

印象語をクエリとするリアクションに基づく ウェブ情報検索

莊 司 慶 行^{†1} 田 中 克 己^{†1}

本研究では、ウェブコミュニケーションデータを用いることで、「面白い」や「泣ける」などの印象語に基づいたウェブ情報検索を実現する。「C言語」に関する“分かりやすい”文書」を探す際、既存の検索エンジンでは、文書の本文中に「分かりやすい」という語が登場するかに基づき検索するため、望んだ文書を発見できない。ウェブ2.0サービスのログに含まれる、ウェブページに対するユーザのリアクションを利用し、クエリとして入力可能にすることで、このような読み手の印象による検索を可能にする。

Impression-based Web Search Using Readers' Reactions in Web2.0 Contents

YOSHIYUKI SHOJI^{†1} and KATSUMI TANAKA^{†1}

This paper proposes the “Impression-based web search” as a new web search model by using Web2.0 contents. Users can input impression queries into the system, such as “interesting,” “tear-jerker” and so on. Traditional web search engines could not find “easy-to-understand document about C-language” by the query “easy-to-understand, C-language” because they only focused on terms in main text of the document. The proposed system achieve above search with impression term queries by using readers' reactions in online communication sites like Twitter.

1. はじめに

現在一般的に用いられているウェブ検索エンジンは、発見対象となる文書の本文中にどの

ような語が含まれているかに基づいて検索を行う。そのため検索ユーザは、自分が望んでいるウェブページに登場するであろう語を推測し、クエリを作成する必要がある。

しかしながら、このような従来のウェブ検索を利用して近年のウェブで文書を発見するのは容易ではない。商用サイトやウェブコミュニケーションサイトが多く含まれ、ページの多様性もページ数も増している現代のウェブにおいて、必要とされる文書にのみ含まれるような語を推測し、適切なクエリを作成することは難しい。加えて、現在ではウェブの利用者も増加しており、コンピュータに詳しくない人もウェブ検索を利用する機会が増えている。しかしながら、このような人たちは、検索エンジンの仕組みに詳しくなく、文書中に登場する語という形式のクエリになじみがない。加えて、検索対象への理解が足りない場合もあり、自分の欲しい文書中にどのような語が含まれているか推測することが困難で、クエリが作れない場合がある。

初心者でも容易に利用可能な検索エンジンのクエリの形式として、“スカっとする”、“日記”、“分かりやすい”、“C言語ポイント”など、印象語とトピック語のペアからなるクエリを提案する。インターネットやコンピュータの初心者が、文書を探す際の例として、現実世界で書籍を探している場合をあげる。本文中にどのような単語が出現するかに基づくウェブではごく自然な検索は、インターネットの外の現実世界だと不自然である。友人や図書館の司書にどのような本が必要か聞かれて、「マニュアルとマイクロフトという登場人物の出てくる本が欲しい」や、「ルーベンスの絵を見て犬が昇天する話を教えてほしい」など、登場人物名や作品中に登場する固有名詞を伝えることは稀である。多くの場合、「面白いSFの本が欲しい」、「泣ける犬の話が読みたい」など、その本を読んだら実際にどのように感じるかを伝えることによって検索を行っている。このような検索要求を、

- (“かわいい”, “猫”)
- (“行ってみたい”, “イベント”)
- (“よくある”, “大学生活”)
- (“賛成”, “法案”)

というような印象語とトピック語の2種類のキーワードからなるクエリで表現する。このような形式のクエリであれば、コンピュータに明るくなかったり、また欲しい文書に含まれているであろう語を推測するのに十分な先行知識を持たないユーザでも、どのようなページが必要なのかをウェブ検索エンジンに十分に伝達可能になると考える。

しかし、このような印象語とトピック語をペアにしたクエリを取り扱いたいと考えた場合、現状の検索エンジンのアルゴリズムでは、満足な検索結果を得ることが難しい。

^{†1} 京都大学大学院情報学研究科社会情報学専攻

Department of Social Informatics, Graduate School of Informatics, Kyoto University

例として、一般のインターネットユーザの書いた、“面白い日記”を読みたい場合を考える。現状のウェブ検索エンジンに、“面白い、日記”というクエリを入力した場合、発見されるウェブページは“面白い”および“日記”という語を本文中に含むウェブページである。一方で、一般的に、本文中に“面白い”という語を含む文書が、実際に閲覧者に“面白い”という印象を与えるとは限らない。たとえば、「今日はとても面白かった」という、“面白い”やそれに類する語が多数含まれる日記が、閲覧者にとってまったく“面白い”と感じられない場合はままある。また逆に、「今日は散々な目にあった」という、“面白い”とは正反対の、陰惨な単語が多く出現するような日記の方が、閲覧者にとって“面白い”と感じられることも多い。このように、本文中に登場する印象語と、閲覧者が実際にコンテンツを閲覧した際に感じる印象の間には、本質的に関係がない。現在のウェブを、発見したいページの本文中に含まれる語に基づいて検索する従来の検索エンジンで検索した場合、この問題は特に顕著に現れる。商業用サイトは製品を売り込むために良いイメージの印象語をページ内に多く含み、また個人の日記や体験談は閲覧者がどう感じるかは無関係に印象語を多く含む。これらのウェブページの存在は、適合率の低下を招く。その一方で、淡々と事実を述べるような、ニュースや解説記事のようなページをユーザが求めることが多いが、これらのページは印象語を含まず、再現率の低下の原因となる。

印象をクエリとして入力可能なウェブ検索エンジンを実現するために、ウェブ 2.0 サービス上のウェブコミュニケーションに着目する。本提案手法では、ウェブ 2.0 サービス内で行われているオンラインコミュニケーションのログに含まれるリアクションを収集し、検索に利用する。Twitter^{*1}や SNS といったコミュニケーション目的のウェブ 2.0 サービスにおいて、他人にウェブページを紹介することは一般的な行為である。ウェブページを紹介されたユーザは、コミュニケーションの一環として、“面白い”や“泣ける”などといった反応（リアクション）をとる場合が多い。たとえば、あるユーザが友人たちに猫の画像の掲載されたウェブサイトを紹介した場合を考える。紹介されたユーザたちが、そのウェブページを閲覧したうえで、“キュートだ”、“抱きしめたい”といったリアクションをとっていた際、閲覧者に“かわいい”という印象を与えるウェブページだと推測することができる。あるウェブページを閲覧したユーザのリアクションを収集することで、ウェブページの本文中に出現する語とはまったく無関係に、その文書が読み手に与える印象を推測可能である。

本研究では、このような印象語をクエリとして入力可能なウェブ検索システムを実現する

手法を提案し、プロトタイプを実装したうえで、実験、評価を行った。以下で本稿の構成について述べる。2 章で本稿で提案する検索手法について、3 章で提案手法の有用性を検証する 2 つの実験とその結果についてそれぞれ述べ、4 章で実験結果の考察を行う。また、5 章で本研究と関連の深い他の研究について述べ、6 章でまとめと今後の課題について述べる。

2. 印象語をクエリとする検索

システムへの入力、読んだ際にどう感じるかを表す印象語クエリと、何に関する文書かを表すトピッククエリである。これら 2 つのクエリは文字列として入力される。検索対象はウェブページであり、スコアリングにはウェブコミュニケーションのログおよびウェブページの内容を利用する。システムの出力は、クエリにマッチするウェブページの順位付きのリストである。ページの順位は、ページと印象語クエリ、トピッククエリとの適応度によって決定される。この検索システムの概要を図 1 に示す。

システムの受け付ける印象語クエリは、読み手が文書を読んだ際にどう感じるかを表す語である。ここで注意しなければならないのは、ある 1 つの印象があった際に、それを表現する語は無数に存在することである。例として、あるウェブページを見て読み手が“分かりやすい”という印象をいだいた際、その印象をそのまま“分かりやすい”と表現することは稀である。一般的な形容詞的な印象語が入力された際、印象語そのものを含まないが同じ印象を意味する別表現は、大別して以下の 3 種がある。

- 換言表現：“簡単だ”、“難しくない”など、同じ意味の語に置き換えた、形容詞的な表現
- 因果に基づく表現：“詳細に書かれている”、“おかげで理解した”など、なぜ分かりや

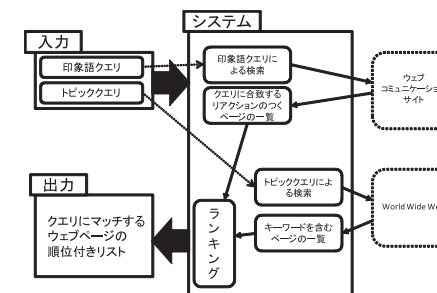


図 1 ウェブ検索システムの概要
Fig.1 Outline of the system.

*1 <http://twitter.com/>

すいのか、分かりやすいからどうなったのかを表す、助詞や動詞をともなう表現

- 感嘆表現：“へえ！”，“なるほど”など、単独では意味を持たない、感動詞による表現
- ウェブページに対するユーザのリアクションを情報検索に利用するためには、入力される印象語と実際に付いたリアクションに含まれる語が、同じ印象を表すものであったら異なった表現でも検索可能にする必要がある。またユーザが入力する印象語クエリについても、ユーザがツェにある印象を直接表す形容詞や形容動詞だけでクエリを作るとは限らず、“キュンとするウェブページ”や“ドキドキなウェブページ”など印象語の別表現をクエリとして入力することが考えられる。そのため、印象語クエリとリアクション双方において、ある印象とそれを表す語を適切にマッチングしてスコアを計算する必要がある。

本稿における提案手法では、印象語をそのまま含むリアクションを足がかりとして、同じ印象を表す別のリアクションを推定し、またその推定を利用してウェブページのランキングを行う。

2.1 検索モデル

すべての検索対象のウェブページ集合 P 、全単語を V とした際、入力されるクエリ q は $q = (q_i, q_t)$ で表される。ただしこの際、 $q_i \subset V$ 、 $q_t \subset V$ である。入力されたクエリ q を構成する印象語クエリ q_i 、トピッククエリ q_t について、提案する検索手法は q_i というリアクションの付く q_t に関するウェブページ p の集合をランクづけして返す。

あるウェブページ p が印象語クエリ q_i およびトピッククエリ q_t とどれだけ適合するかを表すランキング関数 $rank(p, q_i, q_t)$ を式 (1) のように定義する。

$$rank(p, q_i, q_t) = score_T(p, q_t)^\alpha \cdot score_R(p, q_i)^\beta \quad (1)$$

ここで $score_T(p, q_t)$ はあるウェブページ p がトピッククエリ q_t に対してどれだけ適合するかの度合いであり、 $score_R(p, q_i)$ はウェブページ p が印象語クエリ q_i に対してどれだけ適合するかの度合いである。これらの関数に関してはそれぞれ 2.2 節、2.3 節で述べる。 α 、 β はトピッククエリと印象語クエリそれぞれの重要度を表す重みである。 α の値を上げるとトピッククエリとの適合度のより高い結果が検索結果上位に出現しやすくなる。このようなモデルを実現するうえで、最も大きな技術的課題は、印象語と実際にとられるリアクションとの対応付けである。あるウェブページ p と入力された印象語クエリ q_i の適合性計算を行う際、 p に付くあるリアクション $r \in R(p)$ について、まったく別の表現がなされていても、同じ印象を表すリアクションかどうかを判定したうえでスコアリングを行う。

2.2 トピッククエリに対するウェブページの適合性

あるページ p がトピッククエリ q_t に対してどれだけの適合性を持っているかの計算を、以下の式 (2) で求める。本稿では、最も単純化しスムージングを排したクエリ尤度モデル¹⁰⁾を用いた。

$$score_T(p, q_t) = Pr(p|q_t) \propto \prod_{t \in q_t} Pr(t|M_p) \quad (2)$$

ここで M_p はウェブページ p を生成した言語モデルである。 $Pr(t|M_p)$ は、トピッククエリ q_t に含まれる語 t が、ウェブページ p 中に登場する回数を $tf_{t,p}$ 、ウェブページ p の長さを L_p としたとき、式 (3) で推定される。

$$Pr(q_t|M_p) = \prod_{t \in q_t} Pr_{mle}(t|M_p) = \prod_{t \in q_t} \frac{tf_{t,p}}{L_p} \quad (3)$$

2.3 印象語クエリに対するウェブページの適合性

あるウェブページ p に付いているリアクションが、どれだけ印象語クエリ q_i に対して適合性が高いかをページごとにとりまとめたスコア $score_R(p, q_i)$ について説明する。

リアクションとは、ある文書を読んだ際に読み手が受け取った印象が具体性をともなって自然言語で記述されたものであり、印象そのものではない。そこで、印象語クエリ q_i に対するウェブページ p の適合性を計算するためには、入力された印象の別表現によるリアクションについて考慮する必要がある。例として、“かわいい”という印象語クエリに対し、自然言語で記述された“わぁ！とてもキュートな猫だね”、“ぎゅっとしたい(*´`*)”など、“かわいい”印象を表す多様なリアクションを検索対象とする。本手法では、印象語をそのまま含むリアクションを足がかりに、印象語を直接含まないが同じ印象を表すリアクションを推測し、スコアリングに利用する。

ウェブページ p に付くリアクション集合を $R(p) = \{r_1, \dots, r_k\}$ としたとき、それぞれ個別のリアクション $r \in R(p)$ の q_i に対する適合度に関して、平均をとったものをページ p と印象語クエリ q_i の適合度 $score_R(p, q_i)$ として用いる。これを以下の式 (4) で表す。

$$score_R(p, q_i) = \frac{\sum_{r \in R(p)} s_r(r, q_i)}{|R(p)|} \quad (4)$$

これら個別のリアクション r の入力された印象語クエリ q_i に対する適合度 $s_r(r, q_i)$ は、リアクション r を構成する複数の語 $W(r) = \{w_1, w_2, \dots, w_j\}$ についてそれぞれ印象語クエ

3. 実験

提案するウェブ検索手法について、その有用性に関して評価を行うため、実際に動作するウェブアプリケーションを実装し、実験を行った。

3.1 実装

実験のため、2章で提案したアルゴリズムにのっとって動作するプログラムを、ウェブブラウザからアクセス可能なウェブアプリケーションプログラムとして実装した。作成したウェブアプリケーションのスクリーンショットを図3に示す。2つ並んだテキストボックスの左側に印象語クエリを、右側にトピッククエリをそれぞれ入力することで、上述のランキング関数に従って順序付けされた検索結果が得られる。この際、発見したウェブページのタイトル、URL、スニペットのほかに、実際にそのウェブページに対してなされたリアクションも表示される。

リアクションはウェブコミュニケーションサイト内でウェブページを紹介しあっているような局面であればどこからでも取得可能であるが、本稿における実験のための実装では、Twitter内の発言の中で、あるユーザが他人のURLを含む発言を引用しつつ、そのURLについてコメントしている形式のものを抽出し、コメント部分のみをリアクションとして利用した。収集したTwitter内の書き込みの具体例を図4に示す。この例において、検索の対象となるURLは“<http://www.dl.kuis.kyoto-u.ac.jp/>”であり、このウェブページを読んだ際のリアクションが“こんなに面白いWebサイト見たことないwww”である。実際にアプリケーションとして実装する際、処理の流れは以下のようにした。

- (1) 前処理としてウェブコミュニケーションサイトからリアクションを収集し、URLとリアクションの対応付けを行い利用可能な状態にする。
- (2) ユーザの入力した印象語クエリとトピッククエリを受け取る。
- (3) 印象語クエリより、2.3節の手法を用いて類似した意味の別表現を探し、単語のスコアを決定する。
- (4) リアクションの付いているウェブページの一覧に対して、トピッククエリで文中語による検索を行う。
- (5) 手順(3)より印象語クエリとの適合度を、手順(4)よりトピッククエリとの適合度を算出し、ランキングを行い出力する。

実装にはスクリプト言語であるRubyを使用し、作成したシステムはDebian GNU/Linux上のApache2によりCGIとして動作する。リアクションおよびウェブページ本文の形態

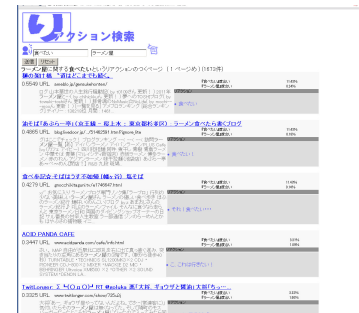


図3 実装したアプリケーションのスクリーンショット
Fig.3 A screenshot of the implemented system.



図4 Twitterから取得したリアクションの例
Fig.4 An example of the reaction in Twitter.

素解析にはMeCabを、MeCabの辞書にはIPA辞書を用いた。この際、文字数のきわめて少ないものや意味不明な記号などのストップワードを除き、すべての品詞の語をスコア計算に利用した。

3.2 実験1: 同じ印象を表す語の推定の評価実験

2.3節で述べた同じ印象を表す語の推定手法の有用性について、既存手法との比較実験を行った。同じウェブページに付いた複数のリアクションを1つの文書と見なした場合、検索結果を利用して印象語クエリを拡張しているといえる。そこで本稿で提案した手法の有用性を明らかにするため、擬似適合性フィードバックに基づくクエリ拡張モデルを作成し、それぞれの出力結果を手で評価することで比較した。

ベースライン手法として擬似適合性フィードバック¹⁸⁾に基づき、フィードバックに用いる語を拡張クエリとして使用するクエリ拡張システムを作成した。このシステムでは、拡張前の印象語クエリをそのままの形で含むリアクションによりウェブページをランキングし、そのうち上位 k 件の結果 P_K について適合する文書と見なし、これらのページに対するリアクションに多く含まれる語を拡張クエリとして推薦する。 k 位のウェブページ $p_k \in P_K$ のスコアは、以下の式(7)のように定義される。このスコアは上位 k 件を発見するためだけに用いられる。

$$score_{BL}(p_k) = |R_{include}(q_i) \cap R(p_k)| \quad (7)$$

語 w_j のスコア $s_{BLw}(w_j)$ は $p_k \in P_K$ へのリアクションに語 w_j を含むリアクションがいくつ含まれているかであり、式(8)によって定義される。

表 1 語ごとの評価結果
Table 1 Score of each term.

印象語	提案手法		ベースライン手法	
	p@10	p@30	p@10	p@30
なるほど	0.60	0.33	0.10	0.03
納得	0.20	0.10	0.00	0.00
簡単	0.20	0.07	0.00	0.00
泣ける	0.80	0.67	0.60	0.41
面白い	0.50	0.50	0.60	0.31
感動	0.70	0.47	0.30	0.31
怖い	1.00	0.60	0.10	0.03
意外	0.10	0.03	0.00	0.00
素敵	0.10	0.07	0.30	0.10
酷い	0.60	0.30	0.20	0.17
MAP@10	0.86		0.37	
MAP@30	0.73		0.31	

$$s_{BLw}(w_j) = \sum_{k=1}^K |R(p_k) \cap R_{include}(w_j)| \quad (8)$$

本実験では、 $K = 10$ とし、上位 10 件のウェブページへのリアクションに多く含まれる語を拡張クエリとして利用した。

評価対象は与えられた印象語クエリ q_r に対して $R_{reaction}(q_r)$ に含まれるすべての w_j のうち、印象語クエリ q_r と同じ印象を表す語であると推測された語、上位 30 件である。実際に評価を行った q_i の一覧を表 1 に、両システムの出力結果の一例を表 2 に示す。評価者は 2 名であり、それぞれの語について、印象語クエリと同じ意味を表す語であるかを 0 から 2 の 3 段階で評価した。2 名の評価者の評価を加算して、3 を超えたものについて適合として扱う。

3.3 結果 1

結果を表 1 にまとめる。評価尺度はそれぞれ、10 件目および 30 件目までにおける適合率を評価する尺度である p@10, p@30, 全クエリを通してのランキングの適合率に基づく評価尺度である MAP@10 および MAP@30 である。MAP@10, MAP@30 において、提案手法はベースライン手法よりもきわめて高い値を出した。これより、全体を通しての精度は提案手法の方が優れているといえる。また、ほぼすべてのクエリのほぼすべての k において、クエリごとの適合率指標である p@ k で提案手法はベースライン手法よりも高く評価さ

表 2 “感動”と同じ印象を表すと判定された語
Table 2 Term expressing “impressive” suggested by the system.

提案手法	ベースライン手法
感動	感動
工房	いい
涙腺	ない
ぐっと	泣い
泣け	涙
鳥肌	T
泣き	すごい
号泣	素敵
再び	けど
泣く	たい
泣ける	泣ける
一言	って
熱く	この
・(じゃ
不覚	!!
素直	宇多田
(;_;	すげ
もつ一度	読ん
泣い	泣き
;)`)	なっ
文章	子供
ぶさ	素晴らしい
必読	なり
;’)	良い
改めて	泣け
止まら	こういう
ブワッ	泣
涙	ライブ
安心	号泣
すばらしい	たら
秀逸	こんな

れた。一方で、 $q_i = \text{“面白い”}$ 、 $q_i = \text{“素敵”}$ において 0 から 30 までのすべての k において、提案手法がベースライン手法を下回った。 $q_i = \text{“面白い”}$ では、結果の上位に“試み”、“アイデア”、“視点”、“発想”などの語が並んだ。これはこれらの語が Twitter 内で使われる局面に限られており、“面白い”という語と特徴的に強く共起する語だったからと考えられる。一方でベースライン手法の上位で評価者に高く評価された語は“笑”、“www”、“ww”などである。これらの語はそもそも使われることが多く、様々な印象を表すときに用いられ

るため、提案手法では低くランク付けされたと考えられる。 $q_i = \text{“素敵”}$ では両手法ともに適切な同じ感動を表す語が取得できていない。原因として、“素敵”自体がきわめて漠然とした印象を表した語であることが考えられる。提案手法ではことさら強く共起する語を発見できなかったため、特に低い結果が出た。また、 $q_i = \text{“意外”}$ においても同様のことがいえ、両手法ともに低い精度の結果となった。

3.4 実験 2: 印象語によるウェブ検索ランキングの評価実験

提案手法について、システム全体を通して結果の精度に関して、人手で作成した正解セットをもとに既存手法との比較実験を行った。

本研究で提案したリアクションを利用した印象語に基づくウェブ検索エンジンは、以下の 2 つの性質を有する。

- (1) 検索対象となる本文中に含まれる語ではなく、そのウェブページに対するウェブコミュニケーションサイトにおけるリアクションを利用することで印象語による検索を行う。
- (2) 入力された印象語と同じ印象を表す別種の表現をともなったりアクションについて同じウェブページへの異なるユーザからの言及に着目することで推定し、検索に利用する。

これら 2 点について比較するため、提案手法と 2 つのベースライン手法を含む 3 つの検索システムを実装し、評価を行った。

- 手法 1: 本文中に出現する語に基づく検索システム
- 手法 2: 印象語そのものを含むリアクションのみを利用する検索システム
- 提案手法: 印象の多様な表現をともなうリアクションを利用する検索システム

手法 1 は本文中に出現する語に基づく検索システムであり、現在一般的に利用されている従来のウェブ検索エンジンと同じである。このシステムでは、受け取った印象語クエリ、トピッククエリの両クエリについて、対象文書の本文中に含まれるかどうかに基づく検索の結果を返す。ランキングはクエリ尤度モデルに基づき決定され、各ページのスコアは式 (2) のとおり、印象語クエリ、トピッククエリは両方ともトピッククエリとして扱われる。

手法 2 は提案手法と同じように、オンラインコミュニケーションサイトにおけるウェブページへのリアクションを検索に利用する。ただしこの際に、入力された印象語について、それをそのまま含むリアクションが付いたウェブページのみを適合とする。結果の順位は式 (9) で表されるランキング関数 $score(p, q_i, q_t)$ で決定される。

$$score(p, q_i, q_t) = score_T(p, q_t)^\alpha \cdot |R_{include}(q_i) \cap R(p)|^\beta \quad (9)$$

$score_T(p, q_t)$ 、 α および β は提案手法における記述と同じである。この値はウェブページ p の本文中にトピック語 q_t が出現する回数とページ本文の長さ、 p に対するすべてのリアクション $R(p)$ 中に q_r が登場する回数によって計算される。

これら両システムに同一の検索対象ページ群とクエリの組を与え、検索結果のランキングに含まれる各ウェブページについて、クエリとの適合度を人手で評価することで、検索システムの評価を行った。

評価のために 28 組のクエリをあらかじめ作成した。クエリの形式は印象語クエリ、トピッククエリ 1 つずつの組合せである。実際に利用したクエリを表 3 に記す。評価者は全 8 名であり、それぞれ 14 件のクエリを担当し、すべてのクエリについて 4 名の評価者が正解セットを作成した。評価対象となったウェブページは、与えられたクエリから各システムが出力した結果の上位 20 件である。ただし、検索結果が 20 件に満たない場合には、出力されたすべての検索結果に対して評価を行った。評価時には、3 つのシステムの出力はクエリごとに 1 つのページにまとめられ、評価対象のウェブページがどのシステムのどの順位の出力か分からないようにした。

各ウェブページは、トピッククエリとの適合度と印象語クエリとの適合度を別々に評価された。ページと印象語クエリの適合度判定の基準は以下のとおりである。

- 1: 与えられた印象語と同じ印象を受ける。
- 2: 人によっては与えられた印象語と同じ印象を受けると思う。
- 3: 与えられた印象語と同じ印象は受けない。
- 4: 与えられた印象語とむしろ逆の印象を受ける。

ページとトピッククエリの適合度判定の基準は以下のとおりである。

- 0: トピッククエリと一致する。
- 1: トピッククエリと一致しない。

実験を行う際に用いたデータについて述べる。2010 年 11 月より 2010 年 12 月にかけて、Twitter ストリーミング API で無作為に取得した URL を含むツイートのうち、日本語で記述され、なおかつ先述した形式に則ったものをすべて利用した。収集した日本語ツイート総数は 30,134,604 件であり、このうち 2,370,222 件がリアクションとして利用可能であった。リアクションの付いた URL の指すウェブページは、短縮 URL による重複を取り除いたうえで 1,286,752 種類であり、本実験ではこれらのウェブページとそれに付いたリアクション

表 3 クエリと発見件数
Table 3 Query and the number of page retrieved.

q_r	q_t	発見ページ数			q_i を含んだ ページ数	q_i を含んだ 正解ページ数	P@20		
		手法 1	手法 2	提案手法			手法 1	手法 2	提案手法
なるほど	風邪 仕組み	52	6	215	3	0	0.00	0.10	0.15
納得	失業 原因	93	9	457	3	0	0.10	0.05	0.25
なるほど	語源 由来	20	6	190	4	2	0.15	0.25	0.70
分かりやすい	クラウド	331	13	*4,031	1	0	0.10	0.05	0.25
簡単	ダイエット	*3,000	28	3,847	13	0	0.10	0.00	0.10
参考になる	勉強法	53	6	762	1	1	0.35	0.10	0.40
泣ける	デート 日記	54	6	1,237	3	0	0.00	0.00	0.00
面白い	デート 日記	346	13	1,231	7	1	0.00	0.00	0.00
感動	犬 思い出	437	10	1,238	5	4	0.20	0.00	0.25
泣ける	受験 日記	45	6	1,434	3	3	0.05	0.10	0.10
怖い	病院 体験	455	35	1,827	10	0	0.10	0.00	0.00
痛い	中学 思い出	240	9	713	7	0	0.15	0.00	0.00
かわいい	犬 画像	*3,000	290	*4,001	11	0	0.20	0.35	0.35
ムカつく	逮捕 ニュース	219	6	*3,699	2	0	0.05	0.05	0.10
意外	芸能ニュース	964	14	*3,506	4	0	0.30	0.05	0.25
素敵	年賀状	745	135	*4,006	6	0	0.20	0.40	0.20
驚いた	ニュース サッカー	747	18	*4,063	0	0	0.00	0.10	0.15
作りたい	レシピ トマト	47	7	670	2	1	0.00	0.00	0.05
飲んでみたい	カクテル	7	8	1,228	1	2	0.15	0.20	0.55
面白い	ミステリ 小説	124	8	321	8	1	0.20	0.05	0.30
美味しそう	ラーメン 京都府	19	5	1,085	0	0	0.00	0.05	0.20
行ってみたい	レストラン 京都	51	19	1,839	12	0	0.15	0.20	0.50
懐かしい	ゲーム RPG	357	37	*3,904	5	2	0.15	0.20	0.15
面白い	アクション映画	58	5	187	3	0	0.30	0.00	0.25
欲しい	アイデア 商品	641	85	2,505	4	1	0.15	0.10	0.15
かわいい	猫 画像	*3,000	526	*4,038	8	0	0.95	0.60	0.60
ぶさいく	猫 画像	14	1	226	0	0	0.10	0.00	0.00
酷い	犬 画像 ペット	223	15	1,666	6	1	0.10	0.00	0.10

を対象として取り扱った．また，印象語クエリとトピッククエリそれぞれの重要度を表す重みは $\alpha = 0.3$, $\beta = 1$ とした．

3.5 結 果

本実験における評価項目および利用した尺度について述べる．本実験では適合率に基づく評価基準として $P@k$ を各クエリごとにそれぞれ@5, @10, @30 について計算した．本実験においては，対象となるウェブページ数が 1,286,752 ページと膨大であるため，再現率を計算することが困難である．再現率の代わりに評価尺度として，発見したページ数

および擬似適合率を算出した．この評価尺度では，3 システムの検索結果の中にすべての正解が含まれると仮定し，上位 k 件での再現率を議論する．また当手法に限られた記述として，提案手法で得られた上位 20 件のウェブページの本文中に印象語クエリ q_t が現れていた数を記載する．ベースライン手法と比較した際に，同じウェブページがランキングに登場しており，さらにそのウェブページがクエリに適合すると判定された数についても記載する．

はじめに，印象語クエリによる適合度とトピッククエリによる適合度の両方に基づくランキングについて評価を行う．このランキングでは，評価されたすべての結果について，トピッククエリに適合しておりなおかつ印象語クエリにも適合しているものを正解ページと見なす．トピッククエリとの適合は，4 名の評価者のうち 3 名以上がトピッククエリと適合していると判断したのものについて適合と見なす．印象語クエリとの適合は，4 名の評価者が付けた 1 から 4 の適合度について，全員の平均が 3 以上だったものについて適合文書と見なす．各クエリについてシステムが発見したページ数について，表 3 に記す．本実験ではベースライン手法において 3,000 件，提案手法において 5,000 件を上限としてページを取得している．表 3 中の * は，検索結果数が上限を超えており，実際にはもっと多いことを表す．表 3 中の q_i を含んだページ数， q_i を含んだ正解ページ数はそれぞれ提案手法で得られたウェブページの中に印象語 q_i が含まれていた数，および提案手法で適合文書と判断された文書に印象語 q_i が含まれていた数である．各システムの $p@20$ について表 3 に表す． $p@20$ は適合率ベースの評価尺度であり，システムを通した評価としてこの平均をとると提案手法が 0.22，手法 1 が 0.15 および手法 2 が 0.11 となった．全ユーザ，全クエリを平均しての $p@20$ について，T 検定を行った．この際 P 値は手法 1 と提案手法の間で 0.03，手法 2 との間で 0.00 である．これにより， $p@20$ においては提案手法の適合率はベースラインとして設定した 2 つの手法に対して優位であるといえる．

本稿で提案する手法は印象語による検索を主眼としている．そこでトピッククエリによって不適合として判断されたウェブページを取り除き，順位を上詰めたランキングを作成した．このランキングについて，印象語クエリとの適合性のみに基づいた評価を行う．結果を表 4 に表す．20 位まで評価されていた結果からトピッククエリで不適合と判断されたページを除去したことで，各クエリの結果件数は表 4 の件数の数まで減った．ここで，手法 1，手法 2 と適合率ベースでの比較を行うために， $P@k$ による評価を行った．ただし，検索結果が何位まであるかがクエリ，システムによって異なるので，各クエリにおいて比較する 3 システムの検索結果のうち最も少ない件数を k 件とし，クエリごとに 3 システムの上位 k 件の結果について評価を行った．次に，先ほどと同様に擬似再現率を計算した．ここでの擬似再

表 4 トピッククエリ適合ページ内での評価
Table 4 Results of topic matched pages.

q_r	q_t	件数			p@k			擬似再現率			
		手法 1	手法 2	提案手法	k	手法 1	手法 2	提案手法	手法 1	手法 2	提案手法
なるほど	風邪 仕組み	0.00	2.00	3.00	0.00	-	-	-	0.00	0.22	0.33
納得	失業 原因	6.00	2.00	6.00	2.00	0.00	1.00	1.00	0.10	0.10	0.30
なるほど	語源 由来	6.00	5.00	17.00	5.00	0.60	1.00	1.00	0.18	0.23	0.73
分かりやすい	クラウド	5.00	1.00	8.00	1.00	0.00	1.00	1.00	0.13	0.06	0.31
簡単	ダイエット	12.00	2.00	11.00	2.00	0.50	0.00	0.00	0.27	0.00	0.45
参考になる	勉強法	12.00	2.00	9.00	2.00	1.00	1.00	1.00	0.43	0.10	0.38
泣ける	デート 日記	0.00	0.00	0.00	0.00	-	-	-	0.00	0.00	0.00
面白い	デート 日記	0.00	1.00	2.00	0.00	-	-	-	0.00	0.00	0.00
感動	犬 思い出	8.00	0.00	5.00	0.00	-	-	-	0.67	0.00	0.83
泣ける	受験 日記	1.00	2.00	2.00	1.00	0.50	1.00	1.00	0.10	0.20	0.20
怖い	病院 体験	4.00	0.00	1.00	0.00	-	-	-	0.25	0.00	0.00
痛い	中学 思い出	5.00	2.00	4.00	2.00	1.00	0.50	0.00	0.33	0.08	0.08
かわいい	犬 画像	4.00	7.00	7.00	4.00	1.00	1.00	1.00	0.17	0.30	0.30
ムカつく	逮捕 ニュース	11.00	1.00	6.00	1.00	1.00	1.00	1.00	0.38	0.05	0.24
意外	芸能ニュース	16.00	7.00	14.00	7.00	0.57	0.14	0.43	0.38	0.05	0.33
素敵	年賀状	18.00	17.00	13.00	13.00	0.23	0.54	0.54	0.29	0.48	0.33
驚いた	ニュース サッカー	3.00	3.00	5.00	3.00	0.33	1.00	1.00	0.06	0.17	0.28
作りたい	レシピ トマト	8.00	1.00	7.00	1.00	0.00	0.00	0.00	0.33	0.00	0.33
飲んでみたい	カクテル	3.00	6.00	14.00	3.00	1.00	0.67	1.00	0.18	0.29	0.76
面白い	ミステリ 小説	14.00	4.00	15.00	4.00	0.75	0.25	0.50	0.46	0.08	0.69
美味しそう	ラーメン 京都府	0.00	1.00	4.00	0.00	-	-	-	0.00	0.05	0.19
行ってみたい	レストラン 京都	3.00	4.00	10.00	3.00	1.00	1.00	1.00	0.09	0.12	0.29
懐かしい	ゲーム RPG	6.00	6.00	7.00	6.00	0.50	0.67	0.33	0.14	0.19	0.14
面白い	アクション映画	9.00	1.00	12.00	1.00	1.00	0.00	1.00	0.43	0.00	0.36
欲しい	アイデア 商品	6.00	5.00	5.00	5.00	1.00	1.00	0.60	0.43	0.29	0.21
かわいい	猫 画像	20.00	13.00	15.00	13.00	0.92	1.00	1.00	0.40	0.27	0.29
ぶさいく	猫 画像	4.00	0.00	10.00	0.00	-	-	-	0.75	0.00	0.00
酷い	犬 画像 ベット	4.00	0.00	2.00	0.00	-	-	-	0.07	0.00	0.07

表 5 トピッククエリ適合ページ内での MAP および MRR
Table 5 MAP and MRR of topic matched pages.

	MAP	MRR
手法 1	0.57	0.67
手法 2	0.56	0.55
提案手法	0.60	0.64

表 6 “なるほど”、“語源 由来”の提案手法による結果例
Table 6 Example of the result by query (“なるほど”、“語源 由来”).

ページの概要	本文中	リアクション中	印象評価
オンライン辞書の語源記事	含まない	含む	4.0
一般サイトの語源の記事	含まない	含む	3.8
差別語の啓蒙記事	含む	含む	3.5
オンライン辞書の語源記事	含まない	含む	4.0
個人の辞書サイトの語源記事	含まない	含む	3.5
オンライン辞書の語源記事	含まない	含まない	3.0
オンライン辞書の語源記事	含まない	含まない	3.8
オンライン辞書の語源記事	含まない	含まない	4.0
名づけに関する一般記事	含む	含む	2.5
オンライン辞書の語源記事	含まない	含まない	3.3

現率は、各システム各クエリの最後の順位まで評価している。このランキングはトピッククエリによる不適合の影響を受けず、トピッククエリとの適合度のみによるランキングである。このランキングについて、ランキングの総合的な適合率ベースの指標 MAP と、必要なページがいかに上位にランキングされるかを評価する MRR で評価した。結果を表 5 に記す。

4. 考 察

本実験で得られた結果について考察を行う。

全体を通して、結果の取得件数に関して、最も多くのページを発見できるのは提案手法であった。これはベースラインであった手法 1 に対しては、本文中に印象語を含まないページを検索できるからであり、手法 2 と比較した場合、直接印象語そのものを含まないリアクションについてもランキングの対象とできるからである。検索結果の取得数が大幅に増えているにもかかわらず、検索結果の適合率は既存手法に勝っている。手法 2 および提案手法では、従来の文中に含まれている語をクエリとする検索エンジンでは検索できないページが発見された。

特にこのような兆候が見られたクエリの例として、“なるほど”、“語源 由来”のタスクがあげられる。このタスクでは提案手法によって 16 件の適合ウェブページが発見されたが、これらのうち“なるほど”という語を本文中に含んでいるページは 2 ページであった。評価者が高く評価したウェブサイトには一般のオンライン辞書サイトの記事が多く含まれた。これらのサイトには“なるほど”や、それに類する語はいっさい含まれていない。提案手法の結果の上位 10 件のウェブページについて、その概要と本文中に“なるほど”という語を含むか、およびリアクション中に“なるほど”という語を含むかを表 6 に示す。このタスクで検索上位のウェブサイトが付いたリアクションで“なるほど”を含まないものは、

- へー！
- へえへえへえ
- 勉強になります
- ふむ。

であった。

逆に、提案手法がベースライン手法に対して適合率的に劣ったタスクとして、“痛い”、“中学 思い出”、“ムカつく”、“逮捕 ニュース”などがあつた。これらのクエリにおける検索結果では、ウェブコミュニティサイトをまとめたブログサイトが多数、正解ページとして評価された。こういったウェブコミュニティサイトのまとめサイトは、あるニュース記事に対するウェブフォーラムの反応を紹介しているものが多く、単独のウェブページ内にニュース記事とそれに対する閲覧者の印象の両方を含んでいる。また、ブログのコメント欄に書かれた書き込みが印象語クエリとマッチしてページが発見された。ブログのコメント欄は作者ではなくそのブログの閲覧者が書いたものであるため、この場合も単独のウェブページがコンテンツと閲覧者の印象の両方を含んでいる。こういったコンテンツについては、既存の文中に含まれる語に基づくウェブ検索エンジンでも、入力された印象語と一致するウェブページが取得可能であった。

一方で、実験を通していくつかの欠点も発見された。今回評価に用いたクエリの中で、“ぶさいく”や“酷い”というクエリで、ベースラインとなる手法1が、提案手法およびベースライン手法2を適合率で上回った。今回はリアクションの収集元をTwitterに限定し、誰かにページを紹介するというコミュニケーションの中からリアクションを抽出した。そのため、ここでのリアクションは紹介された人のリアクションとなるが、この際、紹介されたユーザは、紹介してくれた相手がいるため、あまり否定的なリアクションをとりにくい。多くの場合、気に入らないウェブページを紹介するユーザは、否定的なリアクションをとることなくその発言を無視する。リアクションはページに対する直接的なアノテーションとは異なり、コミュニケーション中に副産物として生じるものである。そのため、人間関係やコミュニケーションのためのデータの偏りとして、嘘やお世辞などが含まれる。特に、印象語クエリ“ぶさいく”のケースでは、テストセット内に“ぶさいく”を含むリアクションが合計15件しか登録されていなかった。そのため同じ印象を表す別表現のリアクションの推定もきわめて限られたデータのみに基づいて行われ、精度が低かった。

5. 関連研究

本研究を行ううえで、テーマ的、技術的な関連を持つ研究について説明する。

5.1 ソーシャルコンテンツの情報検索への活用

オンラインコミュニケーションサイトや消費者生成メディアといったソーシャルコンテンツを情報検索に活用するソーシャルサーチの研究について述べる。

Twitter や SNS をはじめとする、オンラインコミュニケーションサービスから利用可能な知識を抽出する研究は多数行われており、とりわけ Twitter を対象とする研究は Sakaki ら¹⁵⁾、Pak ら¹³⁾ などさかんに行われている。こういった研究の中で、Twitter から抽出した知識をもとに情報検索を改善する研究の一例として、Dong ら¹⁾の研究がある。この研究では、Twitter における発言に含まれる URL からウェブページの鮮度を判定し、ウェブページのランキングに活用している。

ほかにオンラインコミュニケーションと一体化した情報検索システムの研究として、McNally ら¹¹⁾の研究があげられる。彼らはウェブコミュニケーションと一体化したソーシャルサーチシステム HeyStack^{*1}を開発、公開している。

読み手の意見を直接的にウェブページに付与する手法として、ソーシャルブックマークやタギングに代表されるソーシャルアノテーションがあげられる。ソーシャルアノテーションは多数のユーザがページに付けたタグを利用して、従来の Taxonomy に基づくディレクトリ型検索エンジンに対し、Folksonomy による階層的なウェブページの分類を行う。ソーシャルブックマークの情報を活用することでロボット型検索エンジンの精度を向上させる試みは、Heyman ら¹⁴⁾、Yanbe ら¹⁹⁾によって研究がなされている。本研究におけるリアクションも、広義ではソーシャルアノテーションの一種であるが、タグとリアクションではその性質が大きく異なる。明示的に人手で付与されるタグと異なり、ウェブコミュニケーションから抽出したリアクションはより瑣末な内容が多く、また多彩な表現をともなう。またタグが単語からなるのに対し、リアクションは自然文であるため、ユーザの生の声が反映されやすい反面、検索に利用するには2.3節で述べたような処理が必要となる。

5.2 評判情報検索

本研究においては、ウェブページを読んだ際にどう感じるかというユーザの記述をコミュニケーションサイトに含まれるリアクションからマイニングした。これと同様に、商品を購入

*1 <http://www.heystaks.com/>

入, 利用した際にどう感じるかというユーザの記述をブログや個人のウェブページなどから抽出する研究として, 評判情報検索がある⁷⁾. 評判情報検索は, その目的によって2種類に分類される.

- ある評価対象に対する評判を検索するもの
- 評判から評価対象を検索するもの

前者のような研究として, 藤村ら⁴⁾, 竹内ら¹⁷⁾の研究がある. これらの研究では, ブログなどに含まれる評価情報を収集し, ある商品や個人などを評価対象として, どのような評価がなされているかを抽出することを目的とする. 入力の商品名や個人名であり, 出力は入力されたオブジェクトがどのように評価されているかを表す評価極性である. 後者のような, 評価表現をクエリとする商品検索の研究として杉木ら¹⁶⁾の研究などがある. これらの研究では, ブログなどに含まれる評価情報を利用して, ある評価を受けているような製品や個人を発見することを目的とする. 入力は, 例としてテレビを探す際には“画質が良い”や“画面サイズが大きい”など, 属性名と評価値のセットであり, 出力は商品など評価対象となるオブジェクトである.

本研究は後者に近いが, ある製品やオブジェクトの発見を目的とするこれらの研究と違い, 検索対象はウェブ上にあるすべてのウェブページである. オブジェクトと異なりウェブページは属性が無数に存在するため, あらかじめ属性を設定しておくことができない. 本研究で提案する手法では, 事前に評価軸を設けず, どのようなクエリが入力された場合でもその場でスコアリングを行い文書をランキングしており, これらの研究とは異なる.

5.3 本文中に含まれない語による検索

検索対象となる本文中に直接的に含まれない語によって文書を検索する研究として, クエリ拡張, 潜在意味インデックス, センチメント分析についてそれぞれ述べる.

クエリ拡張とは, 入力されたクエリに対して関係の深い別の語をクエリとして利用することで, ページ発見の精度を上げたり, 元のクエリが直接含まれていないページを発見可能にしたりする手法である. ウェブ検索エンジンでよく見かけられるクエリ拡張として他ユーザのクエリログに基づくクエリ拡張および元のクエリによる検索結果を利用したクエリ拡張があげられる. 後者の手法は擬似適合性フィードバックと類似しており, 元の検索クエリと適合度の高いウェブページに含まれる語を関連語と見なし, 拡張クエリとして利用する. 近年では Wikipedia³⁾ や CQA サイトなど外部リソースを解析することで自動的にオントロジやシソーラスを構築しクエリ拡張を行う研究もさかんである.

潜在意味インデックスを用いた検索結果の拡張は, 潜在的意味解析²⁾に基づいており, 本

文中の語をもとに文書を分類することで, 同じ意味を持つ語であれば異なる表現でも文書を検索可能にする. しかしながら, ここでいう意味とは文書の持つ意味であり, コンテンツ閲覧者が受ける印象とは異なる.

文書を読んだ際に読み手が受ける印象と類似した概念として, 文書の持つ感情情報(センチメント)があげられ, Naら¹²⁾などにより情報検索利用が研究されている. また, 文書中の語を検索に利用できない絵画や音楽といったマルチメディアコンテンツの検索クエリとして感性語を利用する研究も栗田ら⁹⁾, 池添ら⁶⁾によって行われている. 商品レビューにおけるセンチメントを検索に利用する研究として Tangら⁵⁾の研究がある. これらの手法では, ウェブページやニュース記事を内容から分析し, “明るい 暗い”, “承認 拒否”などのいくつかの定められた軸ごとに感情の度合いを算出する⁸⁾. ある文書の持つセンチメントの算出には文中の要素を利用しているため, 現れるセンチメントは読み手ではなく書き手によって大きく左右される. この点が本稿における, 読み手がどう感じるかに基づいたウェブ検索と大きく異なる.

6. ま と め

本稿では, 文書を読んだ際にどのように感じるかに基づいた読み手目線の検索を実現する, 新しい検索モデルについて提案を行った. 提案手法ではウェブコミュニケーションデータに含まれるウェブページに対する読み手のリアクションを利用し, クエリとして印象語を入力可能にすることでこのような検索を可能にした.

提案手法を実装したウェブアプリケーションによる実験により, クエリ尤度モデルに基づく既存の検索手法と比較して提案手法が有用であると明らかにした.

今後は検索ランキングアルゴリズムのさらなる洗練化, ソーシャルファクターに基づく精度向上とランキングのパーソナライゼーション, オンライン掲示板や SNS へのリアクション収集対象の拡大など, 改良を行う.

謝辞 本研究の一部は, グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」, 文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」, 計画研究「情報爆発に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者: 田中克己, 課題番号 1809041) によるものです. ここに記して感謝の意を表します.

参 考 文 献

- 1) Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z. and Zha, H.: Time is of the essence: Improving recency ranking using twitter data, *Proc. 19th International Conference on World Wide Web*, pp.331–340 (2010).
- 2) Thomas H.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol.42, No.1-2, pp.177–196 (2001).
- 3) 堀憲太郎, 大石哲也, 峯 恒憲, 長谷川隆三, 藤田 博, 越村三幸: Wikipedia からの拡張クエリ生成による Web 検索とその評価, 第 20 回セマンティックウェブとオントロジー研究会, SIG-SWO-A803-13 (2009).
- 4) 藤村 滋, 豊田正史, 喜連川優: 電子掲示板からの評価表現および評判情報の抽出, 第 18 回人工知能学会全国大会, 3F1-03 (2004).
- 5) Tang, H., Tan, S. and Cheng, X.: A survey on sentiment detection of reviews, *Expert Systems with Applications*, Vol.36, No.7, pp.10760–10773 (2009).
- 6) 池添 剛, 梶川嘉延, 野村康雄: 音楽感性空間を用いた感性語による音楽データベース検索システム, *情報処理学会論文誌*, Vol.42, No.12, pp.3201–3212 (2001).
- 7) 乾 孝司, 奥村 学: テキストを対象とした評価情報の分析に関する研究動向, *自然言語処理学会論文誌*, Vol.13, No.3, pp.201–241 (2006).
- 8) Kumamoto, K. and Tanaka, K.: Proposal of Impression Mining from News Articles, *Lecture Notes in Computer Science*, Vol.3681, pp.901–910 (2005).
- 9) 栗田多喜夫, 加藤俊一, 福田郁美, 坂倉あゆみ: 印象語による絵画データベースの検索, *情報処理学会論文誌*, Vol.33, No.11, pp.1373–1383 (1992).
- 10) Manning, C.D., Raghavan, P. and Schütze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).
- 11) McNally, K., O’Mahony, M.P., Smyth, B., Coyle, M. and Briggs, P.: Towards a reputation-based model of social web search, *Proc. 14th International Conference on Intelligent User Interfaces*, pp.179–188 (2010).
- 12) Na, J.C., Khoo, C.S.G., Chan, S. and Hamzah, N.B.: Sentiment-based search in digital libraries, *Proc. 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005)*, pp.143–144 (2007).
- 13) Pak, A. and Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining, *Proc. 7th Conference on International Language Resources and Evaluation (LREC 2010)*, pp.1320–1326 (2010).
- 14) Heymann, P., Koutrika, G. and Garcia-Molina, H.: Can social bookmarking improve web search?, *Proc. International Conference on Web Search and Web Data Mining*, pp.195–206 (2008).
- 15) Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proc. 18th International World Wide Web Conference (WWW2010)*, pp.851–860 (2010).
- 16) 杉木健二, 松原茂樹: 消費者の意見に基づく商品検索, *情報処理学会論文誌*, Vol.49, No.7, pp.2598–2603 (2008).
- 17) 竹内啓祐, 浦島 智, 畑田 稔, 安宅彰隆: 電子掲示板からの評判情報抽出における P/N 判断, *電子情報通信学会技術研究報告, 知能ソフトウェア工学*, Vol.106, No.473, pp.55–60 (2007).
- 18) Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern information retrieval*, ACM Press, New York (1999).
- 19) Yanbe, Y., Jatowt, A., Nakamura, S. and Tanaka, K.: Can Social Bookmarking Enhance Search in the Web?, *Proc. 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2007)*, pp.107–116 (2007).

(平成 23 年 4 月 11 日受付)

(平成 23 年 9 月 12 日採録)



莊司 慶行

昭和 60 年生。京都大学大学院情報学研究所博士後期課程在学中。平成 20 年青山学院大学理工学部情報テクノロジー学科卒業。平成 22 年青山学院大学大学院社会情報学研究所社会情報学専攻修士課程修了。ウェブ検索, ソーシャルサーチ, ウェブコミュニケーションからの知識抽出の研究に従事。日本データベース学会学生会員。



田中 克己 (正会員)

京都大学大学院情報学研究所社会情報学専攻教授。昭和 51 年京都大学大学院修士課程修了。博士 (工学)。主にデータベース, マルチメディアコンテンツ処理, ウェブ検索の研究に従事。IEEE Computer Society, ACM, 人工知能学会, 日本ソフトウェア科学会, 日本データベース学会各会員。