

推薦論文

コンセプト追跡を用いたマイクロブログ検索

宮西 大樹^{1,a)} 関 和広^{1,b)} 上原 邦昭^{1,c)}

受付日 2013年12月22日, 採録日 2014年4月7日

概要: マイクロブログ検索には、単語を用いた疑似適合フィードバックによるクエリ拡張が有効である。しかし、単語は意味的・時間的な曖昧性を持つため、単語を用いたクエリ拡張は有効に機能しない場合がある。そこで、本稿では、単語や2語以上の単語の組合せであるコンセプトを用いた疑似適合フィードバックによるクエリ拡張手法を提案する。さらに、提案手法は検索クエリと同時期に出現するコンセプトの頻度の時間遷移に関する情報を疑似適合フィードバックに組み入れることで、マイクロブログサービスのリアルタイム性を考慮する。代表的なマイクロブログデータである Tweets2011 コーパスを用いた実験から、提案するコンセプトを用いたクエリ拡張によって、検索クエリに適合し、かつ情報量の豊富な文書を効果的に検索できることを示す。

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

Concept Tracking for Microblog Search

TAIKI MIYANISHI^{1,a)} KAZUHIRO SEKI^{1,b)} KUNIAKI UEHARA^{1,c)}

Received: December 22, 2013, Accepted: April 7, 2014

Abstract: Incorporating the temporal property of words into query expansion methods based on relevance feedback has been shown to have a significant positive effect on microblog searching. In this paper, we propose a concept-based query expansion method based on a temporal relevance model that uses the temporal variation of concepts (e.g., terms or phrases) on microblogs. Our model naturally extends an extremely effective existing concept-based relevance model by tracking the concept frequency over time. Moreover, the proposed model produces important concepts that are frequently used within a particular time period associated with a given topic, which have more power to discriminate between relevant and non-relevant microblog documents than words. Our experiments using a corpus of microblog data (the Tweets2011 corpus) show that the proposed concept-based query expansion method improves search performance significantly, especially when retrieving highly relevant documents.

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

1. はじめに

マイクロブログは興味深いイベントが発生した際、そのイベントに関連する文書が多くのユーザによって作成されるリアルタイム性を持つ。そのため、マイクロブログ検索には時間情報を考慮した検索モデルが有効であ

る [10], [11], [12], [20], [21], [30]。特に、初期検索の結果から得た文書中の単語の頻度やその時間変化を手がかりとして、検索トピックに関連する語彙や時間帯を自動的に推定し、その語彙・時間情報からユーザクエリに意味的、時間的に関連する単語を推定してクエリ拡張に用いる疑似適合フィードバックが有効である [8], [23], [24], [27], [28]。これら時間情報を考慮した疑似適合フィードバックによるクエリ拡張手法は、主に疑似適合文書中の単語の頻度を語彙情報、単語

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, Hyogo 657-8501, Japan

a) miyanishi@ai.cs.kobe-u.ac.jp

b) seki@cs.kobe-u.ac.jp

c) uehara@kobe-u.ac.jp

本稿の内容は 2013 年 11 月の WebDB フォーラム 2013 にて発表され、同シンポジウムプログラム委員会により情報処理学会論文誌データベースへの掲載が推薦された論文である。

の頻度の時間変化を時間情報として利用する。しかし、単語を用いた疑似適合フィードバックは、検索クエリに適合し、かつ情報量の多い文書の検索に対して有効に機能しない場合がある。主な理由は、単語は意味的・時間的曖昧性を持ちやすく、単語を用いた検索モデルはユーザクエリに関連のない単語を誤って関連すると判定するためである [27]。

この単語を用いた検索の欠点を克服するため、複数の単語の組合せからなる概念を用いた検索手法が教師なし、または教師ありの枠組みで提案されている [1], [2], [3], [4], [5], [16], [18], [25], [26]。概念は単語よりも一般的に識別能力が高く、この概念の重要度の推定を行うことで、より多くの適合文書を上位に順位付けることができる。しかし、これまでの概念を用いた情報検索の枠組みは、時間情報があまり重要でない Web 検索で主に研究されてきたため、マイクロブログ検索にとって重要な時間情報を考慮しなかった。そのため、時間情報を効果的に利用して概念の重み付けを行う方法は未解決の問題である。

この問題を解決するため、我々は時間に基づく適合モデルを用いた新たな概念の重み付け手法を提案する。本稿では、最新の概念を用いたクエリ拡張手法である潜在概念拡張 (Latent Concept Expansion; LCE) [26] をマイクロブログが持つ時間情報の追加によって拡張した検索モデルを提案する。

従来の単語を用いた検索モデル (wRM [17]) と提案手法である語彙と時間情報を考慮した概念を用いた検索モデル (cTRM) の違いを見るため、TREC マイクロブログトラックで使用された検索トピック “White House spokesman replaced” (トピック番号: MB044) について、各手法を用いて推定した意味に関連する語と概念を表 1 に示す。この検索トピックは、アメリカ合衆国の副大統領 Joseph R. Diden Jr. の主任広報担当官だった Jay Carney 氏がホワイトハウスの報道官を引き継ぐニュースに関する文書の検索を意図している。概念の推定に使用した手法は、単語を用いた検索モデル wRM と提案手法の概念の語彙情報だけを用いた検索モデル cTRM (Lexical)、提案手法の概念の時間情報だけを用いた検索モデル cTRM (Temporal) である。本稿では、便宜上、単語や概念の頻度や文書頻度を語彙情報、頻度の時間変化を時間情報と呼ぶことにする。

表 1 より、単語を用いた検索モデル (wRM) は、「jay」や「carney」といった検索トピックに関連する単語を推定できていることが分かる。しかし、両単語とも多くの文書に出現する単語であるため、これらの単語を検索に用いれば、非適合文書が上位に検索されてしまう可能性が高い。一方、概念の語彙情報だけを用いる cTRM (Lexical) は、「news spokesman」、「press secretary」、「jay carney」といった適合文書の検索に役立つ概念を推定できて

表 1 単語に基づく疑似適合フィードバックと概念に基づく疑似適合フィードバックが推定した検索クエリ MB044: “White House spokesman replaced” に関連する概念の例

Table 1 Example of expanded words and concepts about a topic MB044: “White House spokesman replaced” suggested by a word-based PRF (wRM) and a concept-based temporal one (cTRM).

wRM	cTRM (Lexical)	cTRM (Temporal)
jay	jay	carney
carney	carney	jay
qantas	qantas	press secretary
new	new spokesman	jay carney
obama	new	biden spokesman

いる。特に、この例では、cTRM の時間情報だけを用いた方法 cTRM (Temporal) が cTRM (Lexical) によって上位に予測できなかった関連概念を推定できている。よって、概念の語彙情報と時間情報の両者を組み合わせた cTRM を用いれば、単語を用いた標準的な手法や、LCE のような概念の時間情報を用いない手法よりも有用な概念を推定できると考えられる。

提案する概念追跡を用いた疑似適合フィードバックの有効性を検証するため、TREC 2011 と 2012 のマイクロブログトラックで使用された 1,600 万件の tweet (代表的なマイクロブログサービス Twitter で用いられるマイクロブログ文書) から構成される Tweets2011 コーパス*1を用いる。実験結果から、我々の提案する疑似適合フィードバックによるクエリ拡張はマイクロブログ検索の性能を向上させることができ、特に、適合度の高いマイクロブログ文書の検索に有効であることを示す。

本稿は、次のように構成されている。2 章で既存研究の紹介と提案手法との違いについて説明する。3 章で言語モデルに基づく情報検索の枠組みについて説明し、その拡張として概念の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提案する。4 章で提案手法の有効性を検証するための実験設定と実験結果について詳述する。最後に、5 章で本研究のまとめと結論について述べる。

2. 関連研究

2.1 時間情報を考慮した情報検索

マイクロブログ検索を利用する人々は、ニュース速報やリアルタイム情報といった特定の時間に関連する情報を探そうとする [35]。そのため、話題の新鮮さや時間変化といった時間に関する特徴がマイクロブログ検索にとって重要となる。現在まで、特定の時間に関連する情報の検索を目的として、多くの手法が時間の特徴を検索の枠組みに取り入れてきた。Li ら [19] は文書の作成日時が最近であるほど文書の生起確率が高くなると仮定して、これを文書の事前確率として言語モデルに基づく情報検索の枠組みに導入

*1 <http://trec.nist.gov/data/tweets/>

した。Efron ら [10] は各クエリの持つ時間的性質に応じて新近性の影響を自動的に決定する言語モデルの平滑化手法を提案した。Dakka ら [9] は所与の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む手法を提案している。Keikha ら [14] 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。また、Lin ら [21] は、検索対象の話題に関連する時間帯を適切に推定することで、時間情報を考慮したマイクロブログ検索の検索性能を向上させることができると報告している。さらに、Miyanishi ら [27] は、Jones ら [13] によって提案された時間プロファイル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とで検索クエリに関する時間的な特徴を表現した。そして、単語の新近性と時間変化をそれぞれ考慮した 2 種類のクエリ拡張をクエリが表す時間の性質に応じて加重結合する手法を提案している。Efron ら [12] はクラスタ検索の考えを用いて、語彙情報と時間情報を組み合わせた文書拡張手法を提案し、彼らの手法が短い文書の検索に有効であることを示した。また、Miyanishi ら [28] は、初期クエリとその適合 tweet 1 つを組み合わせたクエリを作成し、この疑似クエリで検索した結果に対してクエリと文書に依存した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適用することで、頑健な Twitter 検索ができることを示した。しかし、これらの手法は主に単語の頻度やその時間遷移を用いており、単語よりも適合文書の識別能力に優れた複数の単語の組合せを考慮したコンセプトを利用していない。一方、提案手法はコンセプトの頻度とその時間変化を考慮することで、コンセプトの語彙と時間情報を組み合わせたクエリ拡張を実現する。

2.2 コンセプトを用いた情報検索

近年、コンセプトを用いた情報検索の手法が単語を用いた手法と比較して、複数の検索課題で優れた検索性能を示している。特に、コンセプトの重み付けとして、ユニグラムやバイグラムまたは任意の単語間の依存関係をモデル化するマルコフ確率場 (Markov Random Field; MRF) を用いた手法の有効性が報告されている。MRF は、特に上位の検索結果を重視する Web 検索において検索性能の良さが報告されている。たとえば、Metzler ら [25] は、MRF を用いて単語間の依存関係を表現して、クエリ中の接続する単語の組合せや近接する単語の組合せを考慮したクエリ改良手法を提案している。さらに、彼らは MRF に基づく単語間の依存関係の表現を用いて、文書から複数の単語の組合せであるコンセプトを取得し、このコンセプトを潜在コンセプトとして疑似適合フィードバックに基づくクエリ拡張に用いる潜在コンセプト拡張 [26] を提案している。彼らの手法は、単語を用いた検索モデルよりも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た。しかし、LCE はコンセプトの重み付けにコンセプトの文書中での頻度やコンセプトを含む文書の頻度を主に用いており、外部ソースのコンセプトの情報を利用していない。そこで、Bendersky ら [3] はコンセプトの重みを内部だけでなく Wikipedia やクエリログといった外部ソースのユニグラムやバイグラムの統計情報を利用して推定する手法を提案している。さらに、Bendersky ら [4], [5] は、長いクエリ中のコンセプトとクエリに関連する潜在コンセプトの重みを検索対象以外のデータから取得した語彙情報を特徴として教師あり学習により求め、重み付けしたコンセプトで文書の順位付けを行った。最近では、Bendersky ら [2] はハイパーグラフを用いてコンセプトどうしの組合せを表現し、クエリ改良を行う手法を提案している。これらのコンセプトを用いた検索手法は、いずれも TREC の標準的な Web 検索用のテストコレクションにおいて高い検索性能を示している。しかし、既存のコンセプトを用いた情報検索の枠組みは、文書、クエリログ、Wikipedia 中でのコンセプトの語彙情報を用いるだけで、マイクロブログ検索にとって重要な要素である時間情報を考慮していない。提案手法は、マイクロブログ文書の検索にコンセプトの語彙情報と時間情報を同時に利用できるため、コンセプトの語彙情報だけを用いる手法よりも検索性能の向上が期待できる。

3. 提案手法

提案するコンセプトを用いた疑似適合フィードバック手法は、Ponte ら [31] によって提案された言語モデルに基づく情報検索モデル (クエリ尤度モデル) の拡張である。そこで、まず単語を用いた言語モデルについて説明し、次に、提案手法の拡張元となる言語モデルの枠組みによる単語を用いた疑似適合フィードバックについて説明する。

3.1 言語モデルに基づく情報検索

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, \dots, q_n からなるクエリ Q のクエリ尤度 $P(Q|D) = P(q_1, q_2, \dots, q_n|D)$ には、次のユニグラム言語モデルを使用する。

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

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 はコーパス中の語彙集合とする.

式 (1) の定義から, クエリ中の単語が少なくとも 1 つ文書に含まれなければ, $P(Q|D)$ の値は 0 になってしまう. 通常, この問題に対処するため, 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 [36] 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, $P(w|C)$ は文書コレクション C 中の単語 w のユニグラム, μ は平滑化パラメータである.

3.2 単語の語彙情報に基づくクエリ拡張

本節では, 単語の語彙情報だけを用いた疑似適合フィードバックについて紹介する. Lavrenko ら [17] 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.

$$\begin{aligned} P(w|\mathcal{R}) &\approx P(w|Q) = \sum_{D \in \mathcal{R}} P(w, D|Q) \\ &= \frac{1}{P(Q)} \sum_{D \in \mathcal{R}} P(w, Q|D) P(D) \\ &\propto \sum_{D \in \mathcal{R}} P(D) P(w|D) \prod_i^n P(q_i|D), \end{aligned} \quad (2)$$

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 (疑似適合文書) を表す. 式の導出には, $P(Q)$ が $P(w|Q)$ に対して定数と見なすことができ, 文書 D を与えたときに単語 w とすべてのクエリ語 q_i どうしが独立になるという仮定を用いている. 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, 式 (2) の $P(w|Q)$ の値の降順に単語 w を順位付けし, 上位 k 個の単語を元のクエリに加えて再検索する. ただし, この疑似適合フィードバックは単語の語彙情報のみを用いている.

3.3 語彙情報に基づくコンセプト拡張

単語間の依存関係を考慮したコンセプトをモデル化するため, Metzler ら [26] はコンセプトの語彙情報を用いた疑似適合フィードバックとして潜在コンセプト拡張 (LCE) を提案している. LCE は元のクエリに関連する 1 つ以上の単語からなるコンセプトを用いる. LCE ではこのコンセプトを潜在コンセプトと名付け, 単語間の依存関係を考

慮したコンセプトをモデル化するためマルコフ確率場の考えを用いている [25]. LCE を使えば, ユーザの入力したクエリに明示されないコンセプトを用いて, 元のクエリを改良することができる. 本稿では, 比較手法と提案手法との要素ごとの違いを見るため, Bendersky ら [4] による LCE の定式化を用いる. 彼らの LCE は疑似文書集合 \mathcal{R} (初期検索の上位 M 件の文書) から取得したコンセプト c の重みを下式に従って求める.

$$\begin{aligned} S_{LCE}(c, Q) \\ \propto \sum_{D \in \mathcal{R}} \exp\{\gamma_1 \phi_1(Q, D) + \gamma_2 \phi_2(c, D) - \gamma_3 \phi_3(c, C)\} \end{aligned} \quad (3)$$

ここで, $\phi_1(Q, D)$ はクエリ Q 中のコンセプトと文書 D 中のコンセプト c との一致度を表す関数, $\phi_2(c, D)$ はコンセプト c の文書 D 中での生起確率, $\phi_3(c, C)$ はコーパス C 中でのコンセプト c の生起確率を表す関数, $\gamma_1, \gamma_2, \gamma_3$ は上記関数の重みを表すパラメータである.

さらに, 式 (3) に対して, $\gamma_1 = \gamma_2 = \gamma_3 = 1$, $\phi_1(Q, D) = \log P(Q|D)$, $\phi_2(c, D) = \log P(c|D)$, $\phi_3(c, C) = 0^{*2}$ とする. 文書 D を与えたときにクエリ Q 中のコンセプト $\hat{q}_1, \hat{q}_2, \dots, \hat{q}_m$ とクエリ拡張に用いるコンセプト c が独立だと仮定する *bag-of-concepts* の考えを用いた場合, 下式のコンセプト c とクエリ Q の関連度を表すスコア関数 $S_{cRM}(c, Q)$ を得る.

$$S_{cRM}(c, Q) \propto \sum_{D \in \mathcal{R}} P(D) P(c|D) \prod_j^m P(\hat{q}_j|D) \quad (4)$$

ここで, \hat{q}_j はクエリ Q 中の j 番目のコンセプトを表す. 式 (4) は, 式 (3) からコーパス中のコンセプトの出現頻度による罰則を省略したモデルとなっている. また, 式 (4) は, 式 (2) の Lavrenko ら [17] による単語を用いた疑似適合フィードバックと同形式である. しかし, 式 (2) と違い, 式 (4) は単語だけでなく複数の単語の組合せも扱うことができる. ただし, ここで紹介した疑似適合フィードバックは, 文書中の単語やコンセプトの生起確率のような語彙情報を用いるだけで, マイクロブログ検索にとって重要な時間情報を考慮していない.

3.4 語彙と時間情報に基づくコンセプト拡張

マイクロブログは, 興味深い出来事が起こったときにその出来事に対するメッセージが大量に作成されるリアルタイム性を持つ. そのため, 多くの研究で時間情報を考慮した疑似適合フィードバックがマイクロブログ検索に有効であることが報告されている [8], [23], [24], [27], [28]. そこで, 本稿では, コンセプトの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に基づくクエリ拡張を提案する.

*2 コーパス中のコンセプトの頻度は検索結果にあまり影響を与えないため [22], $\phi_3(c, C) = 0$ とする.

コンセプトと語彙情報を時間情報を適合モデルに組み込むため、コンセプトを用いた適合モデル $P(c|\mathcal{R})$ は時間情報と語彙情報の2つの情報源から構成されると仮定する。そこで、コンセプト c が語彙情報を表す疑似適合文書 \mathcal{R}_l と時間情報を表す疑似適合文書 \mathcal{R}_t から独立して生成されたと考え、以下の適合モデル $P(c|Q)$ を導出する。

$$P(c|Q) = \sum_{D_l \in \mathcal{R}_l} \sum_{D_t \in \mathcal{R}_t} P(c, D_l, D_t|Q) \\ = \sum_{D_l \in \mathcal{R}_l} \sum_{D_t \in \mathcal{R}_t} P(D_l|c, D_t, Q)P(c, D_t|Q) \quad (5)$$

ここで、 D_l を \mathcal{R}_l 中の文書、 D_t を \mathcal{R}_t 中の文書のタイムスタンプが表す各時間（本稿では各日）と見なす。次に、Efronら [10] の手法と同様に時間 D_t が文書 D_l と独立であるという仮定を用いて、式 (5) の条件付き確率から D_t を除去し、 $P(Q)$ が $P(c|Q)$ から見て定数であることを用いて、下式を得る。

$$P(c|Q) = \sum_{D_l \in \mathcal{R}_l} P(D_l|c, Q) \sum_{D_t \in \mathcal{R}_t} P(c, D_t|Q) \\ = \frac{1}{P(c|Q)} \sum_{D_l \in \mathcal{R}_l} P(c, D_l|Q) \sum_{D_t \in \mathcal{R}_t} P(c, D_t|Q) \\ \propto \frac{1}{P(c|Q)} \sum_{D_l \in \mathcal{R}_l} P(D_l)P(c, Q|D_l) \\ \times \sum_{D_t \in \mathcal{R}_t} P(D_t)P(c, Q|D_t)$$

次に、bag-of-concepts の考えを用いて、文書 D_l と時間 D_t をそれぞれ与えたときにクエリ Q 中のコンセプト $\hat{q}_1, \hat{q}_2, \dots, \hat{q}_m$ とクエリ拡張に用いるコンセプト c が独立であると仮定すると、

$$P(c|Q) \propto \frac{1}{P(c|Q)} \sum_{D_l \in \mathcal{R}_l} P(D_l)P(c|D_l) \prod_j^m P(\hat{q}_j|D_l) \cdot \\ \sum_{D_t \in \mathcal{R}_t} P(D_t)P(c|D_t) \prod_j^m P(\hat{q}_j|D_t)$$

が得られる。さらに、 $P(c|Q)$ が非負の関数であることを用いると、以下のコンセプト c とクエリ Q の関連度スコア $S_{cTRM}(c, Q)$ に従いコンセプト c を順位付けすることができる。

$$S_{cTRM}(c, Q) \stackrel{rank}{=} \left\{ \underbrace{\sum_{D_l \in \mathcal{R}_l} P(D_l)P(c|D_l) \prod_j^m P(\hat{q}_j|D_l)}_{\text{Lexical}} \cdot \underbrace{\sum_{D_t \in \mathcal{R}_t} P(D_t)P(c|D_t) \prod_j^m P(\hat{q}_j|D_t)}_{\text{Temporal}} \right\}^{1/2}, \quad (6)$$

ここで、 $P(c|D_t)$ は時間 t におけるコンセプト c の生起確率であり、 $P(\hat{q}_j|D_t)$ は時間 t におけるコンセプト \hat{q}_j の生

起確率である。よって、式 (6) 中の $P(c|D_t) \prod_j^m P(\hat{q}_j|D_t)$ の値は、候補となるコンセプト c とクエリ中のコンセプト $\hat{q}_1, \hat{q}_2, \dots, \hat{q}_m$ が同時期に使用されると大きくなる。つまり、式 (6) は疑似適合文書 \mathcal{R}_t が表す時間帯において、コンセプト c の生起確率 $P(\hat{q}_j|D_t)$ 、 $P(c|D_t)$ を同時に追跡することで、コンセプトのリアルタイム性を表現している。一方、 $\sum_{D_l \in \mathcal{R}_l} P(D_l)P(c|D_l)P(Q|D_l)$ は式 (4) に示したコンセプトに基づく疑似適合フィードバック LCE と同じである。よって、式 (6) は、各コンセプトの語彙情報を時間情報で重み付けしたものと解釈できる。

式 (1) のクエリ尤度モデルの場合と同様に、各時間に対してクエリ中のコンセプトが出現しない場合、 $P(c|D_t)$ の値が0になってしまう。この問題に対処するため、 $P(c|D_t)$ に対しても以下のディリクレ平滑化を行う。

$$P(c|D_t) = \frac{|D_t|}{|D_t| + \mu_t} \hat{P}_{ml}(c|D_t) + \frac{\mu_t}{|D_t| + \mu_t} P(c|C) \quad (7)$$

ここで、 $|D_t|$ は時間 D_t における疑似適合文書中の文書数、 $P_{ml}(c|D_t) = \frac{f(c;D_t)}{\sum_{c' \in V_c} f(c';D_t)}$ 、 V_c はコーパス中のコンセプト、 $f(c;D_t)$ は時間 t での c の頻度を表す。 μ_t は平滑化パラメータである。最後に、クエリと潜在コンセプトの関連度を表す $S_{cTRM}(c, Q)$ の値が高い上位 k 件の潜在コンセプトをクエリ拡張に用いる。

4. 評価

4.1 実験設定

4.1.1 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。Tweets2011 コーパスは 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 個の検索トピックを持つ。検索トピックの例を図 1 に示す。図中の <num> がトピック番号、<title> がユーザクエリ、<querytime> がクエリの発行された時間 (クエリ時間) を示している。本実験では、TREC 2011 と 2012 のマイクロログトラックと同様に <title> を検索クエリとして用いる。

TREC の検索トピックには、任意の情報検索システムの評価を行うため、各検索トピックについて適合する tweet が明示されている。各 tweet は、所与のクエリに内容が関連していれば適合と判定される。また、適合度は非適合

<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.

表 2 評価に用いる TREC のテストコレクションの要約

Table 2 Summary of TREC collections and topics used for evaluation.

Name	Type	#Topics	Topic Numbers
TREC 2011	<i>allrel</i>	49	1-49
	<i>highrel</i>	33	1, 10-30, 32, 36-38, 40-42, 44-46, 49
TREC 2012	<i>allrel</i>	59	51-75, 77-110
	<i>highrel</i>	56	51, 52, 54-68, 70-75, 77-104, 106-110

(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 の種類を *allrel*, 非常に適合と判定された tweet の種類を *highrel* とし, 両者を分けて評価する. 実験に使用する検索トピックの数およびトピック番号をそれぞれ TREC 2011 と TREC 2012 の *allrel* と *highrel* に分けて表 2 にまとめる.

4.1.2 索引の作成

全 tweet データは, 各検索トピックのクエリ時間以前の tweet について索引を作成する. この索引を各検索トピック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ないようにするためである. この設定は, 実際のマイクロブログ検索を行う状況と同じである. 検索エンジンには, Indri [34] 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 [15] による接辞の除去 (stemming) を行う.

4.1.3 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 [36] を用いる. 語彙情報に対する平滑化パラメータは過去の研究 [12] に倣い, $\mu = 2500$ とする. この検索方法を LM とする. また, 時間情報に対する平滑化パラメータ μ_t については, 予備実験の結果から, *allrel* に対応するクエリで検索する場合は $\mu_t = 150$, *highrel* に対応するクエリで検索する場合は $\mu_t = 350$ とした. 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*^{*3} 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)^{*4} は TREC のマイクロブログトラックの指針 [29], [33] 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, ある話題に関する retweet と適合文書のタイムスタンプは類似することがあるため [8], retweet は適合文書と同じ話題を示している可能性があり, 適合文書の同定に役立つ文書だと考えられる. また, 予備実験の結果から, retweet を含めた上位の検索結果を疑似適合文書として用

^{*3} <https://github.com/shuyo/ldig>

^{*4} 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.

いる疑似適合フィードバックは, retweet を除いた疑似適合文書を用いる疑似適合フィードバックよりも検索性能が優れていると分かった. そこで, すべての疑似適合フィードバックを用いた手法に対して retweet を含む上位の検索結果を疑似適合文書として用いる. 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.

4.2 検索モデル

4.2.1 比較手法

まず, 提案手法のコンセプト追跡を用いた疑似適合フィードバックの実験設定について紹介し, 次に提案手法の各構成要素ごとの有効性を検証するために用意した比較手法について説明する. 提案手法では, 1 語以上の単語の組合せである潜在コンセプトを用いる. 単語を用いた疑似適合フィードバックと区別するため, 断りのない限り, コンセプトを用いた疑似適合フィードバックでは, 1 単語および 2 単語の組合せを潜在コンセプトとして用いる. 文書からコンセプトを抽出する際は, tweet 中の連続して出現する単語の組合せ (sequential-dependence) [25] だけを考慮する. 使用するコンセプトの種類は, 単語 w_1 と w_2 が文書中に連続して出現することを表すコンセプト $\#1(w_1, w_2)$ と, 単語 w_1 と w_2 が文書の 8 語以内に存在することを表すコンセプト $\#8(w_1, w_2)$ とする. また, 式 (7) において, コンセプトのコーパス中での頻度 $P(c|C)$ の計算コストが高いため, $df(c)/N$ を $P(c|C)$ の代替として用いる. ここで, $df(c)$ はコンセプト c の文書頻度, N を総文書数とする. $df(c)/N$ と $P(c|C)$ の値は異なるが, 平滑化パラメータ μ_t によって調整する. 式 (4) のコンセプトを用いた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によるクエリ拡張を cTRM とする.

さらに, 2 単語からなるコンセプトの使用の有効性を確かめるため, クエリが示す時間変化を考慮した疑似適合フィードバック手法 (wTRM) を提案する. wTRM は式 (4) において, コンセプト内の単語数が 1 つ, つまり単語だけを用いる手法である. wTRM は, cTRM と違い, 複数の単語の組合せであるコンセプトを考慮できない. wTRM と cTRM を比較することで, 時間情報を用いた検索モデルでのコンセプトの有用性を検証する.

提案手法の wTRM および cTRM の検索性能を評価するため, 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フィードバック手法 (wRM) [17] である. wRM は単語だけを用いるため, 2 語以上の単語からなるコンセプトを考慮せず, 時間情報も考慮しない.

2 つめの比較手法は, コンセプトの語彙情報を用いる疑

表 3 検索モデルの概要. Lexical, Temporal, Concept はそれぞれ語彙情報, 時間情報, コンセプトの使用を意味する

Table 3 Summary of the evaluated retrieval methods.

Method	Lexical	Temporal	Concept
wRM	✓		
cRM	✓		✓
wTRM	✓	✓	
cTRM	✓	✓	✓

似適合フィードバック (cRM) である. これは潜在コンセプト拡張 (LCE) と同等の手法である. ただし, 提案手法と比べてコンセプトの時間情報が有効であるかを検証するため, 式 (3) の単純化した LCE を用いる. この手法は, Bendersky ら [4] の LCE からコーパス中でのコンセプトの頻度に対する罰則を除去したものである. cRM と cTRM は, 両者とも潜在コンセプトを使用するものの, cTRM と違い, cRM は語彙情報だけを用いるため, 疑似適合文書 \mathcal{R}_t による時間情報を考慮していない.

表 3 に比較手法 wRM, cRM と提案手法 wTRM と cTRM の各構成要素の違いをまとめる. たとえば, 表 3 から, cRM と cTRM はコンセプトを使用でき, wRM と wTRM は使用できないことが分かる. また, wTRM と cTRM は単語またはコンセプトの時間情報を使用でき, wRM と cRM は使用できないことが分かる. 各手法について, 語彙情報だけを使う疑似適合フィードバック wRM と cRM がベースラインであり, 単語およびコンセプトの語彙情報と時間情報の両方とも用いる wTRM と cTRM が提案手法である.

なお, 本実験では tweet の検索を行う前にあらかじめ Tweets2011 コーパス中にあるすべての tweet に対してコンセプトの抽出を行った. そのため, 疑似適合フィードバックに要した時間は全手法 wRM, cRM, wTRM, cTRM とほぼ同程度であった. また, コンセプト抽出に要した時間は全コーパスで 1 日以内であった.

4.2.2 クエリ拡張の設定

すべての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. 候補語にはストップワードを用いず, 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. また, すべての疑似適合フィードバックに対して, 文書生起確率 $P(D_i)$ と時間の生起確率 $P(D_t)$ は一様とした. 次に, 各手法のクエリと候補語の関連度 $S_{cRM}(c, Q)$, $S_{cTRM}(c, Q)$ の値が大きい候補語の中から k 個の単語またはコンセプトを選択する. これをクエリ改良に使用するコンセプト (拡張コンセプト) とする. 各拡張コンセプトは疑似適合フィードバックによるクエリと拡張コンセプトの関連度 $S_{cRM}(c, Q)$, $S_{cTRM}(c, Q)$ を用いる. 各コンセプトの重みはそれぞれ上位 k 件の関連度のスコアで正規化を

```
#weight(
  λ1 #combine(bbc world service staff cuts)
  λ2 #weight( s1 #1(service outlines)
              s2 #uw8(bbc outlines)
              s3 outlines
              ...
              sk #1(weds bbcworldservice)))
```

図 2 TREC 2011 マイクロプログトラックで使用された検索クエリ MB001: “BBC world service staff cuts” についてのクエリ改良の例

Fig. 2 Example of query reformulation of topic MB001: “BBC world service staff cuts” from TREC 2011 Microblog track queries.

表 4 検索モデルのパラメータ

Table 4 Parameters for IR models.

Method	Year	allrel			highrel		
		M	N	K	M	N	K
wTRM	2011	10	30	10	10	30	40
cTRM		30	20	20	10	30	40
wTRM	2012	30	10	10	30	10	20
cTRM		20	30	40	10	30	40

行う.

本稿では, 所与のクエリとコンセプトの重み付けのため, Indri 検索エンジンの推論ネットワーク [34] の仕組みを用いる. 図 2 に, i 番目の拡張コンセプトの重みが $s_i = \frac{S_{cTRM}(c_i, Q)}{\sum_j^k S_{cTRM}(c_j, Q)}$ となるように正規化したクエリの例を示す. 実験では, 疑似適合フィードバックから取得した拡張コンセプトと所与のクエリとの組合せを 1:1 の重みで組み合わせる. よって, $\lambda_1 = \lambda_2 = 0.5$ とする.

4.2.3 検索モデルのパラメータ設定

すべての疑似適合フィードバック手法に対して, 語彙情報となる疑似適合文書の数 M , wTRM, cTRM に対する時間情報となる疑似適合文書の数 N や拡張コンセプトの数 k を調整する. パラメータの候補は, M を 10, 20, 30 まで, N を 10, 20, 30 まで, k を 10, 20, 30, 40 まで変化させ, これらのパラメータの組合せから最適なパラメータを見つける. 調整の方法は, すべてのパラメータに対して, 訓練データ中の allrel と highrel ごとに上位 1,000 件の検索結果に対する評価値 Average Precision (AP) がすべてのクエリの平均に対して最大となるパラメータを用いる. ここで, AP をパラメータ調整に使う理由は, Precision より識別と安定性ともに優れた指標であることが報告されているからである [6]. 具体的なパラメータの調整方法は, TREC 2012 のデータセットを用いて上位 1,000 件の AP 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. 表 4 に提案手法 wTRM と cTRM の調整後のパラメータ M, N, k の値を TREC 2011 と 2012 のデータセットごとに掲載する.

表 5 単語を用いた検索モデルの検索精度

Table 5 The performance comparison of the word-based PRFs.

Method	allrel			highrel		
	AP	P@30	bpref	AP	P@30	bpref
LM	0.2936	0.3981	0.3103	0.2130	0.1723	0.1933
wRM	0.3502 ^α	0.4543 ^α	0.3594 ^α	0.2473 ^α	0.2086 ^α	0.2242
wTRM	0.3726^{αβ}	0.4660^{αβ}	0.3872^{αβ}	0.2580^α	0.2094^α	0.2361^α

4.3 評価指標

本手法の目的は、適合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、Average Precision (AP) と上位 30 件の Precision (P@30) と binary preference (bpref) [7] を評価指標として用いる。P@30 は TREC 2011 マイクロブログトラックの公式の評価指標であり、TREC 2012 では P@30 と AP が評価に用いられた。一方、bpref は適合文書のデータセットが不完全な場合でも AP や Precision よりも頑健な評価指標である。TREC 2012 では、非常に適合する文書 (highrel) に対してだけ評価が行われたため、allrel と highrel を分けて評価する。本実験では、t 検定と違い評価値の母集団の分布が正規分布に従うことを仮定しない並べ替え検定 [32] を用いて、並べ替え数 100,000、有意確率 $p < 0.05$ で実験結果の有意差検定を行う。

4.4 実験結果

提案手法 wTRM と cTRM の検索性能を評価するため、wTRM、cTRM を標準的な疑似適合フィードバック wRM、cRM と比較する。

4.4.1 単語を用いた疑似適合フィードバックの比較

表 5 に初期検索 (LM)、単語の語彙情報に基づく疑似適合フィードバック [17] (wRM)、単語の語彙情報と時間情報に基づく疑似適合フィードバック (wTRM) の検索結果に対する AP、P@30、bpref の評価値を手法ごとに掲載する。LM、wRM、wTRM の手法に対して、有意に検索精度が向上した場合、それぞれ α , β , γ を各結果の添字で示す。また、表中の各欄で最も良い結果を太字で表す。

表 5 より、wRM と wTRM とともに allrel, highrel においてすべての評価指標で LM を著しく上回っていることが分かった。また、大半の評価指標で検索結果の差は有意であった。この結果は、単語を用いた疑似適合フィードバックが検索結果の改善に役立つことを意味する。さらに、allrel, highrel とともにすべての評価指標に対して wTRM は wRM を上回った。特に、allrel の AP と bpref において、その差は有意であった。これは提案手法の枠組みを用いて単語の時間情報を考慮することで、マイクロブログ検索の性能をさらに向上させることができることを示している。

4.4.2 コンセプトを用いた疑似適合フィードバックの比較

表 6 にコンセプトの語彙情報を用いた疑似適合フィードバック [4] (cRM)、コンセプトの語彙情報と時間情報に

表 6 コンセプトを用いた検索モデルの検索精度

Table 6 The performance comparison of the concept-based PRFs.

Method	allrel			highrel		
	AP	P@30	bpref	AP	P@30	bpref
LM	0.2936	0.3981	0.3103	0.2130	0.1723	0.1933
cRM	0.3385 ^α	0.4506^α	0.3479 ^α	0.2511 ^α	0.2060 ^α	0.2356 ^α
cTRM	0.3644^α	0.4485 ^α	0.3825^{αβ}	0.2694^β	0.2101^α	0.2527^α

表 7 時間情報を用いた検索モデルの検索精度

Table 7 The performance comparison of the proposed word- and concept-based temporal PRF methods.

Method	allrel			highrel		
	AP	P@30	bpref	AP	P@30	bpref
wTRM	0.3726	0.4660^β	0.3872	0.2580	0.2094	0.2361
cTRM	0.3644	0.4485	0.3825	0.2694	0.2101	0.2527

基づく疑似適合フィードバック (cTRM) の検索結果に対する AP、P@30、bpref の評価値を各手法ごとに掲載する。LM、cRM、cTRM の手法に対して、有意に検索精度が向上した場合、それぞれ α , β , γ を各結果の添字で示す。また、表中の各欄で最も良い結果を太字で表す。

表 6 より、cRM と cTRM とともに allrel, highrel においてすべての評価指標で LM を著しく上回っていることが分かった。また、すべての評価指標で検索結果の差は有意であった。この結果は、複数の単語の組合せをコンセプトとして用いた疑似適合フィードバックの有効性を意味する。特に、cTRM は大半の評価指標において cRM を上回っており、allrel の bpref と highrel の AP においてその差は有意であった。この結果は、提案手法の枠組みを用いてコンセプトの時間情報を考慮することで、マイクロブログ検索の性能をさらに向上させることができることを示している。

4.4.3 単語を用いた提案手法とコンセプトを用いた提案手法の比較

表 7 に単語の語彙情報と時間情報に基づく疑似適合フィードバック (wTRM)、コンセプトの語彙情報と時間情報に基づく疑似適合フィードバック (cTRM) の検索結果に対する AP、P@30、bpref の評価値を手法ごとに掲載する。wTRM、cTRM の手法に対して有意に検索精度が向上した場合、それぞれ α , β を各結果の添字で示す。また、表中の各欄で最も良い結果を太字で表す。

表 7 より、allrel においてはすべての評価指標で wTRM が cTRM を上回っており、P@30 での差は有意であることが分かった。しかし、P@30 はパラメータ調整に使用した評価指標ではない。一方、highrel に関しては有意差はないものの、すべての評価指標で cTRM が wTRM を上回っていることが分かった。この理由は、コンセプトを用いることで単語の曖昧性が解消されたため、より適合度の高い tweet を検索することができたからだと考えられる。たとえば、検索トピック “Gasland” *5 では、wTRM は oscar

*5 “Gasland” (MB109) は 2011 年のアカデミー賞ドキュメンタリー部門にノミネートされた映画。

や *nomination* といった単語をクエリ拡張の候補として用い、cTRM は関連語だけでなく *oscar nomination* といったコンセプトもクエリ拡張の候補として用いることができた。ここで、単語 *nomination* は任意の賞にノミネートされた作品に関する文書を誤って検索する可能性がある。一方で、コンセプト *oscar nomination* はアカデミー賞にノミネートされた作品に関する文書を検索できる可能性が高い。その結果、cTRM は wTRM と比較して、*allrel* より検索トピックに適合する確度が高い *highrel* の検索に有効であったと考えられる。

5. おわりに

本稿では、マイクロブログ上のリアルタイム性を考慮したコンセプト追跡によるマイクロブログ検索のモデルを提案した。本手法は、疑似適合フィードバック手法の考えを用いて、コンセプトが語彙情報と時間情報の2つの異なる情報源から成り立つと考え、語彙情報としての疑似適合文書中でのコンセプトの頻度、時間情報としての疑似適合文書中でのコンセプトの頻度の変化を言語モデルに基づく情報検索の枠組みに導入した。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、従来の語彙情報だけを用いる適合フィードバックよりも本提案手法が複数の評価指標で優れた検索性能を示すことが分かった。特に、コンセプトの語彙情報と時間情報を用いる手法は、非常に適合する文書の検索に対して優れた検索性能を示すことが分かった。

参考文献

- [1] Bendersky, M. and Croft, W.B.: Discovering key concepts in verbose queries, *SIGIR*, pp.491-498 (2008).
- [2] Bendersky, M. and Croft, W.B.: Modeling higher-order term dependencies in information retrieval using query hypergraphs, *SIGIR*, pp.941-950 (2012).
- [3] Bendersky, M., Metzler, D. and Croft, W.B.: Learning concept importance using a weighted dependence model, *WSDM*, pp.31-40 (2010).
- [4] Bendersky, M., Metzler, D. and Croft, W.B.: Parameterized concept weighting in verbose queries, *SIGIR*, pp.605-614 (2011).
- [5] Bendersky, M., Metzler, D. and Croft, W.B.: Effective query formulation with multiple information sources, *WSDM*, pp.443-452 (2012).
- [6] Buckley, C. and Voorhees, E.M.: Evaluating evaluation measure stability, *SIGIR*, pp.33-40 (2000).
- [7] Buckley, C. and Voorhees, E.M.: Retrieval evaluation with incomplete information, *SIGIR*, pp.25-32 (2004).
- [8] Choi, J. and Croft, W.B.: Temporal models for microblogs, *CIKM*, pp.2491-2494 (2012).
- [9] Dakka, W., Gravano, L. and Ipeirotis, P.G.: Answering general time-sensitive queries, *TKDE*, Vol.24, No.2, pp.220-235 (2012).
- [10] Efron, M. and Golovchinsky, G.: Estimation methods for ranking recent information, *SIGIR*, pp.495-504 (2011).
- [11] Efron, M.: Query-specific recency ranking: Survival analysis for improved microblog retrieval, #TAIA (2012).
- [12] Efron, M., Organisciak, P. and Fenlon, K.: Improving retrieval of short texts through document expansion, *SIGIR*, pp.911-920 (2012).
- [13] Jones, R. and Diaz, F.: Temporal profiles of queries, *TOIS*, Vol.25, No.3 (2007).
- [14] Keikha, M., Gerani, S. and Crestani, F.: Time-based relevance models, *SIGIR*, pp.1087-1088 (2011).
- [15] Krovetz, R.: Viewing morphology as an inference process, *SIGIR*, pp.191-202 (1993).
- [16] Lang, H., Metzler, D., Wang, B. and Li, J.-T.: Improved latent concept expansion using hierarchical Markov random fields, *CIKM*, pp.249-258 (2010).
- [17] Lavrenko, V. and Croft, W.B.: Relevance based language models, *SIGIR*, pp.120-127 (2001).
- [18] Lease, M.: An improved Markov random field model for supporting verbose queries, *SIGIR*, pp.476-483 (2009).
- [19] Li, X. and Croft, W.: Time-based language models, *CIKM*, pp.469-475 (2003).
- [20] Liang, F., Qiang, R. and Yang, J.: Exploiting real-time information retrieval in the microblogosphere, *JCDL*, pp.267-276 (2012).
- [21] Lin, J. and Efron, M.: Temporal relevance profiles for tweet search, #TAIA (2013).
- [22] Macdonald, C. and Ounis, I.: Global statistics in proximity weighting models, *SIGIR Web N-gram Workshop* (2010).
- [23] 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).
- [24] Metzler, D., Cai, C. and Hovy, E.: Structured event retrieval over microblog archives, *HLT/NAACL*, pp.646-655 (2012).
- [25] Metzler, D. and Croft, W.B.: A Markov random field model for term dependencies, *SIGIR*, pp.472-479 (2005).
- [26] Metzler, D. and Croft, W.B.: Latent concept expansion using Markov random fields, *SIGIR*, pp.311-318 (2007).
- [27] Miyanishi, T., Seki, K. and Uehara, K.: Combining recency and topic-dependent temporal variation for microblog search, *ECIR*, pp.331-343 (2013).
- [28] Miyanishi, T., Seki, K. and Uehara, K.: Improving pseudo-relevance feedback via tweet selection, *CIKM*, pp.439-448 (2013).
- [29] Ounis, I., Macdonald, C., Lin, J. and Soboroff, I.: Overview of the TREC-2011 microblog track, *TREC* (2011).
- [30] Peetz, M.H., Meij, E., de Rijke, M. and Weerkamp, W.: Adaptive temporal query modeling, *ECIR*, pp.455-458 (2012).
- [31] Ponte, J. and Croft, W.: A language modeling approach to information retrieval, *SIGIR*, pp.275-281 (1998).
- [32] Smucker, M.D., Allan, J. and Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation, *CIKM*, pp.623-632 (2007).
- [33] Soboroff, I., Ounis, I. and Lin, J.: Overview of the TREC-2012 microblog track, *TREC* (2012).
- [34] Strohman, T., Metzler, D., Turtle, H. and Croft, W.: Indri: A language model-based search engine for complex queries, *ICIA*, pp.2-6 (2005).
- [35] Teevan, J., Ramage, D. and Morris, M.: #Twitter-Search: A comparison of microblog search and web search, *WSDM*, pp.35-44 (2011).
- [36] Zhai, C. and Lafferty, J.: A study of smoothing meth-

ods for language models applied to information retrieval, *TOIS*, Vol.22, No.2, pp.179-214 (2004).



宮西 大樹

平成 23 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。平成 26 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修了。情報検索, Web マイニングの研究に従事。



関 和広 (正会員)

平成 14 年図書館情報大学情報メディア研究科博士前期課程修了。平成 18 年インディアナ大学図書館情報学研究科博士課程修了。神戸大学助教, 准教授等を経て, 現在, 甲南大学知能情報学部准教授。知能情報システムの研究に従事。Ph.D.. 人工知能学会, 自然言語処理学会各会員。



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。

(担当編集委員 齊藤 和巳)