）の tweet に含まれる語を用いた．式 (1) に示した Massoudi の手法（IRQE）に関しては，過去の研究に従い $\beta = 1.5 \times 10^{-5}$ に設定した．この手法では，ϕ 以上の tweet に表れる候補単語だけをクエリ拡張に利用していた．しかし，事前の実験結果において，あまり検索精度に差がないことが分かったため，すべての候補単語を用いることにした．TRQE と TSQE に関しては，検索して得られた上位 50 件（$L = 50$）の tweet を使って時間プロファイルを作成した．ここで，3.5 節の Web 類似度の場合と同様に，初期クエリ中の単語と候補単語の双方を含む tweet が 5 個以上なければ，その候補単語を除去した．すべてのクエリ拡張手法は予測したクエリと候補単語の適合度スコアが高い上位 15 個の（$K = 15$）単語を初期クエリに加えて拡張クエリとした．ただし，拡張クエリは初期クエリの単語を含まない．拡張クエリは Indri クエリ言語 [22] を用いて 6 : 4 で初期クエリと候補単語を重み付けした．

クエリ拡張手法を結合するために用いる GBDT には，学習率 α と木の深さ d とイテレーション回数 m の 3 つのパラメータが存在する．それぞれのパラメータを訓練データを用いて 10 分割交差検定で最適値を求めたところ，それぞれ $\alpha = 0.005$，$m = 3000$，$d = 4$ となった．テストクエリでの評価の際には，このパラメータを用いた．また，順位学習に使用する特徴を特徴選択することで検索精度を向上させることができることが知られているが [4]，本稿で扱う特徴は比較的少量であるため，本手法では特徴選択を行わないことにする．

表 2 各クエリ拡張手法の検索精度
Table 2 Retrieval performance of each QE method.

手法	P@10	P@30	Rprec	MAP
baseline	0.4490	0.3864	0.2584	0.2159
WJ	0.4816‡	0.4075†	0.2767†	0.2289
WO	0.4816	0.4156‡	0.2742‡	0.2298‡
WD	0.4816‡	0.4075†	0.2767†	0.2289
WP	0.4857†	0.4211‡	0.2776‡	0.2338‡
NGD	0.4755‡	0.4088‡	0.2745‡	0.2288‡
IRQE	0.4469	0.4184‡	0.2733†	0.2251
TRQE	0.4592	0.4299‡	**0.2828**‡	0.2288†
TSQE	0.4898‡	0.4231‡	0.2788‡	0.2325‡
GBDT	**0.4959**‡	**0.434**‡	0.2820‡	**0.2352**‡

4.4 評価指標

本提案手法の目的は，クエリ拡張によって生成したクエリを用いて検索を行い，適合度順に順位付けが行われた tweet のリストを取得することである．なお，評価の際は，TREC 2011 マイクロブログトラックの規定[*6]により，検索結果をタイムスタンプが新しい順に並べ替えた．評価には，上位 10 件と 30 件の Precision（P@10, P@30），R-precision（Rprec），Mean Average Precision（MAP）を用いた．P@30 は同トラックの公式の評価指標である．著者らは，検索して得られた上位 30 件の tweet だけを評価した．提案手法の統計的な優位性を検証するため，ウィルコクソンの符号順位検定を各クエリ拡張手法の組ごとに行った．一番良い成績を収めた手法は太文字で表し，統計的に $p < 0.05$ で向上すれば「†」を，$p < 0.01$ で向上すれば「‡」を付した．また，Krovetz stemmer を適用した索引と初期クエリを用い，Indri で検索した結果をベースライン（baseline）とした．断りのない限り，有意性については，ベースラインとの比較による．

4.5 実験結果

本節では，先述した 50 個のテストトピック（クエリ）とそのクエリに適合する 2,965 件の tweet（適合する：2,404 件，非常に適合する：561 件）を用いて各検索モデルを評価した．**表 2** に提案手法の結果を示す．最初に Web 類似度とベースラインとを比較する．すべての Web 類似度を用いたクエリ拡張手法（WJ，WO，WD，WP，NGD）はベースラインを上回っており，特に P@10 を著しく改善できた．WP はその定義の簡潔さにもかかわらず，Web 類似度の中で最も P@10，P@30，Rprec，MAP を向上させることができた．また，時間情報を考慮した手法である IRQE，TRQE，TSQE もベースラインからそれぞれ P@30 の値を

[*6] https://sites.google.com/site/microblogtrack/2011-guidelines

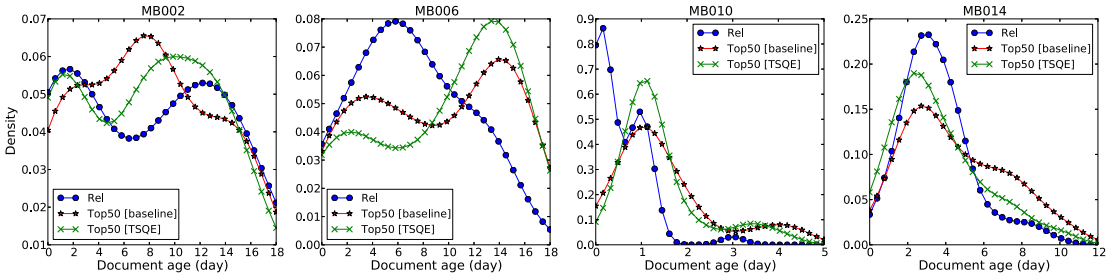


図 6 4 つのトピック，MB002，6，10，14 に対応する tweet の文書年齢のカーネル密度推定値
Fig. 6 The kernel density estimates corresponding to four topics: MB002, 6, 10, and 14.

表 3 クエリ拡張手法が他手法より P@30 を向上できたトピック数
Table 3 Difference of the number of improved topics comparing two QE methods.

	WJ	*WO*	*WD*	*WP*	*NGD*	*IRQE*	*TRQE*	*TSQE*
WJ	0	5	0	4	1	5	6	2
WO	8	0	8	1	2	7	6	6
WD	0	5	0	4	1	5	6	2
WP	8	2	8	0	2	7	7	7
NGD	7	5	7	4	0	9	8	5
IRQE	13	12	13	11	11	0	9	6
TRQE	13	10	13	10	9	8	0	7
TSQE	12	13	12	13	9	8	10	0

8%，11%，9% 向上させることができた．Web 類似度の結果と比べると，時間プロファイルを使用した手法 TRQE や TSQE の方が検索精度を向上させることができた．このことから，時間的に話題と関連のある単語を同定することで話題に関連する tweet を取得できることが分かった．さらに，順位学習を用いた手法である GBDT は，Rprec を除くすべて指標で他の手法よりも優れた検索能力を示し，Rprec に関しても，最も良い精度を示した手法との差はわずかであった．これらの結果より，時間に基づくクエリ拡張と時間に依存しないクエリ拡張を統合することは，マイクロブログの検索精度向上に有効であることが分かる．

続いて，各クエリ拡張手法に得意または不得意なトピックがあるかを調べるため，**表 3** に 2 つのクエリ拡張手法を比べて P@30 をより向上させることのできたトピックの数の差を示す．たとえば，TSQE は WJ より 12 個のトピックに対して P@30 を向上させることができており，その一方で WJ が TSQE よりも P@30 を向上させることができた検索トピックは 2 つある．WD と WJ を除くすべてのクエリ拡張手法について，他の手法と比べて P@30 を向上させることができるトピックが存在した．特に，IRQE，TRQE，TSQE は Web 類似度に対して多くのトピックで P@30 を向上させることができており，時間情報を用いた手法と用いない手法の違いが顕著に現れている．以上の結果から，各クエリ拡張手法が他の手法に対して P@30 を向上させることができるトピックは異なることが分かり，すべての手法を統合した GBDT が P@30 を向上させることができた理由が明らかとなった（表 2 参照）．

4.6 時間的分析

提案手法である TRQE と TSQE の有効性を時間的な観点から分析するため，クエリプロファイル，拡張クエリプロファイル，適合プロファイルの 3 種類の時間プロファイルを用意して比較する．**図 6** に，特徴的な 4 つのトピック，“2022 FIFA soccer（MB002）”，“NSA（MB006）”，“Egyptian protesters attack museum（MB010）”，“release of “The Rite”（MB014）” の時間プロファイルに関する文書年齢のカーネル密度の推定値を示す．図中の Rel，Top50 [baseline]，Top50 [TSQE] は適合 tweet，初期クエリで検索した上位 50 件の tweet，TSQE を用いて作成した拡張クエリで検索した上位 50 件の tweet の推定値である．トピック MB002，MB014 に関しては，TSQE は P@30 の検索精度をベースラインからそれぞれ 0.3000 から 0.4333，0.4000 から 0.8000 に向上させることができた．一方，トピック MB006 や MB010 に対して TSQE は P@30 の値を 0.0667 から 0.000 と 0.3667 から 0.2000 に低下させた．興味深いことに，図 6 より前者のトピックに対する拡張クエリプロファイル（Top50 [TSQE]）は適合プロファイル（Rel）に類似しており，逆に後者のトピックに関しては Top50 [TSQE] は Rel と類似しなくなっていることが分かった．この原因は，TSQE が初期検索の結果から作成した時間プロファイルに大きく依存するためである．よって，TSQE はクエリプロファイル（Top50 [baseline]）と適合プロファイルの間で小さい KL 情報量を持つトピック MB002，MB014 に関して，適合プロファイルに似た拡張プロファイルを予測することができた．また，それとは逆に，大きい KL 情報量を持つトピック MB006，MB010 に関しては，TSQE は正しい拡張プロファイルを予測できなかった．その結果，TSQE は前者のトピックに対して検索精度を向上させることができ，後者のトピックに対しては検索精度を低下させてしまったと考えられる．

一方，新近性を自身のモデルに取り入れた TRQE は，MB006 や MB010 に関して P@30 をそれぞれ 0.0667 から 0.2000 と 0.3667 から 0.6000 へと向上させることができた．

表 4 各クエリ拡張手法によって予測したクエリとの意味的関連度が高い上位 15 件の候補単語
Table 4 Top 15 candidate terms suggested by each QE method.

Number	手法	P@30	クエリ拡張で使用する候補の単語
MB002	NGD	0.4000	fifa soccer soccers blatter hammam uefa sepp daihyo jfa qatar matsuki epl socceroos goalkeeper evolution
	IRQE	0.4000	fifa soccer qatar cup blatter secclinton cheaptweet sellers world jogojusto sepp ebay marriage held verano
	TRQE	0.3667	rip verano neck ps governing body plans stadiums torres sunderland ban stage never followmejp sepp
	TSQE	0.4333	fifa qatar cup cups world check de el want club winter mundial sport sports best
	GBDT	0.4000	fifa soccer blatter soccers game games qatar cups cup world governing de football futbol updates
MB006	NGD	0.0667	nsa nsas ung analyst security ako former ng hires hired global apple relationships relationship secret
	IRQE	0.0333	nsa aprevious officialkevinp tlgang katulad kamay abot bhay vvip delivered niya ung palestinian nila opportunities
	TRQE	0.2000	de ng rt ung haha google watch ako nsa nsas na com relationships relationship ko
	TSQE	0.0000	nsa nsas com rt na ako ng security new news former sa apple analyst ko
	GBDT	0.2000	nsa nsas na security watch com sa ko ng new google relationships ako former relationship
MB010	NGD	0.2667	egyptian egyptians museums museum snipper protesters attacking protester pharaonic antiquities artefacts looters looted looting tutankhamun
	IRQE	0.5000	museum egyptian jan egypt looters protect protesters mummies destroy jazeera cairo looting army thousands national
	TRQE	0.6000	looters stealing pharaonic mummies broke destroyed artefacts cabinet storm egipto king museums museum soldiers tanks
	TSQE	0.2000	protester protesters force forces ndp important support ring prayers egypt youth joined shot shots security
	GBDT	0.3000	attacking protester protesters looting egyptians egyptian cairo looted protests sultanalqassemi police museums museum protestors jazeera
MB014	NGD	0.4333	riting rites released releases release rite hopkins exorcist exorcism sdk press beta xvid anthony ios
	IRQE	0.3000	rite sheso mjhainee somefin adeyossie hea hollywood aid poison feel fathilahnazri beyonces sdk moneymaker tooo
	TRQE	0.9333	credit ios topped apple developers office thriller box lookin chips dept nt zone per hopkins
	TSQE	0.8000	heard weekend single top scary films film win wins things little takes take taking anthony
	GBDT	0.5667	rite riting rites release releases released exorcism press hopkins newly news new im lays five

その理由は，TRQE はその定義から時間的にクエリ時間に近い時間に作成された文書を好む性質を持ち，MB006 や MB010 に関連する tweet の文書年齢の平均が初期クエリで検索した tweet の文書年齢の平均よりもクエリ時間に近いためと考えられる．その結果，TRQE は検索精度をこれらのトピックに関して向上させることができた．しかし，TRQE は多峰性の時間プロファイルを扱えないため，適合プロファイルの形が多峰であるトピック MB002 に関しては P@30 の値を 0.4000 から 0.3667 へと減少させた．

4.7 拡張したクエリの例

表 4 に 5 つのクエリ拡張手法（NGD，IRQE，TRQE，TSQE，GBDT）を用いて予測したトピック MB002，MB006，MB010，MB014 に対する 15 個の候補単語の例を掲載する．候補の単語は各クエリ拡張手法で計算された適合度の順に並べられている．Web 類似度の NGD を用いることで，トピック MB002 において「uefa」，「jfa」，「goalkeeper」，MB006 において「security」や「secret」といったクエリに関連する語が同定された．しかし，これらの語はある時間において話題に関連する語ではなく，このトピックに関しては適切でない．これは Web 類似度は時間情報を考慮していないことが原因として考えられる．TRQE は文書のタイムスタンプだけに依存するので，他のクエリ拡張手法が上位に予測できているクエリ語を上位に予測できていない．たとえば，TRQE はトピック MB002 に関連する「qatar」や「blatter」といった語を予測できていない[*7]．この理由は，TRQE（式 (6) を参照）が文書のタイムスタンプの分布は正規分布に従うと仮定しており，MB002 のような多峰性の適合プロファイルを予測できず，誤った単語を関連語として予測したためである．しかし，新近性を考慮したクエリ拡張手法である IRQE と TRQE は，適合文書がクエリ時間付近に存在する MB010 に関しては，「looters」や「mummies」といった話題に関連する

[*7] Sepp Blatter は現在の FIFA の会長であり，2022 年の FIFA ワールカップは Qatar で開催される．

語[*8]を予測できており，P@30 の値を向上させることができた．一方，新近性を考慮しない TSQE と NGD はこれらの語を予測できていない．さらに，初期検索の結果が悪い MB006 に関しても，TRQE は話題に関連する重要な単語「google」を予測できている[*9]．この原因はクエリプロファイルと適合プロファイルの文書年齢の平均が近いためである．IRQE は MB002，MB006，MB014 といった適合文書がクエリ時間付近にないトピックに対して有効に働かなかった．そのため，IRQE は TSQE と違い「anthony」，「hopkins」といった MB014 にとって重要な関連語を予測できなかった[*10]．GBDT は，他のクエリ拡張手法が上位に予測できている話題の関連語をすべての検索トピックに対して予測できた．

5. 関連研究

近年，マイクロブログ検索は情報検索コミュニティでさかんに研究されている比較的新しい研究分野である [19], [23]．現在まで，Duan ら [4] の手法にならい，Web 検索で使用される順位学習の手法をそのままマイクロブログ検索に適応する手法が，TREC 2011 マイクロブログトラックで良い成績を収めている [17], [18]．従来の順位学習を用いたマイクロブログ検索では主に URL の有無，フォロワの数，ユーザがリストに入れられた数など，時間情報を使わない特徴が用いられてきた．一方，Efron [6] はマイクロブログの文書を擬似クエリとして取得した時間情報を用いることにより，マイクロブログ検索の精度を向上させた．Li ら [13] は新近性を情報検索のための言語モデル [20] に取り入れる方法を提案しており，Efron ら [5] はこの手法を発展させ，時間的な特性を言語モデルに取り入れたマイクロブログ検索手法を提案しており，新近性を持つトピックに対して有効であることを示している．Dakka ら [3] は，時間的な特性を言語モデルに取り入れ，トピックに対して重要な時間を自動的に推定しながらニュース記事の検索を行う手法を提案している．著者らの手法は，検索精度の向上を図るため，文書（tweet）の作成日とクエリ時間との新近性，話題の時間変化，Web 類似度を用いた時間に依存しない特徴を順位学習の特徴として統合し，マイクロブログ検索に対するクエリ拡張を行った．

本提案手法はクエリ拡張に焦点を当てており，与えられたクエリに対して，自動的に関連語や同義語，単語のスペルミスの補正を行う手法を提案している．クエリ拡張の手法として，これまで Lavrenko ら [12] の関連モデルが標準的な方法としてよく用いられている．マイクロブログ検索のクエリ拡張としては Massoudi ら [16] がクエリ時間に近い単語を選び出す新近性のみを考慮したクエリ拡張手法を提案している．また，TREC 2011 では上位チームのほとんどが，本手法と同様に何らかのクエリ拡張を行っている．しかし，前処理や初期検索の結果が異なるため本手法との直接の比較はできない．また，本手法のように検索精度を著しく向上させることのできる時間情報を考慮したクエリ拡張手法はいまだ提案されていない．TREC 2011 の後，Liang ら [14] は，擬似関連フィードバックを用いてクエリ拡張を行い，クエリと文書によって重要な時間と組み合わせることで，時間を考慮したマイクロブログ検索を行っている．しかし，この手法はクエリ拡張と話題の時間情報の推定を分けて行っているため，ある時間に話題と関連する単語を予測することができない．著者らが知る限り，本稿で提案した手法は，新近性と話題の時間変化と時間情報以外の特徴を統合し，話題に時間的に関連した単語を同定できる初めてのクエリ拡張手法である．また，3.1 節で示したように，話題とその関連語の時間変化を時間プロファイルを用いて表現することで，トピックの関連語がいつ注目されたかといった時間的な解析を可能にした．

6. おわりに

本稿では，新近性に基づく手法（IRQE，TRQE），話題の時間変化に基づく手法（TSQE），Web 類似度に基づく手法（WJ，WO，WD，WP，NGD）の計 8 つのクエリ拡張手法を用いた．そして，個々のクエリ拡張手法の欠点を補うため，すべてのクエリ拡張手法を順位学習器を用いて組み合わせた．Tweets2011 コーパスを用いた実験により，時間の特性は話題に関連する単語を見つけるための重要な手がかりとなり，時間情報を用いたクエリ拡張手法は多くのトピックに対して検索精度を向上させることができることが分かった．また，クエリ拡張に時間プロファイルの概念を導入することで，所与のクエリに関する話題や話題の関連語がいつ注目されたかを時間的に解析できるようになった．さらに，各クエリ拡張手法は異なるトピックに対して P@30 を向上させることを示し，すべてのクエリ拡張手法を順位学習（GBDT）を用いて統合することで様々な話題の関連語を適切に予測することができ，クエリ拡張を行わない場合と比べ，P@30 で 12%の検索精度を向上させることができた．

[*8] 2011 年の 1 月末にエジプトで起きたデモの最中，略奪者たち（looters）によってカイロ市の有名なエジプト考古博物館のミイラ 2 体（mummies）の頭がもげ，いくつかの遺物が損壊したというニュースが報道された．

[*9] 米国の NSA（国家安全保障局）と Google との関係が 2 月初旬に話題となった．

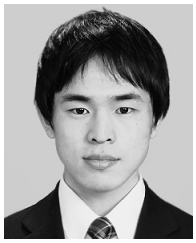
[*10] Anthony Hopkins 主演の “The Rite” という映画が 2011 年 1 月 28 日に公開された．

参考文献

[1] Bollegala, D., Matsuo, Y. and Ishizuka, M.: Measuring semantic similarity between words using web search engines, *WWW*, pp.757–766 (2007).

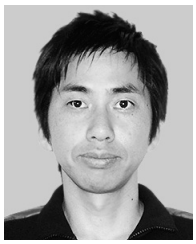
[2] Cilibrasi, R.L. and Vitanyi, P.M.B.: The google similarity distance, *IEEE Trans. Knowledge and Data Engineering*, Vol.19, No.3, pp.370–383 (2007).

[3] Dakka, W., Gravano, L. and Ipeirotis, P.: Answering general time sensitive queries, *CIKM*, pp.1437–1438 (2008).
[4] Duan, Y., Jiang, L., Qin, T., Zhou, M. and Shum, H.: An empirical study on learning to rank of tweets, *COLING*, pp.295–303 (2010).
[5] Efron, M. and Golovchinsky, G.: Estimation methods for ranking recent information, *SIGIR*, pp.495–504 (2011).
[6] Efron, M.: The University of Illinois' Graduate School of Library and Information Science at TREC 2011, *TREC* (2011).
[7] Friedman, J.: Greedy function approximation: A gradient boosting machine, *Annals of Statistics*, Vol.29, No.5, pp.1189–1232 (2001).
[8] Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques, *TOIS*, Vol.20, No.4, pp.422–446 (2002).
[9] Java, A., Song, X., Finin, T. and Tseng, B.: Why we twitter: Understanding microblogging usage and communities, *SNA-KDD*, pp.56–65 (2007).
[10] Jones, R. and Diaz, F.: Temporal profiles of queries, *TOIS*, Vol.25, No.3 (2007).
[11] Kwak, H., Lee, C., Park, H. and Moon, S.: What is Twitter, a social network or a news media?, *WWW*, pp.591–600 (2010).
[12] Lavrenko, V. and Croft, W.B.: Relevance based language models, *SIGIR*, pp.120–127 (2001).
[13] Li, X. and Croft, W.: Time-based language models, *CIKM*, pp.469–475 (2003).
[14] Liang, F., Qiang, R. and Yang, J.: Exploiting real-time information retrieval in the microblogosphere, *JCDL*, pp.267–276 (2012).
[15] Manning, C., Raghavan, P. and Schutze, H.: *Introduction to information retrieval*, Vol.1, Cambridge University Press, Cambridge (2008).
[16] Massoudi, K., Tsagkias, M., de Rijke, M. and Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts, *ECIR*, pp.362–367 (2011).
[17] Metzler, D. and Cai, C.: USC/ISI at TREC 2011: Microblog Track, *TREC* (2011).
[18] Miyanishi, T., Okamura, N., Liu, X., Seki, K. and Uehara, K.: TREC 2011 Microblog Track Experiments at Kobe University, *TREC* (2011).
[19] Ounis, I., Macdonald, C., Lin, J. and Soboroff, I.: Overview of the TREC-2011 Microblog Track, *TREC* (2011).
[20] Ponte, J. and Croft, W.: A language modeling approach to information retrieval, *SIGIR*, pp.275–281 (1998).
[21] Soboroff, I., McCullough, D., Lin, J., Macdonald, C., Ounis, I. and McCreadie, R.: Evaluating Real-Time Search over Tweets, *ICWSM*, pp.579–582 (2012).
[22] Strohman, T., Metzler, D., Turtle, H. and Croft, W.: Indri: A language model-based search engine for complex queries, *ICIA* (2005).
[23] Teevan, J., Ramage, D. and Morris, M.: #TwitterSearch: A comparison of microblog search and web search, *WSDM*, pp.35–44 (2011).



宮西 大樹 （学生会員）

平成 21 年神戸大学大学院工学研究科情報知能学専攻博士前期課程修了．同年同大学院システム情報学研究科計算科学専攻博士後期課程進学．情報検索，Web マイニングの研究に従事．人工知能学会学生会員．



関 和広 （正会員）

平成 14 年図書館情報大学修士課程修了．平成 18 年インディアナ大学博士課程修了．現在，神戸大学大学院システム情報学研究科講師．情報検索，自然言語処理，機械学習の研究に従事．Ph.D. 自然言語処理学会会員．

上原 邦昭 （正会員）

昭和 53 年大阪大学基礎工学部情報工学科卒業．昭和 58 年同大学院博士後期課程単位取得退学．同年産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授，同都市安全研究センター教授等を経て，現在，同大学院システム情報学研究科教授．工学博士．人工知能，特に機械学習，マルチメディア処理の研究に従事．人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員．