情報処理学会論文誌 Vol.55 No.5 1585–1594 (May 2014)

マイクロブログ文書の選択による適合フィードバックを用いた疑似適合フィードバックの検索性能改善

宮西 大樹[1,a)] 関 和広[1,b)] 上原 邦昭[1,c)]

受付日 2013年8月5日, 採録日 2014年2月14日

概要：マイクロブログ検索には疑似適合フィードバックを用いたクエリ拡張が有効であることが知られている．疑似適合フィードバックでは，初期検索の上位の検索結果は適合文書であり，この適合文書の中にユーザクエリの補強に役立つ単語が含まれていると仮定している．しかし，上位の検索結果の多くが非適合文書である場合，疑似適合フィードバックを用いると，ユーザクエリに関係のない単語が選ばれてしまう可能性がある．そこで，提案手法は上位の検索結果の中からマイクロブログ文書を 1 つだけユーザが選び，この文書をクエリ拡張に用いることで選んだ適合文書と類似した適合文書を上位の検索結果に集める．次に，再検索した上位の結果に対して疑似適合フィードバックを適用することで，検索精度の向上を図る．マイクロブログの代表的なサービスである Twitter のデータを用いて提案手法と従来の疑似適合フィードバックとを比較する．

キーワード：Twitter，マイクロブログ検索，疑似適合フィードバック，クエリ拡張

Improving Pseudo-relevance Feedback via Micro-document Selection

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

Received: August 5, 2013, Accepted: February 14, 2014

Abstract: Query expansion methods using pseudo-relevance feedback have been shown effective for microblog search because they can solve vocabulary mismatch problems often seen in searching short documents such as Twitter messages (tweets), which are limited to 140 characters. Pseudo-relevance feedback assumes that the top ranked documents in the initial search results are relevant and that they contain topic-related words appropriate for relevance feedback. However, those assumptions do not always hold in reality because the initial search results often contain many irrelevant documents. In such a case, only a few of the suggested expansion words may be useful with many others being useless or even harmful. To overcome the limitation of pseudo-relevance feedback for microblog search, we propose a novel query expansion method based on two-stage relevance feedback that models search interests by manual tweet selection and integration of lexical and temporal evidence into its relevance model. Our experiments using a corpus of microblog data (the Tweets2011 corpus) demonstrate that the proposed two-stage relevance feedback approaches considerably improve search result relevance over almost all topics.

Keywords: Twitter, microblog search, pseudo-relevance feedback, query expansion

1. はじめに

適合フィードバックに基づくクエリ拡張を用いることで，マイクブログ検索性能を著しく向上させることができることが知られている [2], [14], [19], [23]．クエリ拡張が有効な理由は，元のユーザクエリをクエリに意味的に関連する単語を用いて拡張することで，マイクロブログ文書の内容の短さに起因するクエリと文書間の語彙の不一致を克服できるからである [14], [19]．ほかにクエリを改善する方法として，ロッキオアルゴリズム [27] がある．ただし，ロッキオアルゴリズムに基づく古典的な適合フィードバックは

[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

各クエリごとに複数の文書に対して適合判定を行わなければならない．一方，疑似適合フィードバックに基づくクエリ拡張では，上位の検索結果を適合文書と見なすため適合文書の判定を行う必要はない [3], [9], [12], [13], [14], [20]．しかし，疑似適合フィードバックは上位の検索結果に適合文書が含まれていなければ，不適切な単語をクエリ拡張に用いてしまう可能性がある [21]．さらに，クエリ拡張に用いられる単語のうち，役に立つ単語はわずかであり，多くは有害であったり役に立たなかったりすることがある [1]．

疑似適合フィードバックに基づくクエリ拡張を用いてマイクロブログ検索の検索性能を向上させるためには，疑似適合フィードバックに用いる上位の検索結果を改善し，かつマイクロブログ検索に適した疑似適合フィードバックを行う必要がある．そこで，本稿では**マイクロブログ文書選択による適合フィードバック（MSF）**とマイクロブログのリアルタイム性を考慮した疑似適合フィードバックによる 2 段階の適合・疑似適合フィードバックを用いたマイクロブログ検索の枠組みを提案する．**MSF** では，まずユーザが 1 つの適合文書をユーザが選び，これを新たなクエリの一部として再検索することで，選択した適合文書に類似した適合文書を検索結果の上位に集め，疑似適合フィードバックに用いる検索結果の改善を行う．さらに，改善された検索結果に対してマイクロブログサービス特有のリアルタイム性を考慮した疑似適合フィードバックを適用することで，さらなる検索性能の向上を目指す．

本稿の貢献は以下のようになる．従来のマイクロブログ検索では，疑似適合フィードバックに代表される人手を介さない手法が主に研究されてきたため，人手による適合性判定が必要な適合フィードバックの有効性が不明であった．本稿は，マイクロブログ検索での疑似適合フィードバックの限界を明らかにし，適合フィードバックの有効性を立証した初めての論文である．また，代表的なマイクロブログサービス Twitter のマイクロブログ文書である tweet のデータ・セットを用いた実験から，適合フィードバックと疑似適合フィードバックを組み合わせることで，従来のマイクロブログ検索で用いられる疑似適合フィードバックと比較して，検索精度の低下するクエリの数を抑えつつマイクロブログ検索の性能を大幅に向上させることができることを示した．

2. 提案手法

提案する 2 段階の適合・疑似適合フィードバックを用いたクエリ拡張手法の概要を**図 1** に示す．まず，初期検索の上位の結果からユーザが自身の意図と一致する適合文書を 1 つだけ選び，その文書を所与のクエリに加えて再検索を行う．次に，再検索した上位の結果に対して疑似適合フィードバックによるクエリ拡張を適用し，再々検索を行う．本章では，これら提案手法の詳細について説明する．

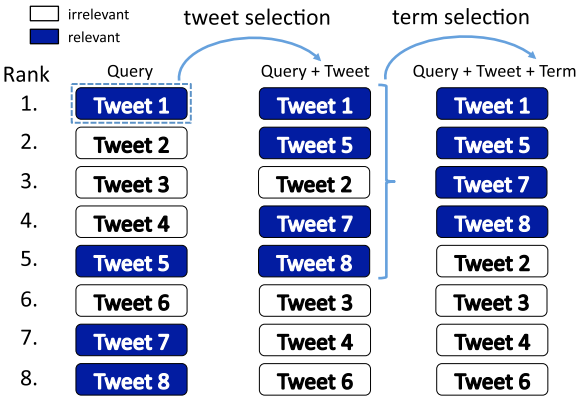


図 1 2 段階の適合フィードバックの概要図
Fig. 1 Overview of two-stage relevance feedback.

2.1 マイクロブログ文書選択による適合フィードバック

提案手法の第 1 の適合フィードバックでは，ユーザの意図に適合するマイクロブログ文書を選択し，選択した文書中にある単語を元のクエリに加えて再検索を行う．ここで，適合文書は他の適合文書を検索するために有効な単語を含むと仮定している．よって，ユーザの選択した適合文書にある単語をクエリ拡張に用いることで，選択した文書に意味的に類似した複数の適合文書を上位に順位付けできる．さらに，マイクロブログのリアルタイム性から意味的に類似した文書どうしは時間的に近いタイムスタンプを持つと考えられるため，タイムスタンプの分布を見ることで，検索対象の話題が盛り上がった時間帯をより正確に推定できるようになる．MSF の再検索よって推定した時間帯を用いることで，2.2.3 項で詳述する時間情報を考慮した疑似適合フィードバックの検索性能向上が期待できる．

本手法の前提として，ユーザが上位の検索結果から適合文書を見つけなければならない．そこで，実データを用いてこの仮定を検証する．予備実験として，上位 M' 件の検索結果に適合文書が少なくとも 1 つあるクエリの割合

$$\frac{1}{N'}\sum_{i=1}^{N'}\psi(P_i@M') \tag{1}$$

を実データで検証する．ここで，$\psi(\cdot)$ は入力変数が 0 より大きければ 1，それ以外は 0 となる関数である．N' はクエリ数，$P_i@M'$ は i 番目のクエリの上位 M' 件の Precision である．検索手法には，標準的な言語モデルとディリクレ平滑化を用いた検索を用いる．検索手法の詳細は，3.1 節の実験設定で述べる．**図 2** に TREC 2011 と 2012 のマイクロブログ検索課題で使用されたクエリを用いて，複数の M' に対する式 (1) の値を示す．

図 2 より，両データ・セットに対して，上位 30 件の中に少なくとも 1 つの適合文書を含むクエリの割合は 95%を超え，30 件を超えた場合，適合文書を含むクエリの割合は変化しないことが分かった．また，マイクロブログ文書の長さは比較的短いため（tweet であれば 140 文字以下），

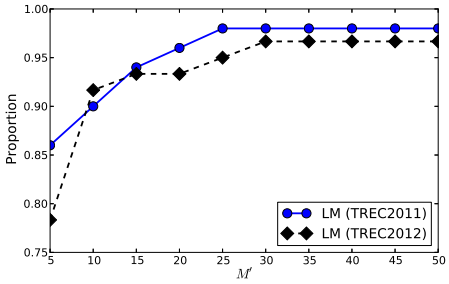


図 2 初期検索の上位 M' 件のうち，少なくとも 1 つの適合文書が含まれる検索クエリの割合

Fig. 2 Probability that at least one relevant document is contained among initial search results across different values of the cut off parameter M'.

上位 30 件のデータであれば容易に読み通すことができる．つまり，ユーザは容易に上位の検索結果からマイクロブログ文書を 1 つ選び，その文書中にある単語を適合フィードバックに基づくクエリ拡張に利用できることが分かった．

2.2 クエリ依存の時間を考慮した適合モデル

本稿で提案する疑似適合フィードバック手法は，言語モデルに基づく情報検索の枠組み（クエリ尤度モデル）を適合モデルに導入した疑似適合フィードバックの拡張である．そこで，まずクエリ尤度モデルについて説明し，次に言語モデルに基づく適合モデルについて紹介する．最後に，マイクロブログのリアルタイム性を考慮した疑似適合フィードバックを行うため，単語の頻度だけでなく文書のタイムスタンプも利用した言語モデルに基づく適合モデルについて詳述する．

2.2.1 クエリ尤度モデル

Ponte ら [26] によって提案されたクエリ尤度モデルは，クエリ Q が文書 D の単語分布から生成された過程をモデル化している．すべての文書はクエリに対する文書の事後分布 $P(D|Q)$ の値によって順位付けされる．$P(D|Q)$ はベイズ規則に基づき，下記に示すクエリ尤度 $P(Q|D)$ と文書の事前分布 $P(D)$ に分解することができる．

$$P(D|Q) \propto P(Q|D)P(D)$$

ここで，$P(D)$ は文書 D が任意のクエリに適合する確率を表している．n 個の単語 $q_1, q_2, \ldots, q_n$ からなるクエリ Q のクエリ尤度 $P(Q|D) = P(q_1, q_2, \ldots, q_n|D)$ には，文書 D に対してクエリ語どうしが独立だと仮定する次のユニグラム言語モデルを使用する．

$$P(Q|D) = \prod_{i=1}^{n} P(q_i|D) \tag{2}$$

ここで，n はクエリ中に含まれる語彙の数であり，q_i はクエリ Q 中の i 番目のクエリ語である．$P(q|D)$ には，最尤推定量 $P_{ml}(w|D) = \frac{f(w;D)}{\sum_{w' \in V} f(w';D)}$ を用い，$f(w;D)$ を文書 D 中の単語 w の出現頻度，$\sum_{w' \in V} f(w';D)$ は文書 D 中の全単語数，V をコーパス中の語彙集合とする．

式 (2) の定義から，クエリ中の単語が少なくとも 1 つ文書中に含まれなければ，$P(Q|D)$ の値は 0 になってしまう．この問題に対処するため，通常，言語モデルの平滑化が行われる．本稿ではクエリ尤度の平滑化のため，下記に示すディリクレ平滑化 [32] を用いる．

$$P(w|D) = \frac{|D|}{|D| + \mu} P_{ml}(w|D) + \frac{\mu}{|D| + \mu} P(w|C)$$

ここで，μ は平滑化パラメータを表す．

2.2.2 言語モデルに基づく適合モデル

Lavrenko ら [12] は言語モデルによる情報検索の枠組みを適合モデルに導入した．適合モデル $P(w|\mathcal{R})$ は，初期検索の結果において単語 w がクエリ中の単語と同時に観測される確率の結合分布で表すことができる．適合モデルは下記の式に従い上位の検索結果の語彙情報に基づいて単語 w を重み付けする．

$$P(w|\mathcal{R}) \approx P(w|Q) \propto \sum_{D \in \mathcal{R}} P(D)P(w|D) \prod_{i}^{n} P(q_i|D) \tag{3}$$

ここで，$\mathcal{R}$ はクエリ Q によって検索された結果の上位 M 件の文書集合を表す．このように，実際の適合文書の代わりに，検索結果の上位の結果を適合モデルに用いる方法を疑似適合フィードバックという．クエリ拡張を行う際は，$P(w|\mathcal{R})$ の値が高い上位 K 件の単語 w を元のクエリに加える．

ここで紹介した疑似適合フィードバックは，文書中の単語の出現頻度のような語彙情報を用いるだけで，文書のタイムスタンプのような時間情報を考慮していない．マイクロブログは，興味深い出来事が起こったときに大量のメッセージが作成されるリアルタイム性を持つため，ユーザが興味を持つ話題が盛り上がった時間帯に出現する文書や文書中の単語を用いることでマイクロブログ検索に適した疑似適合フィードバックが行えると考えられる．また，適切な時間情報を検索モデルに組み込むことで，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 [15]．そこで，本手法では語彙情報と時間情報を組み合わせた疑似適合フィードバック手法を提案する．

2.2.3 語彙と時間情報を考慮した適合モデル

まず，時間情報を適合モデル $P(w|Q)$ に取り込むため，Dakka ら [3] の定式化に従い，$P(w|Q)$ を各文書ごとに分解した $P(w, D|Q)$ の文書 D を文書中の単語を表す語彙情報 D_w と文書に付与されたタイムスタンプを表す時間情報 D_t に分け，以下の式を導出する．

$$P(w|Q) = \sum_{D \in \mathcal{R}} P(w, D|Q)$$

$$= \sum_{D\in\mathcal{R}} P(w, D_w, D_t|Q)$$
$$= \sum_{D\in\mathcal{R}} P(w, D_w|D_t, Q)P(D_t|Q) \quad (4)$$

次に，Efron ら [4] の手法で用いられた仮定を用いる．彼らは文書 D の語彙の集まりを表す語彙情報 D_w が文書 D のタイムスタンプを表す時間情報 D_t に依存しないと仮定し，式 (4) の条件付き確率から D_t を除去する．ここで，各語彙 w とタイムスタンプを表す D_t が独立だとは仮定していない．さらに，$P(Q)$ が $P(w|Q)$ に対して定数と見なすことができることを用いると，

$$P(w|Q) = \sum_{D\in\mathcal{R}} P(w, D_w|Q)P(D_t|Q)$$
$$\propto \sum_{D\in\mathcal{R}} P(D_w)P(w, Q|D_w)P(D_t|Q) \quad (5)$$

を得る．ここで，$P(D_t|Q)$ はクエリを所与としたときの文書 D のタイムスタンプ（本稿では各日）の生起確率とする．さらに，式 (5) について，文書を表す変数 D_w を与えたとき，クエリ拡張に使う候補となる単語 w（拡張語）とすべてのクエリ語 q_i どうしが独立であるという仮定を用いると，

$$P(w|Q) \propto \sum_{D\in\mathcal{R}} P(D_w)P(w|D_w)\prod_{i}^{n} P(q_i|D_w)P(D_t|Q) \quad (6)$$

を得る．ここで，式 (6) の $P(D_w)P(w|D_w)\prod_i^n P(q_i|D_w)$ は，式 (3) に示した通常の疑似適合フィードバックと同じ形式となる．つまり，式 (6) は，各文書ごとに文書 D 中で単語 w とクエリ語 q が同時に生起する確率を各クエリに対する文書 D のタイムスタンプの生起確率で重み付けしたものだと解釈できる．

確率 $P(D_t|Q)$ はクエリ Q と文書のタイムスタンプ D_t との関連度と見なせるため，クエリ Q と文書 D が類似した時間性質を持つとき，確率 $P(D_t|Q)$ の値が高くならなければならない．そこで，この時間的な性質を Q と D の時間モデルの距離で表現する．時間モデルには Jones ら [7] が提案した時間プロファイルの考えを用いて，クエリ Q に対する日単位の時間モデル $P''(t|Q)$ を

$$P''(t|Q) = \frac{1}{Z}\sum_{D\in\mathcal{R}} P(t|D)P(Q|D)$$

と定義する．ここで，$P(t|D)$ は文書 D のタイムスタンプと時間 t が一致したときに $P(t|D) = 1$ となり，それ以外は $P(t|D) = 0$ になる関数である．$P(Q|D)$ は文書 D に対するクエリ Q のクエリ尤度で，Z は正規化項である．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，以下のようにコーパス上での単語の統計値を用いて平滑化を行う．

$$P'(t|Q) = \lambda P''(t|Q) + (1-\lambda)P(t|C)$$

ここで，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．また，パラメータ λ は従来手法 [7] に従い 0.9 と定める．さらに，Dakka ら [3] が提案している各隣接する日どうしの $P'(t|Q)$ の値を用いた平滑化も同時に行う．平滑化の式は，

$$P(t|Q) \approx \frac{1}{2\cdot\phi+1}\sum_{i=-\phi}^{\phi} P'(t+i|Q) \quad (7)$$

である．ここで，ϕ は平滑化のための前後の日数を表す．また，Miyanishi ら [23] の調査で，マイクロブログ検索のデータ・セットにおいて，複数の検索トピックの適合文書のタイムスタンプの分布が 1 日ごとに変化していることを報告しているため，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と考えられる．そこで，$\phi = 1$ とする．式 (7) によって得られた $P(t|Q)$ を最終的なクエリの時間モデルとして用いる．

次に，文書の時間情報の生起確率を下式で定義する．

$$P(D_t|Q) \propto P(X > d) = e^{-\gamma d} \quad (8)$$

ここで，d は Q と D の時間モデル $P(t|Q)$ と $P(t|Q_D)$ の距離，γ は指数分布の尺度パラメータである．Q_D は Efron ら [5] の考えに基づき，文書 D を疑似クエリとしたものである．$P(t|Q_D)$ を用いることで，各文書の持つ時間情報を表現する．式 (8) は，2 つの時間モデル $P(t|Q)$ と $P(t|Q_D)$ の距離は指数分布に従うと仮定している．これは，上位の検索結果を疑似クエリ Q_D として検索した文書のタイムスタンプの多くは所与のクエリ Q によって検索された文書のタイムスタンプと類似しているという疑似適合フィードバックの考えに基づくためである．また，類似した時間モデルは類似した時間的特質を持つと考えられるため，バタチャリア係数を用いて，2 つの時間モデル $P(t|Q)$ と $P(t|Q_D)$ の距離 d を

$$d = -\ln \mathcal{B}(Q, D) \quad (9)$$

と定義する．ここで，バタチャリア係数 $\mathcal{B}(Q, D) = \sum_{t\in\mathcal{T}}\sqrt{P(t|Q)P(t|Q_D)}$ とする．このバタチャリア係数は 2 つの分布の類似度を表し，0〜1 の値をとる．さらに，式 (9) を式 (8) に代入することで，

$$P(D_t|Q) = \{\mathcal{B}(Q, D)\}^{\gamma} \quad (10)$$

を得る．この確率 $P(D_t|Q)$ は $P(t|Q)$ と $P(t|Q_D)$ が類似していれば値が大きくなり，類似していなければ値が小さくなる．

クエリ Q をユーザが興味のある話題だと考えると，時間モデルは話題に関連する文書のタイムスタンプの分布と見なすことができ，マイクロブログのリアルタイム性を表現

していると考えられる．さらに，$P(t|Q)$ はユーザが興味のある時間帯，$P(t|Q_D)$ は文書 D と関連のある時間帯を表していると考えられるので，$P(t|Q)$ と $P(t|Q_D)$ の類似度を表す $P(D_t|Q)$ を用いることで，疑似適合フィードバックにリアルタイム性を考慮することができる．

また，既存研究 [4] において，クエリに依存した新近性を検索モデルに組み込むことで，マイクロブログ検索の性能を向上させることができることが知られている．ここで，新近性とはクエリが発行された時間と文書のタイムンプの時間的近さを表す．そこで，指数分布のパラメータ γ を下式のようにクエリが持つ時間的な性質に応じて自動的に変化するように設計する．

$$\gamma = 1 - \sum_{t \in \mathcal{T}_Q} P(t|Q) \tag{11}$$

ここで，$\mathcal{T}_Q = \{t \in \mathcal{T} : t_Q - t < \alpha\}$，$\mathcal{T}$ は文書コレクション全体の時間範囲であり，t_Q はクエリ Q がユーザによって発行された時間（クエリ時間）である．α は各クエリごとの新近性の度合いを調整するハイパーパラメータである．つまり，このパラメータ γ は，クエリ時間の α 日前までの時間モデルの累積分布の相補的な値である．この式の定義より，与えられたクエリによって検索された上位の文書のタイムスタンプがクエリ時間付近に集中しているほど，γ の値が小さくなり，逆に離れていれば，γ の値は大きくなる．また，$P(D_t|Q)$ は γ の値が高いとき，つまり，与えられたクエリが多くの古い文書を上位に順位付けるとき，バタチャリア係数に対して線形的に増加する．一方，クエリが最近の話題を示し，γ の値が低くなるとき，$P(D_t|Q)$ は急激に増加する．最後に，式 (10) と (11) を式 (5) に代入して得た $P(w|Q)$ の値に基づきクエリ拡張を行う．

3. 評価

3.1 実験設定

3.1.1 評価データ

提案手法の有効性を評価するため，TREC 2011 と 2012 のマイクロブログトラックで使用された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．このテストコレクションは 2011 年 1 月 23 日から同年 2 月 8 日までに収集された 1,600 万件の tweet から構成されており，110 個のクエリを持つ．さらに，任意の情報検索システムの評価を行うため，各クエリについて，適合する tweet が明示されている．各 tweet は，所与のクエリに内容が一致していれば適合と判定される．また，適合度は非適合（ラベル 0），やや適合（1），非常に適合（2）の 3 つのカテゴリで構成される．本実験では，やや適合と非常に適合と判定された tweet を適合文書とする．

3.1.2 索引の作成

全 tweet データは，各クエリのクエリ時間以前の tweet について索引を作成する．検索エンジンには，Indri [29] を用いる．索引付けの際には，大・小文字は区別せず，禁止語（stopword）の除去は行わず，Krovetz stemmer による接辞の除去（stemming）を行う．索引を各クエリごとに作成する方法は，クエリが発行された時間から見て未来の情報を使わず，実際のマイクロブログ検索の状況に合わせるためである．本実験では，TREC のクエリの 1–50 と 51–110 番[*1]のタイトルをテストクエリとして用いる．これらのクエリセットはそれぞれ TREC 2011 と 2012 のマイクロブログトラックに使用されたデータである．

3.1.3 Tweet の検索

Tweet を検索するため，Indri に実装されているディリクレ平滑化を適用した言語モデル [32] を用いる．平滑化パラメータは過去の研究 [5] にならい，$\mu = 2500$ とする．この検索方法を LM とする．また，すべての適合・疑似適合フィードバックは LM の検索結果を用いる．非英語で書かれた tweet は，無限グラムを用いた言語判定器 ldig[*2] を用いて検索結果から除去する．すべての retweet（RT から始まる tweet）[*3] は TREC のマイクロブログトラックで非適合と判定されるため，最終的な検索結果から除去する．しかし，retweet は検索性能の向上に貢献する単語を含むことがあるため [2]，疑似適合性フィードバックを適用する以前の検索結果からは除去しない．また，マイクロブログトラックの指針に従い，301，302，403，404 の HTTP ステータスコードを持つすべての tweet を検索結果から除去する．最後に残った検索結果の上位 1,000 件について評価を行う．

3.2 検索方法

3.2.1 提案手法

提案手法の第 1 段階として，マイクロブログの選択によるフィードバック（MSF）を行う．本実験では，ユーザ評価が目的ではないため，テストコレクションに付与された適合度をもとに上位 L 件の初期検索の結果から自動的に適合文書を 1 つ選択し，MSF に使用する．ここで，2.1 節の予備実験の結果から L が 30 以上であれば，TREC 2011 および 2012 のデータ・セットにおいて 95%以上のクエリに対して上位の検索結果に適合文書が 1 つ以上存在すると分かっており，$L = 30$ 以降からはそのクエリの割合が変化しないため，L を 30 に設定する．選択された適合文書にはやや適合するか非常に適合する tweet を使用する．上位 L 件の検索結果の中に適合文書が複数存在する場合，tweet 中の語彙の種類数が最も多い文書を用いる．これは語彙が多い tweet は内容が豊富であり，ユーザはそのよう

*1 クエリ MB050 と MB076 は適合文書を含まないため，本実験では使用せず，他のクエリだけで評価を行う．
*2 https://github.com/shuyo/ldig
*3 情報の拡散を目的として，他のユーザの tweet を再投稿した tweet．

な長い tweet を選ぶと仮定したからである．初期検索の上位の結果中に適合文書がなければ，所与のクエリだけを用いる．また，選択された tweet から Indri の禁止語リストに対応する語彙，統一資源位置指定子（URL），ユーザ名（例：@trecmicroblog）を表す文字列を除去する．MSF のため，選択された tweet の単語と所与のクエリを組み合わせて新たなクエリとする．提案手法の第 2 段階では，マイクロブログ文書の選択後に語彙・時間情報を利用したクエリ拡張手法（QDRM）を実行する．QDRM では，初期検索で取得した文書を疑似クエリとして再検索を行い，上位 N 件の tweet を時間モデル $P(t|Q_D)$ の作成に用いる．また，この疑似クエリに対しても所与のクエリと同様の前処理を行う．ここで，MSF と QDRM の組合せを MSF + QDRM とする．

3.2.2 比較手法

提案手法の MSF と MSF + QDRM の検索性能を評価するため，複数の比較手法を用意する．最初の比較手法は，語彙情報だけを利用した標準的な疑似適合性フィードバック手法（RM）[12] である．RM の前に MSF を行う手法を MSF + RM とする．RM は時間情報を利用しない点で，QDRM と異なる．ただし，QDRM 中の γ を 0 に設定したとき，式 (11) の定義より，QDRM と RM は等しい．2 つめの比較手法は，文書の事前情報として文書の新近性を考慮した疑似適合性フィードバック [13] である．これを EXRM とする．EXRM は，QDRM と違い，クエリに依存した新近性やクエリが表す時間変化を考慮できない．また，新近性を考慮した手法に対するマイクロブログ選択によるフィードバックの効果を調べるため，他の手法と同様に MSF と組み合わせた MSF + EXRM を用意する．最後に，3 つめの比較手法としてクエリが示す時間変化を考慮した疑似適合性フィードバック手法（TBRM）[9] を用意する．これは語彙情報だけでなく，クエリに依存した時間情報を適合モデルに組み込んでいる．しかし，この手法は新近性や各文書が示すクエリの時間情報を考慮できない．このモデルと MSF による拡張である MSF + TBRM を提案手法である QDRM および MSF + QDRM と比較することで，各文書ごとの語彙・時間情報のモデル化の有効性を検証する．

すべての適合・疑似適合フィードバックを用いたクエリ拡張手法に対して，所与のクエリまたは所与のクエリと選択されたマイクロブログ文書を組み合わせた新たなクエリ（以下，MSF クエリと呼ぶ）で検索した上位 M 件の結果から URL，‘@’ で始まるユーザ名や特殊文字（!, @, #, ’, ” など）を除き，フィードバックに用いる候補語を抽出する．候補語には禁止語を利用せず，すべてのクエリ語，フィードバックに使用する単語，tweet を小文字化する．次に，$P(w|Q)$ の値が大きい候補語の中から K 個の単語を選択する．これをクエリ拡張に使用する単語（拡張語）とする．MSF を用いた手法では，拡張語は所与のクエリの単語を含むことはないものの，選択されたマイクロブログ文書中の語を含む場合がある．各拡張語は疑似適合フィードバックによる $P(w|Q)$ の値で重み付けを行い，クエリ拡張に用いる．最後に，疑似適合性フィードバックの拡張語と所与のクエリまたは MSF クエリを 1 : 1 の重みで組み合わせる．

QDRM と EXRM において，時間プロファイルの長さ N，ハイパーパラメータ α と尺度パラメータ γ の値を調整する．また，すべての手法に対して疑似適合性フィードバックに使用するフィードバック文書の数 M や拡張語の数 K を調整する．調整の方法は，すべてのパラメータに対して訓練データ中で，TREC 2011 マイクロブログトラックの公式の評価指標である上位 30 件の Precision が最大になるように最適化する．たとえば，TREC 2012 のデータ・セットを用いて上位 30 件の Precision が最大になるようにパラメータの調整を行い，TREC 2011 のデータ・セットで評価する．また TREC 2012 のデータ・セットで評価を行う際は，TREC 2011 のデータ・セットでパラメータの調整を行う．

3.3 評価指標

本手法の目的は，適合性フィードバックを用いて文書を順付けることである．検索モデルの評価を行うため，上位 10，30 の Precision（P@10，P@30）と Average Precision（AP）と nDCG [6] を用いる．nDCG は適合度のレベルを考慮できる．P@30 は TREC 2011 のマイクロブログトラックにおける公式の評価手法である．また，本実験では並べ替え検定 [28] を用いて実験結果の有意差検定を行う．

3.4 実験結果

3.4.1 疑似適合性フィードバックの効果

表 1 に P@10, P@30, AP と nDCG@10 の評価値を全 10 個の手法ごとに掲載する．MSF によって疑似適合フィードバックが有意に検索精度が向上した場合，有意確率 $p < 0.05$ と $p < 0.01$ ごとに △ と ▲ で示す．さらに，MSF を使わない手法間では，LM，RM，EXRM，TBRM，QDRM に対して $p < 0.05$ で統計的に有意である場合，各手法の頭文字 l，r，e，t，q を各結果の添字として用いる．さらに，MSF を用いる手法間では，MSF + LM，MSF + RM，MSF + EXRM，MSF + TBRM，MSF + QDRM に対する $p < 0.05$ の有意差を添字 l'，r'，e'，t'，q' で示す．また，表中の各欄で最も良い結果を太字で表す．

表 1 より，TREC 2011 と 2012 の両データ・セットにおいて，QDRM は RM，EXRM，TBRM と同様に初期検索（LM）の結果を大半の評価指標で検索精度を有意に向上させていることが分かった．さらに，QDRM は他の時間情報を考慮した疑似適合フィードバック EXRM と TBRM と同様に標準的な疑似適合フィードバック手法である RM を大半の指標で上回っていることが分かった．この結果から，

表 1 ベースラインと提案手法の検索精度
Table 1 The performance comparison of the proposed methods and baselines.

	TREC 2011				TREC 2012			
Method	AP	nDCG@10	P@10	P@30	AP	nDCG@10	P@10	P@30
LM	0.3571	0.5301	0.4755	0.4143	0.2408	0.4177	0.4814	0.3847
RM	0.4063_{l}	0.5616	0.5673_{l}	0.4741_{l}	0.3024_{l}	0.4592_{l}	0.5475_{l}	0.4503_{l}
EXRM	0.4204_{lr}	0.5725	0.5816_{l}	0.4762_{l}	0.3025_{l}	0.4663_{l}	0.5492_{l}	0.4520_{l}
TBRM	0.4020	0.5573	0.5673_{l}	0.4728_{l}	0.3139_{l}	0.4826_{l}	0.5610_{l}	0.4644_{lq}
QDRM	0.4206_{l}	0.5843	0.5735_{l}	0.4721_{l}	0.3039_{l}	0.4760_{l}	0.5542_{l}	0.4441_{l}
MSF + LM	$0.5040^{\blacktriangle}$	**0.6956**$^{\blacktriangle}$	$0.6388^{\blacktriangle}$	$0.4966^{\blacktriangle}$	$0.3198^{\blacktriangle}$	$0.5309^{\blacktriangle}$	$0.5763^{\blacktriangle}$	$0.4559^{\blacktriangle}$
MSF + RM	$0.5287^{\blacktriangle}$	$0.6730^{\blacktriangle}$	$0.6327^{\triangle}$	$0.5224_{l'}$	$0.3475^{\triangle}_{l'}$	$0.5352^{\triangle}$	$0.6068^{\triangle}$	0.4785
MSF + EXRM	$0.5328^{\blacktriangle}$	$0.6814^{\blacktriangle}$	$0.6449^{\triangle}$	$0.5218^{\triangle}_{l'}$	$0.3476^{\triangle}_{l't'}$	0.5329	$0.6068^{\triangle}$	0.4797
MSF + TBRM	$0.5174^{\blacktriangle}$	$0.6745^{\blacktriangle}$	$0.6429^{\blacktriangle}$	0.5177	$0.3415_{l'}$	0.5331	0.6051	0.4763
MSF + QDRM	**0.5384**$^{\blacktriangle}_{l'}$	$0.6843^{\blacktriangle}$	**0.6571**$^{\blacktriangle}_{r'}$	**0.5354**$^{\blacktriangle}_{l'}$	**0.3584**$^{\blacktriangle}_{l'r'e't'}$	**0.5552**$^{\blacktriangle}_{r'e'}$	**0.6220**$^{\blacktriangle}$	**0.4910**$^{\blacktriangle}_{l'}$

新近性や話題の時間変化などの時間情報を疑似適合フィードバックに組み込むことがマイクロブログ検索において効果的であることが分かった．ただし，これらの検索性能の違いは，RM と EXRM の AP を除き有意差はなかった．

マイクロブログの文書選択を用いた手法の場合，両データ・セットに対して MSF + LM は初期検索 LM をすべての評価指標について有意に上回った．この結果から，マイクロブログ文書を新たなクエリとして用いることは有効であると分かった．さらに，MSF と疑似適合フィードバックを組み合わせた手法は，MSF なしの手法に対してすべての評価指標について上回り，TREC 2011 のデータ・セットでの AP，nDCG@10，P@10 のすべてにおいて有意差があった．より詳細に見ると，MSF + QDRM は両データ・セットのすべての評価指標で QDRM を著しく上回っていることが分かった．また，両データ・セットにおいて，MSF + QDRM は他の MSF を使用する疑似適合性フィードバック手法（MSF + RM，MSF + EXRM，MSF + TBRM）を上回った．特に，TREC 2012 における AP の差は有意であった．これらの結果は，マイクロブログの選択による適合フィードバックの結果が疑似適合性フィードバックにとって有用であることを示している．特に，マイクロブログ文書の選択によるフィードバックは，クエリに依存した語彙や各文書ごとの時間的な情報を用いた疑似適合フィードバック QDRM に対して有効に機能することが分かった．

3.4.2 検索結果の頑健性

本項では，MSF による適合フィードバックと MSF と疑似適合フィードバックとの組合せによる検索の検索結果の頑健性について検証する．ここで，既存研究 [20] に従い，検索結果の頑健性は，疑似適合フィードバックにより初期検索から検索精度が向上または低下したクエリの数と定義する．頑健性が高い検索モデルは，多くのクエリに対して検索精度を向上させることができ，また精度が低下するクエリの数を抑えることができる．

図 3 は MSF を用いない手法（RM，EXRM，TBRM，QDRM），MSF を用いる手法（MSF + LM，MSF + RM，MSF + EXRM，MSF + TBRM，MSF + QDRM）について，初期検索の結果からの Mean Average Precision（MAP）の向上率または低下率ごとのクエリ数の度数分布を示している．使用するクエリは TREC 2011 と 2012 の両データ・セットを合わせて 108 個ある．図中の横軸が向上率・低下率の範囲，縦軸がクエリ数を示す．

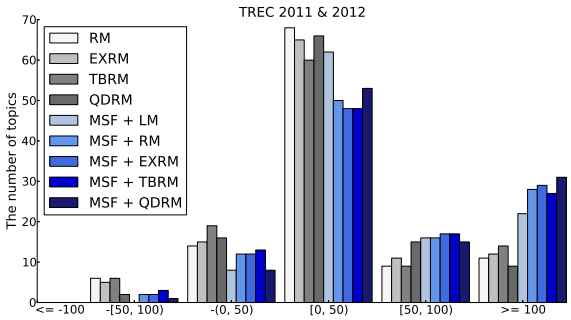


図 3 所与のクエリに対する全適合フィードバック手法の初期検索 LM への頑健さ
Fig. 3 Robustness of all PRF methods for the original queries w.r.t. the LM method.

図 3 より，MSF を用いる手法が MSF を用いない手法よりも頑健性が高いことが分かった．たとえば，MSF を用いない手法は，16〜23%のクエリに対して検索精度を低下させ，73〜79%のクエリに対して検索精度を向上させることができた．一方，MSF を用いた手法は 7〜15%のクエリに対して検索精度を低下させ，82〜88%のクエリに対して検索精度を向上させることができた．特に，提案手法の MSF + LM と MSF + QDRM は，大半のクエリに対して検索精度を低下させずに多くのクエリに対して検索精度を向上させることができた．たとえば，MSF + LM と MSF + QDRM で検索精度が低下したクエリの割合はそれぞれ全体の 7%と 8%である一方で，検索精度が向上したクエリの割合は双方ともに全体の 88%であった．このことから，MSF を用いることで頑健な検索を行うことができ，さらに

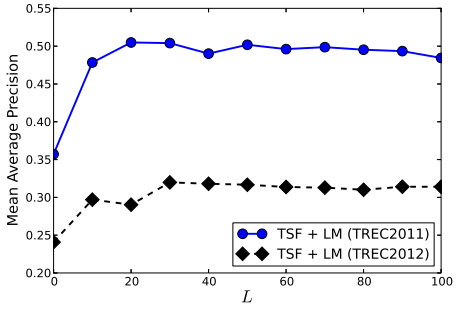


図 4 マイクロブログ文書の選択による適合フィードバックのパラメータ L に対する検索精度の敏感さ．x 軸が L の値，y 軸が TREC 2011 と 2012 のマイクロブログトラックで使用されたクエリに対する平均精度（MAP）

Fig. 4 Sensitivity to the number of top retrieved tweets L used for tweet selection feedback. The x-axis is the values of L, and the y-axis is the value of mean average precision over the TREC 2011 and 2012 Microblog track topics, respectively.

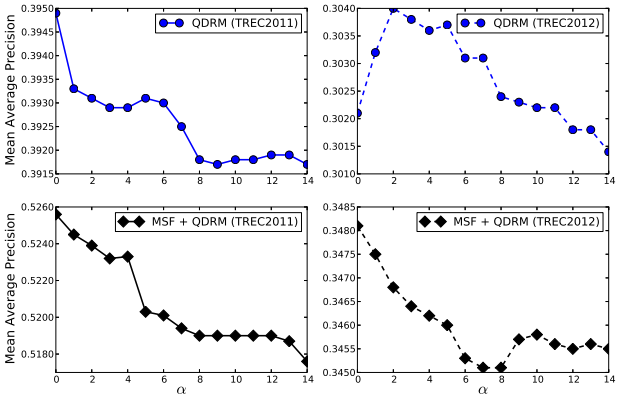


図 5 QDRM の新近性を調整するパラメータ α の敏感さ．x 軸が α の値，y 軸が TREC 2011 と TREC 2012 のクエリに対する MAP 値

Fig. 5 Sensitivity to the recency control parameter α used in QDRM over QDRM and MSF + QDRM at TREC 2011 (left-top and bottom) and QDRM and MSF + QDRM at TREC 2012 (right-top and bottom). The x-axis is the values of α, and the y-axis is the value of mean average precision.

語彙情報と時間情報を組み合わせた疑似適合フィードバックを用いることで，頑健に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，表 1 の結果では複数の評価指標において，MSF + LM よりも MSF + QDRM が優れていることから，MSF と提案する時間情報を用いた疑似適合フィードバックを組み合わせることで，MSF による頑健さを保ちつつ検索精度をさらに向上させることができることが分かった．

3.4.3 パラメータの敏感さ

本実験では，マイクロブログ文書による適合フィードバックとして，初期検索 LM の結果上位 30 件（つまり，$L = 30$）から最も語彙の種類数が多い tweet を 1 つ選んで，所与のクエリと組み合わせて新たなクエリとした．図 4 に，複数の L に対する MSF による MAP 値の違いを示す．両データ・セットに対する結果から，$L = 30$ まで MSF + LM の検索精度が増加し続け，L が 30 以上になると検索精度はほぼ変わらなくなることが分かった．この結果は，初期検索の上位 30 件内に MSF を使って検索精度の向上に役立つ適合 tweet を含みやすいことを示している．つまり，Tweets2011 コーパスの検索において，マイクロブログ検索を利用するユーザは初期検索の 30 件の tweet を読み，その中から 1 つだけ適合文書を選ぶだけで検索精度を効率的に向上させることができると分かった．

QDRM にあるクエリの新近性を考慮するパラメータ α に対する検索精度の敏感さについて見る．図 5 は，QDRM と MSF + QDRM の $L = 30$，$M = 100$，$N = 10$ と $K = 20$ を固定し，複数の α についての MAP 値を示している．QDRM と MSF + QDRM の両データ・セットに対する検索性能は，α の値を大きくした場合，急激に低下していることが分かる．これは式 (10) と (11) の定義から，α が大きくなると γ の値は 0 に近づきやすくなり，時間情報の影響が小さくなるからである．よって，この結果から QDRM 中のクエリと文書に依存した時間情報が効果的に機能していることが分かった．一方，TREC 2012 のデータ・セットにおける QDRM の α の適切な値は $\alpha = 2$ であった．これは新近性を考慮した効果があったことを示している．しかし，MAP 値の変化はわずかである．新近性の効果がわずかであった理由として，Tweets2011 コーパスのデータの期間は 2 週間と短いからだと考えられる．既存研究 [5] も同様の結果を指摘している．一方，両データ・セットに対する MSF + QDRM の α の最適値は 0 であった．これは新近性を無視し，クエリと文書が示す時間変化の情報だけを考慮すればよいことを示している．この理由としては，MSF を用いたことにより，QDRM はより正しいクエリの時間変化をとらえることができ，その結果，新近性の効果を無視できたと考えられる．

4. 関連研究

4.1 クラスタ情報に基づく検索

本稿で提案した手法は，文書の順位付けのためにクラスタの情報を用いるクラスタ検索 [8], [10], [11], [16], [30], [31] から着想を得ている．Kurland ら [10] は語彙情報の類似度を用いた k 近傍で文書のクラスタリングを行い，取得したクラスタ情報を用いて文書の順位付けを行った．Liu ら [16] は K 平均アルゴリズムを用いて類似する文書をクラスタリングし，クラスタの情報を用いて言語モデルに基づく検索モデルの平滑化を行った．また，Wei ら [31] は潜在的ディリクレ配分法（Latent Dirichlet Allocation，LDA）を用いてクラスタ情報を取得し言語モデルの平滑化を行った．一方，Kalmanovich ら [8] は，言語モデルの平滑化で

はなくクエリ拡張にクラスタの情報を利用した．さらに，Efron ら [5] は，k 近傍法で類似文書を集め言語モデルの平滑化を行う Tao ら [30] の考えに基づき，短い文書の検索に対する文書拡張手法を提案した．Efron らは，tweet のような短い文書は 1 つの話題について言及していると考え，文書を疑似クエリとして検索し，類似した文書を集めて時間・語彙情報を取得して検索に利用した．我々の手法は，手動で選んだ（本実験では自動で選択）1 つの適合文書を疑似クエリとして検索することで類似した文書を取得し，クエリに関係のあるクラスタを作成した点でこれらの既存手法とは異なる．つまり，我々の提案手法は，よりユーザの検索意図を反映したクラスタの情報を利用できる．このマイクロブログ文書を新たなクエリの一部として再検索する適合フィードバック手法は，TREC 2012 において Miyanishi ら [22] が提案した手法に基づいている．本提案手法はこの適合フィードバックの結果を疑似適合フィードバックに利用することで，ユーザの検索意図を反映しつつ，検索精度を向上させることができることを可能とした．

4.2 時間を考慮した情報検索

マイクロブログ検索は興味深い出来事が起こると数多くの tweet が多数の人々によって作成されるリアルタイム性を有している．このリアルタイム性を利用した時間情報に基づく検索手法が近年数多く提案されている．Dakka ら [3] は，あるクエリに対して重要な時間帯を自動的に特定し，その時間情報を言語モデルに基づく情報検索手法に組み込んだ．Peetz ら [25] は時間上の単語使用頻度のバーストをクエリ拡張に利用した．Keikha ら [9] は時間情報に基づく疑似適合モデルを提案し，これをブログ検索に適用した．しかし，Dakka，Peetz，Keikha らの手法はクエリごとの新近性と時間変化といった 2 つの時間情報を統合できなかった．Li ら [13] は新近性を言語モデルに基づく情報検索の枠組み [12], [26] に導入した．Peetz ら [24] は認知学に基づく時間情報に関する文書の事前分布を定義し，新鮮な文書を検索する手法を提案している．しかし，これらの手法はクエリごとの新近性を考慮できない．一方，クエリ依存の新近性を考慮した検索モデルの研究もなされている．Massoudi ら [17] はクエリが発行された時間付近に使用された単語に大きい重みを与えるクエリ拡張手法を提案している．Efron ら [4] は各クエリに依存した時間の新近性を言語モデルの枠組みに取り入れ，新鮮な情報を検索するために使用されるクエリに対して有効であることを示した．また，Miyanishi ら [23] は新近性とクエリの時間変化をクエリが表す時間の性質に応じて重み付けして結合するクエリ拡張手法を提案している．しかし，彼らのクエリ拡張手法は文書ごとの時間的性質を考慮していない．本稿で提案した手法は，語彙情報を各文書ごとの時間的性質に応じて重み付けを行い，さらにクエリごとの新近性を同時に考慮している．また，順位学習の枠組みを用いてマイクロブログ検索を行う手法も提案されており，Tweets2011 コーパスを用いた実験において優れた検索性能を示している [18]．順位学習は任意のクエリと文書間の適合度を素性として用いることができるので，本稿で提案した手法および比較手法の検索スコアを素性として統合することができ，検索性能のさらなる向上が期待できる．

5. おわりに

本稿では，マイクロブログ文書の選択と語彙情報と時間情報に基づく 2 段階の適合・疑似適合フィードバックを提案した．本手法は，まずユーザが適合するマイクロブログ文書を 1 つだけ選び，その文書をユーザクエリに加えて，新たなクエリとして再検索を行う．次に，検索精度をさらに向上させるため，クエリに依存した語彙・時間情報を利用した疑似適合フィードバックを適用した．TREC 2011 と 2012 のマイクロブログトラックのデータ・セットを用いた実験から，マイクロブログ文書による適合フィードバックを他の疑似適合フィードバックと組み合わせることで，大幅に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，マイクロブログ文書による適合フィードバックを用いれば，初期検索から検索精度が低下するクエリの数を抑えつつ，多くのクエリ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．今後は，ユーザの選んだマイクロブログ文書がユーザクエリと関係のない語彙を含む問題に対処するため，選択した文書の中から自動的にクエリと関連のある語彙を抽出し，これをクエリ拡張に用いることで，さらなる検索精度の向上を目指す．

参考文献

[1] Cao, G., Nie, J.-Y., Gao, J. and Robertson, S.: Selecting Good Expansion Terms for Pseudo-Relevance Feedback, *SIGIR*, pp.243–250 (2008).

[2] Choi, J. and Croft, W.B.: Temporal Models for Microblogs, *CIKM*, pp.2491–2494 (2012).

[3] Dakka, W., Gravano, L. and Ipeirotis, P.G.: Answering General Time-Sensitive Queries, *TKDE*, Vol.24, No.2, pp.220–235 (2012).

[4] Efron, M. and Golovchinsky, G.: Estimation Methods for Ranking Recent Information, *SIGIR*, pp.495–504 (2011).

[5] Efron, M., Organisciak, P. and Fenlon, K.: Improving Retrieval of Short Texts Through Document Expansion, *SIGIR*, pp.911–920 (2012).

[6] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *TOIS*, Vol.20, No.4, pp.422–446 (2002).

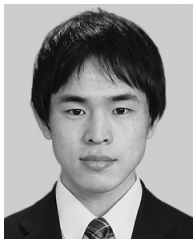
[7] Jones, R. and Diaz, F.: Temporal Profiles of Queries, *TOIS*, Vol.25, No.3 (2007).

[8] Kalmanovich, I.G. and Kurland, O.: Cluster-Based Query Expansion, *SIGIR*, pp.646–647 (2009).

[9] Keikha, M., Gerani, S. and Crestani, F.: Time-Based Relevance Models, *SIGIR*, pp.1087–1088 (2011).

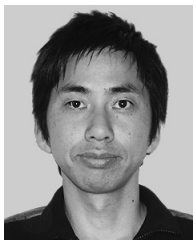
[10] Kurland, O. and Lee, L.: Corpus Structure, Language Models, and Ad Hoc Information Retrieval, *SIGIR*,

pp.194–201 (2004).
[11] Kurland, O., Lee, L. and Domshlak, C.: Better Than the Real Thing? Iterative Pseudo-Query Processing Using Cluster-Based Language Models, *SIGIR*, pp.19–26 (2005).
[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] Lin, J. and Efron, M.: Temporal Relevance Profiles for Tweet Search, *TAIA* (2013).
[16] Liu, X. and Croft, W.B.: Cluster-Based Retrieval Using Language Models, *SIGIR*, pp.186–193 (2004).
[17] 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).
[18] Metzler, D. and Cai, C.: USC/ISI at TREC 2011: Microblog Track, *TREC* (2011).
[19] Metzler, D., Cai, C. and Hovy, E.: Structured Event Retrieval Over Microblog Archives, *HLT/NAACL*, pp.646–655 (2012).
[20] Metzler, D. and Croft, W.B.: Latent Concept Expansion Using Markov Random Fields, *SIGIR*, pp.311–318 (2007).
[21] Mitra, M., Singhal, A. and Buckley, C.: Improving Automatic Query Expansion, *SIGIR*, pp.206–214 (1998).
[22] Miyanishi, T., Seki, K. and Uehara, K.: TREC 2012 Microblog Track Experiments at Kobe University, *TREC* (2012).
[23] Miyanishi, T., Seki, K. and Uehara, K.: Combining Recency and Topic-Dependent Temporal Variation for Microblog Search, *ECIR*, pp.331–343 (2013).
[24] Peetz, M.-H. and de Rijke, M.: Cognitive Temporal Document Priors, *ECIR*, pp.318–330 (2013).
[25] Peetz, M.H., Meij, E., de Rijke, M. and Weerkamp, W.: Adaptive Temporal Query Modeling, *ECIR*, pp.455–458 (2012).
[26] Ponte, J. and Croft, W.: A Language Modeling Approach to Information Retrieval, *SIGIR*, pp.275–281 (1998).
[27] Rocchio, J.J.: Relevance Feedback in Information Retrieval, *The SMART Retrieval System*, pp.313–323 (1971).
[28] Smucker, M.D., Allan, J. and Carterette, B.: A Comparison of Statistical Significance Tests for Information Retrieval Evaluation, *CIKM*, pp.623–632 (2007).
[29] Strohman, T., Metzler, D., Turtle, H. and Croft, W.: Indri: A Language Model-Based Search Engine for Complex Queries, *ICIA*, pp.2–6 (2005).
[30] Tao, T., Wang, X., Mei, Q. and Zhai, C.: Language Model Information Retrieval with Document Expansion, *HLT/NAACL*, pp.407–414 (2006).
[31] Wei, X. and Croft, W.B.: LDA-Based Document Models for Ad-Hoc Retrieval, *SIGIR*, pp.178–185 (2006).
[32] Zhai, C. and Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information Retrieval, *TOIS*, Vol.22, No.2, pp.179–214 (2004).



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准教授．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．