情報処理学会論文誌　データベース　Vol.6　No.4　35–45 (Sep. 2013)

ソーシャルストリーム閲覧時の振舞いを利用したユーザプロファイル構成手法

土岐 真里奈[1]　牛尼 剛聡[2,a)]

受付日 2013年3月20日, 採録日 2013年7月8日

概要：本論文では，Twitter に代表されるソーシャルストリームに対するユーザの閲覧時の振舞いから，ユーザのプロファイルを構成する手法を提案し，被験者実験により有効性を評価する．本手法では，ソーシャルストリームの閲覧時におけるユーザのスクロール操作を利用し，各ツイートを読む時間（滞留時間）を推定する．そして，推定した滞留時間に基づいてユーザプロファイルを構成する．本論文で提案するユーザプロファイルは，単語に対する興味を表す「興味単語プロファイル」と，コンテンツの発信者に対する興味を表す「興味ユーザプロファイル」から構成される．興味単語プロファイルを構成するために，TF-IDF 法を滞留時間によって拡張した TF-IDF-RT 法を提案する．また，興味ユーザプロファイルを構成するために，滞留時間に基づいてユーザが興味を持つ投稿者を推定する手法を提案する．提案手法で構成したユーザプロファイルを利用して推薦されるツイートに対して，ユーザが興味の度合いを評価するタスクに関する被験者実験の結果に基づいて，提案手法の有効性を示す．

キーワード：SNS，Twitter，ソーシャルストリーム，ユーザプロファイル，振舞い

A Method for Composing a User Profile Based on Browsing Behaviors on Social Streams

Marina Toki[1]　Taketoshi Ushiama[2,a)]

Received: March 20, 2013, Accepted: July 8, 2013

Abstract: In this paper, we introduce a method for composing a user profile of a user based on browsing behaviors of the user on social streams such as Twitter, and evaluate the effectiveness of our method based on subjective experiment results. The proposal method estimates the time of reading each tweet (retention time) in a timeline according to scrolling operations of the user on the timeline, then compose a profile of the user based on the estimated retention times. The user profile that is proposed in this paper consists of interest word profile and interest user profile; an interest word profile represents which subjects the user is interested in, on the other hand, an interest user profile represents which users the user is interested in. In order to compose an interest word profile, we introduce the IF-IDF-RT method, which is an expansion of the TF-IDF method with the retention time. On the other hand, in order to compose an interest user profile, we introduced an technique for estimating other users who interests the system based on the retention time. We evaluate the effectiveness of our method based on subjective experiment results on the tasks that extract tweets that interest the subjects from timelines by means of the user profiles that our method composed.

Keywords: SNS, Twitter, social stream, user profile, behavior

1. はじめに

Twitter や Facebook に代表される SNS（ソーシャルネットワーキングサービス）が爆発的に普及し，友人・知人とのコミュニケーション，情報収集等，様々な目的のために日

[1] 九州大学芸術工学部
School of Design, Kyushu University, Fukuoka 815–8540, Japan
[2] 九州大学大学院芸術工学研究院
Faculty of Design, Kyushu University, Fukuoka 815–8540, Japan
[a)] ushiama@design.kyushu-u.ac.jp

常的に利用されるようになった [1]．SNS に投稿されたコンテンツ（ソーシャルコンテンツ）は，投稿したユーザをフォローしているユーザに自動的に配送される．一般的に，SNS ユーザは複数のユーザをフォローするため，1 人のユーザには，断続的に様々なユーザから多様なコンテンツが配信される．配信されたコンテンツは，1 列に並べた形式でユーザに提示され，一種のストリームデータと考えることができるため，「ソーシャルストリーム」と呼ばれる [2]．

従来，Web 上で情報を取得する代表的な手法として Web ページ検索が一般的に利用されてきた．Web ページ検索では，ユーザは情報要求をクエリとして表現し，情報要求に合致する Web ページを検索する．Web ページ検索は，ユーザが必要とする情報があらかじめ明確に分かっている場合には有効である．一方，ソーシャルストリームを閲覧するユーザは，あらかじめ必要な情報が分かっているわけではない．ソーシャルストリームを閲覧するユーザは，自分が興味を持つ情報を提供することが期待できるユーザをあらかじめフォローしておき，配信されてくるコンテンツ列の中から興味を持つコンテンツを選別する．

一般に，ソーシャルストリーム上には多種多様なコンテンツが含まれているため，その中には，ユーザにとって価値の高いコンテンツもあれば，価値の低いコンテンツも存在する．また，ソーシャルストリームは大量のコンテンツから構成され，ユーザはコンテンツの 1 つ 1 つの内容を正確に確認することが困難である．そこで多くのユーザは，忙しい朝や隙間時間等，時間的な余裕がないときには，ソーシャルストリームの中から目についたものだけを拾い読みすることがある．しかし，内容を正確に確認せず読み飛ばされたコンテンツの中に，ユーザにとって価値の高い情報が含まれている可能性がある．このようなコンテンツは，ユーザに時間的な余裕がないために有効に活用されなかったが，ユーザに時間的な余裕があれば，有効に活用できた可能性がある．我々は，このように，ユーザにとって価値が高いにもかかわらず，ユーザが閲覧する際のコンテクストが適切でなかったために，有効に活用されなかったコンテンツを，「見落としコンテンツ」(slipped contents) と呼ぶ．

我々は，ソーシャルストリームにおける見落としコンテンツの存在が，ソーシャルストリームにおける情報取得の効果を減少させていると考えている．もし，見落としコンテンツが取得できれば，それらを時間的に余裕があるときに再提示することによって，ソーシャルストリームにおける情報取得の効果を向上させることができると考えられる．

こうした背景の下，我々は，ソーシャルストリームからの見落としコンテンツの自動抽出システムを開発中である．本システムにおける処理の流れを**図 1** に示す．本システムでは，ソーシャルストリームの閲覧インタフェースに対

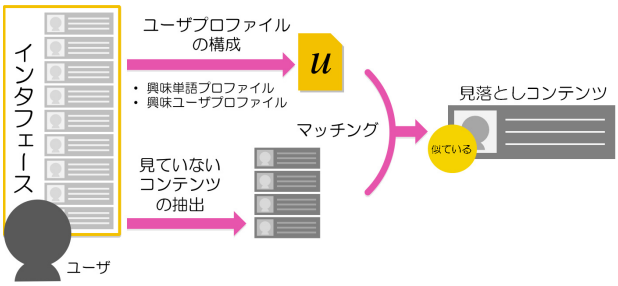

図 1　見落としコンテンツ抽出の処理の流れ
Fig. 1　A process flow of extraction of slipped contents.

する振舞いに基づいて，ユーザのプロファイルを自動的に抽出する．また，ソーシャルストリームの中で，ユーザが読み飛ばし等によって見ていないコンテンツ中から，ユーザプロファイルに基づいて，ユーザにとって価値が高い見落としコンテンツを自動的に抽出する．本論文では，上記のシステムにおいて，ソーシャルストリーム上の振舞いに基づいてユーザプロファイルを自動的に構成する手法を提案する．なお，対象として，代表的な SNS の 1 つである Twitter を想定する．

これまでにも，ユーザの振舞いに基づいて，ユーザの興味を取得する研究が行われてきた．従来の代表的な手法として，ユーザの視線を抽出する手法がある [3]．この手法では，特殊な視線測定装置を利用して，表示領域上でユーザがどこに注目しているかを検出し，注目領域に基づいてユーザの興味箇所や単語を抽出する．この手法では，ユーザの文書中の注目箇所を正確に把握できる．しかし，視線測定には特殊な装置が必要であり汎用性が低い．また，クリックしたコンテンツやマウスの動き等からユーザが興味を持った箇所を推定する手法も提案されている [4]．しかし，これは一般的な Web ページを対象とした手法であり，ソーシャルストリーム閲覧では，一般的に，複数のリンクの中から興味がある内容に関するリンクのみをクリックをするわけではないため，そのまま適用することは困難である．

上記の問題点を解決するために，本論文では，ソーシャルストリームに対するユーザのスクロール操作を利用してユーザプロファイルを構成する手法を提案する．ソーシャルストリームを閲覧する際，ユーザはつねに一定の速度でストリームをスクロールするわけではない．ユーザは，興味を引かれないツイートに対しては読み飛ばしを行い，興味を引かれるツイートは時間をかけて読み，ときにはお気に入り追加や返信等の操作を行うと考えられる．ユーザが個々のツイートの閲覧に要した時間（滞留時間）は，ユーザの興味を反映していると考えられる．そこで，個々のツイートに対する滞留時間を推定することができれば，それらを利用してユーザの興味を推定できる可能性がある．そこで，本手法では，ソーシャルストリームに対するユーザ

の振舞いデータを利用して，それぞれのツイートの滞留時間を推定する．

本手法において，見落としコンテンツを抽出するためのユーザプロファイルは，ツイートに含まれる単語に対するユーザの興味を表す「興味単語プロファイル」と，ツイートの投稿者に対する興味を表す「興味ユーザプロファイル」から構成される．興味単語プロファイルを構成するために，従来の文書中の単語の重み付けに利用されてきた代表的な手法である TF-IDF 法をツイートの滞留時間によって拡張した TF-IDF-RT 法を提案する．また，興味ユーザプロファイルを構成するために，ツイートに対する滞留時間を利用して，ユーザの投稿者に対する興味を構成する手法を提案する．

本論文の構成は以下のとおりである．2 章で関連研究について述べる．3 章では，提案手法について述べる．4 章で提案手法の有効性を検証するための被験者実験の結果を示し有効性を評価する．5 章で提案手法の問題点と今後の発展について議論する．6 章でまとめと今後の課題を述べる．

2. 関連研究

膨大かつ多様なソーシャルストリームの中から，ユーザが効果的に情報取得を行うことを目的として，Facebook では，EdgeRank アルゴリズム [5] を用いて，ユーザと投稿者との「親密度」，他のユーザの「いいね」の数，その投稿の新しさという 3 つの指標に基づいて，ストリームにおけるコンテンツをランク付けし，コンテンツを表示する順番を決定している．しかし EdgeRank では，主にユーザの交友関係やコンテンツに対する他者の評価に重きをおいており，ユーザ自身の嗜好や興味を考慮していない．そのため，ユーザが興味のある投稿が読み飛ばされ，十分に活用されない可能性がある．それに対して提案手法では，ユーザの振舞いに基づいてユーザの嗜好や興味を抽出することにより，EdgeRank とは異なる観点から，ユーザにとって価値が高いコンテンツを抽出することを実現する．

コンテンツに対する振舞いを用いてユーザプロファイルを作成するアプローチは，情報検索や情報推薦の分野で暗黙的フィードバック [6] を実現するための手法としていくつか提案されている．

梅本ら [7], [8] は，ユーザの視線測定結果に基づいて，Web ページにおいて，ユーザが興味を持った箇所を推定する手法を提案している．これは，特殊な視線計測装置を利用して，ユーザが注視しているスクリーンの位置と時間を取得し，分析することにより，ユーザが閲覧中のページで興味がある事項を推定可能である．しかし，この手法では，高価な視線装置が必要であり，汎用性が高いとはいえない．それに対して，我々は，対象をソーシャルストリームに限定することにより，スクロール操作というユーザの自然な振舞いから，視線計測装置のような特殊な機器を必要とせずに，ユーザの興味を反映したプロファイルを抽出可能である．

松尾ら [9] は，ユーザの Web ページアクセスという振舞いから興味を把握し，個人化した情報提示を行う手法を提案している．具体的には，ユーザが閲覧した Web ページの履歴から，ユーザにとって重要度の高い単語を推定し，Web ページを閲覧する際に，表示中の Web ページにおいて重要度の高い単語をハイライトすることで，ユーザの効果的な情報取得を支援する．この手法では，ユーザのアクセス履歴に含まれる Web ページはユーザの興味を反映していると考え，一般的な単語の共起とアクセス履歴に含まれる Web ページにおける単語の共起の偏りに注目して，ユーザが興味を持つ単語を推定する．Web ページのアクセス履歴は，ユーザが能動的に Web ページを選択した結果であり，ユーザの興味が強く反映されていると考えられる．一方で，本研究で対象とするソーシャルストリームにおいては，多種多様なコンテンツが混在しているため，ユーザに配信されたソーシャルストリームを構成するすべてのコンテンツがユーザの興味を強く反映しているわけではない．そこで，本手法では，単にコンテンツが配信されたかだけではなく，それぞれのコンテンツに対する滞留時間を推定し，滞留時間を利用することで，ソーシャルストリームに対する振舞いのみからユーザの興味を効果的に推定可能である．

Morita ら [10] は，ネットニュースに対する閲覧時間を利用してユーザの興味を抽出する手法を提案している．この手法では，ネットニュースでは，ユーザはブラウジング中に単一の記事のみが画面上に現れるインタフェースを利用することが一般的であり，記事の表示時間が簡単に取得可能である．それに対して，本研究で対象とするソーシャルストリームでは，複数の異なる投稿が直列化され，画面上に表示されている時間であっても，ユーザは画面中で別の記事を読んでいる可能性があり，Morita らの手法ではソーシャルストリームを構成する個々の記事に対する滞留時間を求めることは困難である．

Buscher ら [11] では，対象文書中の 1 行ごとの表示時間を測定し，文書中の段落に含まれる行の表示時間の平均に基づいて，その段落に対する滞留時間を推定する手法を提案している．そして，滞留時間が一定時間以上の段落に対して TF-IDF 法を適用することにより，ユーザが興味を持つ単語の重み付けを行う．この手法では，文書の段落を単位として属性を推定しているが，ソーシャルストリームでは，1 つの記事を構成する行数が一般的な文書の段落よりも短く，1 つの画面内に多数の投稿が表示されることが多い．したがって，画面上に表示されている行の表示時間の平均値として個々の投稿に対する滞留時間を考えるのは不適切である．また，文書は前後の段落が意味的に関連しており，文書自体にも意味的な関連が強いことが多いが，ソーシャ

ルストリームは意味的な関連が弱い多種多様な投稿が混在しているために，適切なユーザプロファイルの推定のために，個々の投稿に対する滞留時間をより正確に推定する必要がある．また，ソーシャルストリームにおいては，投稿者に関する注目度が影響すると考えられるが，Buscher らの手法では投稿者に関する興味を考慮していない．

林ら [12] は，ソーシャルコンテンツを含む CGDC（Customer Generated Digital Contents）特有の特徴を考慮し，ドキュメントの発生順序を考慮した TF-IDF 法の提案を行っている．林らは，ニュース記事のように，時間の経過にともなって断続的に生成されるコンテンツには，通常の TF-IDF 法では最新の状況を的確に反映させることができないと考え，文書数の増加に応じて IDF 項の値を変化させることで，イベントの影響により変化した特徴語を TF-IDF 法に取り入れている．この手法では，同一の話題に関する情報が断続的に投稿されるようなコンテンツに対しては，系列中の「変化」に基づいて特徴的な単語を抽出可能である．しかし，本研究で対象とするソーシャルストリームにおいては，ニュースのように同一の話題の経過を表さない独立した話題のコンテンツも多く含まれるため，林らの手法が効果的でない場合が多い．それに対して，我々の手法では，ストリームを構成するコンテンツ間の順序関係は利用していないため，意味的に継続性を持たないコンテンツに対しても適用可能である．

3. 提案手法

3.1 アプローチ

本研究は，ソーシャルストリーム中の見落としコンテンツを抽出することを最終的な目標とする．本論文では，代表的なソーシャルストリームである Twitter のタイムラインを対象として，上記の目標のために効果的な，ユーザプロファイルを，ソーシャルストリームに関する振舞いのみから，特別な装置を必要としない環境で構成可能な手法を提案する．

提案手法を開発するにあたり，我々は，タイムラインを閲覧するユーザは，タイムラインを構成するすべてのツイートの内容をすべて正確に把握するわけではなく，興味を持つ可能性が高いと判断したコンテンツのみを選択的に読むことが多いと考え，選択的に読んだコンテンツがユーザにとって価値が高ければ，お気に入りへの追加，リツイート，返信，サイト閲覧等の操作を行うため，興味がひかれたコンテンツにはそのツイート読む時間（滞留時間）が長くなると考えた．

本手法では，上記の考えに基づいて，ユーザのスクロール操作から各ツイートの滞留時間を推定し，推定した滞留時間に基づいて，ツイートに含まれる単語や，ツイートを投稿したユーザの重みを計算することにより，ユーザの興味や嗜好を反映させたユーザプロファイルを構成する．

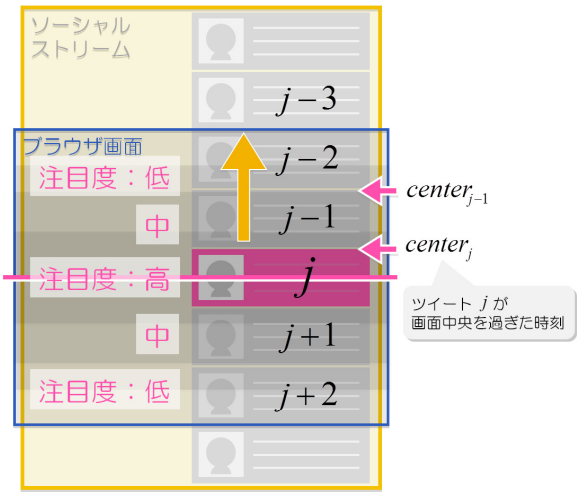

図 2　滞留時間の推定
Fig. 2　Estimation of the retention time.

3.2 滞留時間の測定

一般的に，タイムラインを表示するブラウザ画面には，同時に複数のツイートが表示される．ブラウザ画面上で，ユーザはつねに固定された点を注目しているわけではないため，個々のツイートに対する滞留時間を正確に測定するためには，ユーザがブラウザ上のどのツイートに注目しているかを知る必要がある．ブラウザ画面上のユーザの注目点は，視線計測装置を利用することにより正確に取得可能である．しかし，本研究では，視線計測装置等の特殊な装置を利用せずに，ツイートの滞留時間を推定することを目指す．いま，閲覧中のタイムラインを構成する j 番目のツイートの下端が，スクロールによってブラウザの表示領域の上下中央の線と重なった時刻を $center_j$ と表すことにする（**図 2**）．このとき，仮に，ユーザが中央線上に注目している場合のツイート j に対する滞留時間 $center\_rt_j$ は中央線を過ぎたツイート $j-1$ と次に過ぎたツイート j との時刻の差として，以下の式で計算できる．

$$center\_rt_j = center_j - center_{j-1} \tag{1}$$

ユーザは上下の中央線付近のみに注目しているわけではなく，画面の位置によって注目する可能性が異なると考えられる．そこで，ブラウザ画面上の表示位置によって注目度が異なると考える．そして $center\_rt_j$ を，ブラウザ画面上に表示されているツイートの注目度に応じて割り振る．最後に，ツイートごとに割り振られた $center\_rt_j$ を注目度によって重み付けした合計をツイート j の滞留時間 rt_j としてート $j+n$ の $2n+1$ 個のツイートが表示されているとする．このとき，ツイート j の滞留時間 rt_j を式 (2) を用いて推定する．

$$rt_j = \sum_{k=j-n}^{j+n} \left(center\_rt_k \times \cos\left(\frac{(k-j)\pi}{n} \right) \right) \tag{2}$$

ここで，cos 関数は注目度を表している．

なお，ツイートを閲覧する際には，ツイートの本文を読むだけではなく，ツイートに含まれるURLをクリックしてWebページを閲覧したり，リツイートや，お気に入り登録等の振舞いを行ったりすることがある．したがって，上記で推定される滞留時間には，ツイート本文の閲覧だけでなく，上記のような振舞いに要する時間が含まれる可能性がある．そこで，滞留時間には上限を設け，1つのツイートの影響が大きくなりすぎないようにする．我々は，経験的に，上限の値としては20秒程度が適切であると考えている．

3.3　ユーザプロファイルの構成とコンテンツ価値推定

本手法では，ユーザが閲覧したタイムラインと推定された滞留時間に基づいて，ユーザの興味を表すユーザプロファイルを構成する．本手法では，ユーザの興味は，ツイートの内容自体に対する興味と，投稿者に関する興味の2種類に分類できると考え，前者を興味単語プロファイルで表し，後者を興味ユーザプロファイルで表現する．

3.3.1　興味単語プロファイルの構成

文書中の単語の重要度を推定するために多くの手法が提案されている[13]．代表的な推定手法としてTF-IDF法がある．TF-IDF法では，ある文書では出現頻度が高いが，他の文書に出現する頻度が低い単語は，その文書の特徴を強く表していると考え，文書中に出現する単語の重要度を，対象とする文書 j における単語 i の出現頻度 $tf_{i,j}$ と，単語 i を含む文書の出現頻度 df_i の逆数の積として求める．TF-IDF法による単語の重み付けは以下の式(3)で形式的に定義される．ある文書 j における単語 i が，その文書中で出現頻度が高く，かつ文書集合中で出現頻度の低いものならば，文書 j における単語 i の重み $tfidf_{i,j}$ の値が高くなる．なお，N は文書集合に含まれる文書数を表す．

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{3}$$

TF-IDF法による重み付けは文書検索等，様々な分野で広く利用されている．しかし，Twitterにおいて，個々のツイートを1つの文書としてとらえ，TF-IDF法をそのまま適用して単語の重要度を求めることは適切ではない．なぜなら，ツイートは140字以内であるため，単一のツイート中に同じ単語が複数回出現することは稀である．したがって，大多数のツイートにおいて，そこに含まれる単語数は1となり，idf値が小さい単語の重要度が多くなる．つまり，出現頻度が少ない単語の重要度が高くなり，利用者の興味を正確に反映することが困難になる．さらに，Twitterのタイムラインには多種多様な情報が含まれており，タイムラインを構成するすべてのツイートがユーザにとって価値が高い情報を提供しているわけではない．したがって，単語の重要度を決定する際に，タイムラインを構成するすべてのツイートが同一の重要度を有すると考えるのは不適切である．

上記の問題を解決するために，タイムラインを構成するツイートの中で，「お気に入り」に登録したり，リツイートしたツイートを，ユーザが興味を持つツイートと考え，それらのツイートに含まれる単語を「ユーザが興味のある単語」として単語の重み付けをしたりするアプローチが考えられる．しかし，すべてのユーザが興味のあるコンテンツすべてに対して「お気に入り」への登録やリツイートといった操作をするわけではない．また，タイムラインには「ユーザにとって価値が高い情報が含まれているが1度読めば十分である」ツイートも多く含まれている．そのため，お気に入りツイートにおける重みは，ユーザにより大きく異なる要素となる．

上記をふまえ，我々はユーザのツイートに対する滞留時間を利用して，単語の重み付けを行う手法を提案する．具体的には，TF-IDF法を，ツイートに対する滞留時間を利用して拡張したTF-IDF-RT法を提案する．単語 i の，ツイート j におけるTF-IDF-RT法に基づく重み $\mathrm{tfidfrt}_{i,j}$ を式(4)として定義する．

$$tfidfrt_{i,j} = tfidf_{i,j} \times rt_j \tag{4}$$

この式では，ツイート j 中の単語 i に対しTF-IDF法で抽出された重みに，ツイート j を読むのに要したと推定される滞留時間 rt_j を掛け合わせている．これにより，「お気に入り」への登録やリツイートといった明示的な操作がなくても，個々のツイートに対するユーザの興味を反映させて，単語の重要度を推定できる．

いま，注目するユーザのタイムラインを構成するツイート集合を TL とすると，ユーザの興味単語プロファイル $\mathbf{p}_{\mathrm{word}}$ は，式(5)に示すように，それぞれの単語 i について，TL に含まれるすべてのツイート j の重み $tfidfrt_{i,j}$ の和を，ベクトルとしたものとして定義する．なお，数字・漢数字のみやひらがな1文字等の興味単語として意味を持たない語，記号，「www」や「RT」等のTwitterで頻繁に利用されるが，それ自体に意味のない語はストップワードとして，興味単語プロファイルの構成要素から除外する．

$$\mathbf{p}_{\mathrm{word}} = \sum_{j \in TL} \begin{pmatrix} tfidfrt_{1,j} \\ tfidfrt_{2,j} \\ \vdots \\ tfidfrt_{n,j} \end{pmatrix} \tag{5}$$

3.3.2　興味ユーザプロファイルの構成

ユーザが興味を持つツイートの重要な要素として，「どのユーザが発信したツイートか」という，投稿者の情報がある．多くのTwitterクライアントのインタフェースでは，ツイートの左側に投稿者のアイコンが表示されるデザインを採用しているため，ユーザは投稿者のアイコンでツイートを読むかどうかを判別することがある．そのため，ユー

ザにとっての投稿者の注目度を見落としコンテンツ抽出のためのユーザプロファイルに利用する．

本手法では，ユーザにとって投稿者 k の注目度を，ユーザのタイムラインに含まれる k が投稿したツイートに対する滞留時間をツイート長で除したものの総和として考える．形式的には，ユーザ k の注目度 at_k は式 (6) として定義される．

$$at_k = \sum_{j \in \mathrm{TW}(TL,k)} \frac{rt_j}{\mathrm{len}(j)} \tag{6}$$

ここで，$\mathrm{TW}(TL,k)$ は，ユーザのタイムライン TL に含まれる投稿者 k が投稿したツイートの集合を表し，rt_j はツイート j の滞留時間を表し，$\mathrm{len}(j)$ はツイート j の文字数を表す．一般的に，ツイートによって文字数は異なるが，ユーザがツイートを読み飛ばさなかった場合，文字数が多いツイートの方が，文字数が少ないツイートよりも滞留時間が長くなる傾向がある．しかし，文字数が少なくてもユーザが興味を持つツイートは存在すると考えられるため，滞留時間をツイートの長さで除することにより，滞留時間を正規化し，ユーザの興味を正しく反映できるようにしている．

興味ユーザプロファイル $\mathbf{p}_{\mathrm{user}}$ は，式 (6) で求めた at_k をすべてのユーザに関してベクトルとして表現したものとして式 (7) のように定義する．

$$\mathbf{p}_{\mathrm{user}} = \begin{pmatrix} at_1 \\ at_2 \\ \vdots \\ at_n \end{pmatrix} \tag{7}$$

3.4　コンテンツ価値判定

本手法では，タイムライン中でユーザが読み飛ばしたと考えられるツイート集合に含まれる個々のツイート対して，興味単語プロファイル $\mathbf{p}_{\mathrm{word}}$ と，興味ユーザプロファイル $\mathbf{p}_{\mathrm{user}}$ を利用して，そのツイートの重要度（価値）を推定し，重要度が高いと考えられるツイートが，見落としコンテンツである可能性が高いと考える．

今，興味単語プロファイル $\mathbf{p}_{\mathrm{word}}$ と，興味ユーザプロファイル $\mathbf{p}_{\mathrm{user}}$ が与えられたとき，ツイート j に対するコンテンツ価値 $\mathrm{value}(j, \mathbf{p}_{\mathrm{word}}, \mathbf{p}_{\mathrm{user}})$ を以下の式 (8) のように定義する．

$$\begin{aligned} &\mathrm{value}(j, \mathbf{p}_{\mathrm{word}}, \mathbf{p}_{\mathrm{user}}) \\ &\quad = \mathrm{sim}_{\mathrm{word}}(j, \mathbf{p}_{\mathrm{word}}) + \alpha \mathrm{sim}_{\mathrm{user}}(j, \mathbf{p}_{\mathrm{user}}) \end{aligned} \tag{8}$$

ここで，興味単語プロファイルに基づいた類似度 $\mathrm{sim}_{\mathrm{word}}$ は式 (9) として定義され，興味ユーザプロファイルに基づいた類似度 $\mathrm{sim}_{\mathrm{user}}$ は式 (10) として定義される．また，α はパラメータである．

$$\mathrm{sim}_{\mathrm{word}} = \frac{\mathbf{v}_{\mathrm{word}_j} \cdot \mathbf{u}_{\mathrm{word}}}{|\mathbf{v}_j| \times |\mathbf{u}_{\mathrm{word}}|} \tag{9}$$

$$\mathrm{sim}_{\mathrm{user}} = at_{\mathrm{user}(j)} \tag{10}$$

ここで，$\mathbf{v}_{\mathrm{word}_j}$ は，式 (11) のように表現されるツイート j の特徴ベクトルである．

$$\mathbf{v}_{\mathrm{word}_j} = \begin{pmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{n,j} \end{pmatrix}$$

$$w_{i,j} = \begin{cases} 1 & (i \text{ が } j \text{ に含まれるとき}) \\ 0 & (\text{それ以外のとき}) \end{cases} \tag{11}$$

4.　実験と評価

4.1　プロトタイプシステム

本手法の有効性を検証するため，プロトタイプシステム「Context Catcher」を実装した．本システムは，Twitter クライアントの一種であり，HTML，PHP，JavaScript を用いて開発され，Apache 上の Web サービスとして JavaScript が動作する Web ブラウザ上で動作する．なお，サーバ側でのデータ管理のために MySQL を利用している．

図 3 に Context Catcher の動作画面のスナップショットを示す．本システムでは，ユーザが Twitter のユーザ名とパスワードを入力することにより，ユーザ自身のタイムラインを閲覧できる．本研究では，ユーザがプロファイル構成のための能動的な動作および視線測定装置等の特殊な設備を使用せず，通常の Twitter 閲覧操作のみから，見落としコンテンツを抽出することを目標としている．そこで，インタフェースのデザインは，公式の Twitter クライアントのデザインを踏襲した．システムはユーザの閲覧操作に基づいて様々なユーザの振舞いデータを取得し蓄積可能である．本手法における滞留時間の推定に主に利用したスク

図 3　プロトタイプシステムのスナップショット
Fig. 3　A snapshot of our prototype system.

ロール操作のほか，「外部リンクのアクセス」，「リツイート」，「お気に入り登録」等も取得可能である．

4.2　実験方法

提案手法の有効性を評価するために，プロトタイプを用い被験者実験を行った．被験者は日常的に Twitter を利用している 19 歳から 25 歳の 9 名である．

被験者には，各自のパソコンで本プロトタイプシステムを用いて各自の Twitter のタイムラインを閲覧してもらった．なお，ユーザプロファイルを効果的に生成するためには，興味があるツイートはすべて確認できるだけの時間的な余裕があることが重要であると考えられる．そこで，本プロトタイプシステムの利用は，時間的に余裕があるときのみに限定し，興味があるツイートすべてを確認する時間的な余裕がないときには，普段利用している Twitter クライアントを用いるよう指示した．

被験者に 5 日間本プロトタイプシステムを利用してもらった後に，得られたデータを用い各被験者ごとに，興味単語プロファイルと興味ユーザプロファイルを作成した．

プロトタイプシステムを使用した最終日から 7 日経過した後に，被験者それぞれの最新のタイムラインから 700〜800 件のツイート（対象ツイート）を取得した．そして，提案手法と既存手法を含めた合計 5 つの手法を利用して，対象ツイートすべてのコンテンツ価値を計算し，コンテンツ価値が高いと判断された 10 ツイートを抽出した．そして被験者には，どの手法によって推薦されたか分からないようランダムに提示して，推薦されたツイートを 5 段階によって評価してもらった．今回対象とした 5 つの手法は以下のとおりである．

- 興味単語プロファイルと興味ユーザプロファイルを併用した手法（提案①）
- 興味単語プロファイルのみを用いた手法（提案②）
- 興味ユーザプロファイルのみを用いた手法（提案③）
- TF-IDF 法を用いた手法（既存④）
- TF 法を用いた手法（既存⑤）

本実験では，各手法において上位 10 件，合計 50 件を推薦ツイートとしてユーザに提示した．被験者には，推薦ツイートの中から同一ツイートを除いてランダムな順番で提示し，「面白い」と感じるかどうかを，5 段階で評価してもらった．被験者に提示した評価用アンケートの例を**図 4** に示す．面白さの基準は「リプライ，お気に入り，リツイートしたくなるようなもの」「ネタ的な面白さ」「勉強のために意識的に読もうと思うもの」を含めた「被験者が普段じっくり読むツイート」とし，リンクアドレスを含むツイートに関しては「普段そのツイートを見たときにどう感じるか（興味を持つか，リンク先を開きたくなるか）」の度合いとした．

なお，実験にあたり，式 (8) における α の値は，経験的

図 4　実験に用いた評価用アンケートの例
Fig. 4　Example of questionnaires for evaluation.

表 1　実験により取得された代表的な振舞いデータ
Table 1　Representative behavioral data obtained by user studies.

	フォロー数（人）	実行回数（セッション数）	1 セッションで読むツイート数	ツイートごとの平均滞留時間
平均	218	99.3	29.20	0.28

に 0.45 とした．また，3.2 節で述べたように，提案手法においては，1 つの滞留時間の長いツイートの影響が強くなりすぎないように，滞留時間の上限を 20 秒に設定した．

4.3　実験結果

被験者 9 名のうち，2 名はシステムの利用回数が少なかったため，データ不足のため検証不可能とし除外した．以下，実験結果の検証は 2 名を除く 7 名のデータをもとに進める．5 日間の実験から得られた代表的な振舞いデータについて，全被験者の平均値を**表 1** に示す．

4.3.1　抽出されたツイートに対する被験者の評価

各手法で，ユーザにとって価値の高いツイートを抽出できたかを比較するため，各手法を用いて抽出したツイートに対して，被験者が回答した評価値の平均値を棒グラフとした図を**図 5** に示す．平均値では提案手法①と提案手法②と提案手法③が，既存手法④，既存手法⑤よりも，高い評価を得た．それぞれの平均値に関して，有意な差があるかを t 検定を用いて検証した．3 種類の提案手法と 2 種類の既存手法との間で t 検定を行った際の p 値を**表 2** に示す．提案手法①と②の評価が最も値が高くなっている．また，提案手法①と既存手法④，提案手法①と既存手法⑤において有意水準 1%で有意差が見られた．また，提案手法②と既存手法④，提案手法②と既存手法⑤において，有意

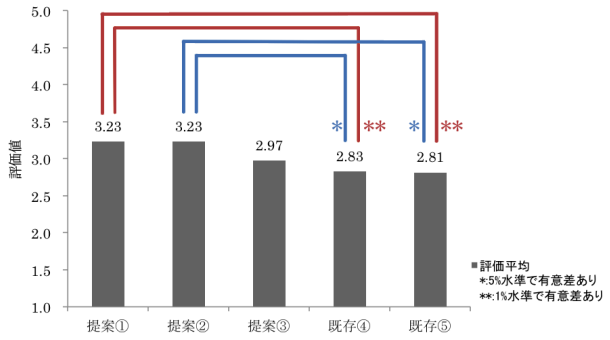

図 5　抽出されたツイートに対して被験者が回答した評価値の平均値
Fig. 5　Average scores of evaluation by subjects for extracted tweets.

表 2　評価値の平均値に対する t 検定における p 値
Table 2　P-values of t-test on average scores of evaluation.

	既存④TF-IDF 法	既存⑤TF 法
提案① 単語+ユーザ	0.009 **	0.009 **
提案② 単語	0.010 *	0.011 *
提案③ ユーザ	0.237	0.205

水準 5%で有意差が見られた．

一方，各手法においてユーザに対するランキングが適切であるかを評価するために，ランキングの評価に一般的に用いられている DCG（Discounted Cumulative Gain）および nDCG（normalized Discounted Cumulative Gain）の値を求めた．DCG はランキングに対する評価を行う指標であり，ユーザによる評価が高いアイテムが上位に順位付けされるほど評価が高いという考え方に基づいている．具体的には，ランキング p 位までの結果に対する DCG は式 (12) で与えられる [14]．

$$\mathrm{DCG}_p = R_1 + \sum_{i=2}^{p} \frac{R_i}{\log_2 i} \tag{12}$$

ここで，R_i は，ランキング i 位のアイテムに対する関連度を表し，今回の実験では被験者が与えた評価値を利用した．各種法における DCG の平均値を棒グラフとして表した図を**図 6** に示す．また，3 種類の提案手法と 2 種類の既存手法との間で t 検定を行った際の p 値を**表 3** に示す．DCG では，値が高いほどユーザにとって評価の高い項目が上位にランキングされることを表している．実験結果により，提案手法①の値が最も評価が高く，既存手法④と既存手法⑤に対して有意水準 5%で有意差が認められた．一方，提案手法②は，提案手法①の次に評価が高く，既存手法④との間では有意水準 5%で有意差が認められ，既存手法⑤との間では有意水準 1%で有意差が認められた．

nDCG は，正規化された DCG であり，各種法で選択されたツイート集合内での順位付けの正しさを表す．具体的には，p 位までの nDCG は式 (13) によって計算される．

$$\mathrm{nDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p} \tag{13}$$

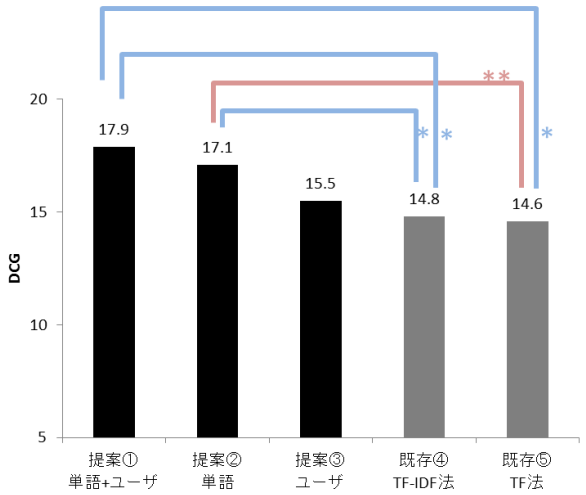

図 6　抽出したツイートに関する DCG の値
Fig. 6　DCG values for extracted tweets.

表 3　手法ごとの DCG に対する t 検定の p 値
Table 3　P-values of t-test on DCG of each method.

	既存④TF-IDF 法	既存⑤TF 法
提案① 単語+ユーザ	0.024 *	0.019 *
提案② 単語	0.046 *	0.008 **
提案③ ユーザ	0.334	0.284

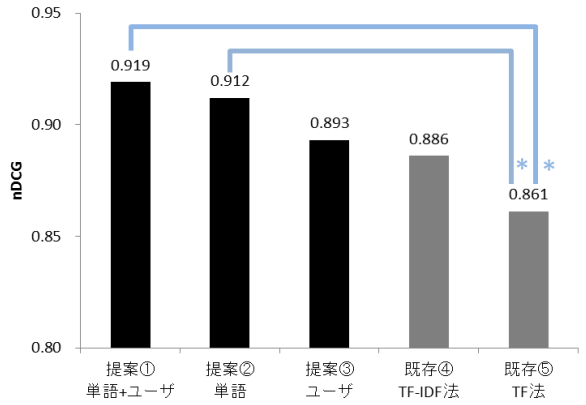

図 7　抽出したツイートに関する nDCG の値
Fig. 7　nDCG values for extracted tweets.

ここで，IDCG_p は，与えらえられた関連性に基づいた理想的な DCG を表している．各種法における nDCG_{10} の平均値を棒グラフとして表したものを**図 7** に示す．また，3 種類の提案手法と 2 種類の既存手法との間で t 検定を行った際の p 値を**表 4** に示す．nDCG では，それぞれの手法で選択された項目内での順位の妥当性を表しており，値が大きいほど有効であると考えられる．実験結果により，提案手法①が最も評価が高く，既存手法⑤に対して有意水準 5%で有意差が認められた．また，提案手法②が 2 番目に評価が高く，既存手法④に対して有意水準 5%で有意差が認められた．

4.3.2　ツイートのとばし読みに関する結果

本研究では，前提として「ソーシャルストリーム上には

表 4　手法ごとの nDCG に対する t 検定の p 値
Table 4　P-values of t-test on nDCG of each method.

	既存④TF-IDF 法	既存⑤TF 法
提案① 単語+ユーザ	0.085	0.035 *
提案② 単語	0.136	0.046 *
提案③ ユーザ	0.400	0.187

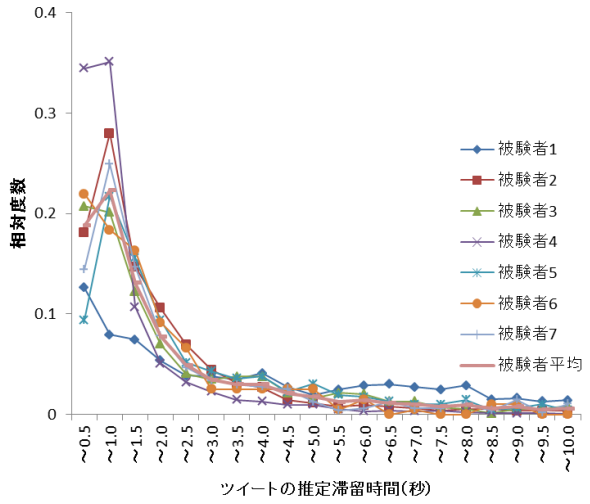

図 8　被験者ごとの推定滞留時間の相対度数分布
Fig. 8　Relative frequency distribution of estimated retention times for each subject.

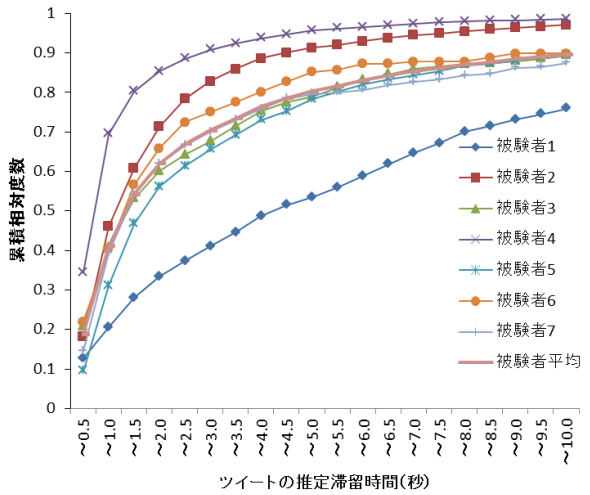

図 9　被験者ごとの推定滞留時間の累積相対度数分布
Fig. 9　Cumlative relative frequency distribution of estimated retention times for each subject.

価値が異なるコンテンツが同列に存在するため，ユーザは選択的に読み飛ばしを行っている」と述べた．この前提を検証するため，実験結果より各被験者が読んだツイートの滞留時間を調査した．7 名の被験者ごとの滞留時間，およびそれらの平均値に関する相対度数分布を表すグラフを**図 8** に，累積相対度数分布を表すグラフを**図 9** に示す．また，実験期間中に被験者の閲覧対象となった総ツイート数を**表 5** に示す．

表 5　被験者が実験期間中に閲覧した総ツイート数
Table 5　The number of tweets by each subject.

	閲覧した総ツイート数
被験者 1	1,450
被験者 2	1,896
被験者 3	686
被験者 4	5,149
被験者 5	1,370
被験者 6	196
被験者 7	470
平均	1,602.4

4.4　考察

4.3.1 項に示した実験結果において，ユーザの評価値の平均，および DCG の 2 種類の指標に関しては，「興味単語プロファイルと興味ユーザプロファイルを併用」した提案手法①，「興味単語プロファイルのみを用いた」提案手法②はともに，既存手法④TF-IDF 法，および既存手法⑤TF 法との間に有意な差が見られた．このことから，提案手法①および提案手法②は，既存手法に比べ，ユーザの興味を反映させたユーザの興味のあるツイートの抽出が可能であることが示された．

nDCG の指標に関しては，提案手法①および提案手法②は，既存手法④および既存手法⑤よりも，高い値を示しているが，有意差は既存手法⑤のみにしか観測できなかった．したがって，抽出されたツイートのランキングの妥当性に関しては，既存手法より有効であることは示せなかった．

すべての評価基準において，「興味ユーザプロファイルのみを用いた」提案手法③と既存手法の間には有意差が観測されなかった．これは，たとえユーザが普段注目する発信者であっても，発信者がつねにユーザにとって興味のあるツイートをするわけではない可能性が高いことを表している．

今回の実験では，すべての評価基準において提案手法①は提案手法②よりも高い値を得たが，有意差は観測されなかった．この理由として，今回の実験では，コンテンツ価値 value を求める際の，興味ユーザプロファイルへのパラメータ α の値が小さかった可能性がある．実験により最適なパラメータ α を検討することは今後の課題である．また，ユーザごとにパラメータ α の値が変化することも考えられるため，ユーザごとにパラメータ α の値を最適化する手法も検討する必要があると考えている．

4.3.2 項に示した実験結果から，平均で 50%以上のツイートの滞留時間が 1.5 秒以下であり，最も読み飛ばしの割合が少ないと思われる被験者でも 30%弱が 1.5 秒以内となっていた．1.5 秒もかけずに読むことが可能なツイートは多数存在するが，ソーシャルストリーム上の多くのツイートが読み飛ばされている傾向にあると考えられる．したがって，読み飛ばしを前提とした提案手法が有効である状況は

多いと考えられる．

5. 議論

5.1 滞留時間の推定における課題と改善案

本手法では，ユーザのスクロール操作に注目し，どのツイートにどのくらいの時間をかけて読むか（滞留時間）をもとに興味単語プロファイルの構成を行った．しかしこの滞留時間は，ユーザが画面中心を読むという前提であり，その誤差による影響は無視できない．被験者数名に聞き取りしたところ，被験者が注目していたツイートと滞留時間の長かったツイートとに差が見られた者もいた．今回は単純な補正と窓関数での平滑化を行ったが，明らかに特定のツイートへのアクションとして認識可能な「お気に入り」，「リツイート」，「返信」等の操作時に，その対象ツイートが画面のどの辺りに位置するかを抽出することで，ユーザごとにソーシャルストリームを読む際の「注視点の位置」を推定することで，精度が向上すると期待できる．

また，滞留時間が長かったものの，「偶然見ただけで興味はなかった」との回答を得たツイートもあった．これは我々の「滞留時間が長いツイートは，ユーザが興味を持ったツイート」という前提と異なるものであり，滞留時間の推定による興味の抽出の限界でもある．しかし，システムの利用期間が長くなれば，そのような偶然見たツイートによる影響は少なくなっていくと期待できる．

5.2 閲覧時のユーザコンテクストの考慮

本論文では，タイムラインに関するユーザの振舞いを利用して，ユーザの興味を反映したユーザプロファイルを構成する手法を提案した．しかし，ユーザのコンテクストに応じてユーザの振舞いは異なると考えられる．たとえば，時間的に余裕があるときには，興味があるツイートの多くの内容を正確に確認することが多いと思われる．しかし，忙しい朝や隙間時間等，時間的な余裕がないときには，読み飛ばしをしながらの「選別」基準は偶然によるものが大きい．大きくスクロールし偶然立ち止まった箇所のみを読むユーザや，読んではいるがほとんど頭に残っていない等，コンテンツ本来の価値にかかわらず，コンテンツを重要なものとして扱わないことが多い．

また，見落としコンテンツを抽出する際にも，ユーザに時間的な余裕がある状況では，見落としコンテンツが比較的少ないと考えられるが，ユーザに時間的な余裕がない状況では，読み飛ばしたコンテンツの中に多くの見落としコンテンツが含まれている可能性が高い．

したがって，ユーザプロファイルの構成および見落としコンテンツ抽出の際に，ユーザコンテクストを考慮することが重要であると考えられる．たとえば，コンテクストを，コンテンツの価値を判断しながらソーシャルストリームを読んでいるコンテクスト「free」と，そうでないコンテクスト「busy」とに区別することを考える．このとき，free のときに読んだツイートの中から，滞留時間の長いものほどユーザが興味を引かれたツイートとし，滞留時間の長短によってユーザプロファイルを構成する．最後にユーザのコンテクストが busy のときの滞留時間の短いツイート群の中から，ユーザプロファイルに適合するツイートを，ユーザにとって価値の高いツイート（見落としコンテンツ）とすることにより，効果的なシステムが実現できると期待できる．この，システムの処理の流れを**図 10** に示す．

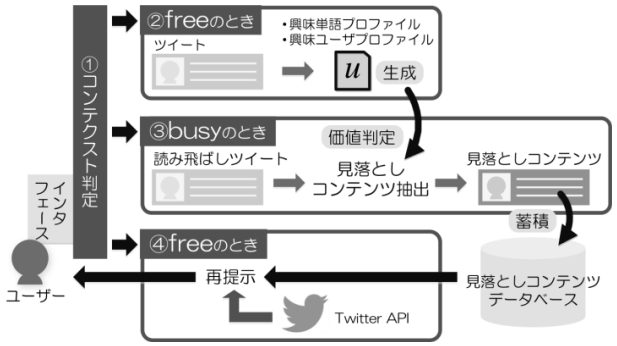

図 10　コンテクストを考慮したシステムの概要
Fig. 10　An overview of context-aware system.

① ユーザがシステムを利用しているとき，システムはユーザの振舞いからコンテクスト（busy/free）を判定する．
② ユーザが free のときは，ユーザの振舞いからユーザの興味を表すユーザプロファイルを生成する．
③ ユーザが busy のとき，ユーザが読み飛ばしたコンテンツを抽出し，ユーザプロファイルに合致する価値の高いコンテンツを見落としコンテンツと判定し，見落としコンテンツデータベースに蓄積する．
④ ユーザが free のとき，見落としコンテンツデータベースに蓄積された見落としコンテンツを新しいソーシャルストリームと融合して再提示する．

以上の流れによって，ユーザは能動的な操作なしに，見落としていた価値の高い情報を活用可能となる．本論文で示したプロトタイプシステムは，ユーザの振舞いからユーザプロファイルの作成手法の開発と検証を行うことを目的とするため，コンテクストの自動判定および処理①，④の実装は行っておらず，②と③のみを実装したものであると考えることができる．コンテクストの自動判定については，ユーザの振舞いを特徴量とした機械学習により実現できると考えている．具体的には，ユーザの閲覧セッションにおける，ツイートの滞留時間の分布，ツイート投稿者ごとの滞留時間の分布，ツイートに含まれるリンクのクリックの頻度，お気に入りへの登録頻度等を特徴量として利用することが有効であると考えている．

6. おわりに

本論文では，ソーシャルストリーム中の見落としコンテンツを抽出することを目的として，ソーシャルストリーム閲覧時の振舞いに基づいてユーザプロファイルを構成する手法を提案した．被験者実験の結果，従来手法と比較して有効性があることが示された．

本研究では，視線計測装置等の特別な装置を用いずに，スクロール等のユーザのブラウザ操作のみをもとに，ソーシャルストリーム上のコンテンツへのユーザの着目度を「滞留時間」という指標でとらえている．作成した興味単語プロファイル，興味ユーザプロファイルの併用によって，読み飛ばしたソーシャルコンテンツ中から価値の高いコンテンツが抽出可能である．

滞留時間によるユーザプロファイル構成手法は，我々の提案するシステムやソーシャルストリームだけでなく，スマートフォン端末におけるニュース配信サイトやｅコマースサイトにおいても活用できると考えられる．

今後は，5 章で示したようなコンテクストに応じた見落としコンテンツの再提示機構を開発し，有効性を評価する予定である．

謝辞　本研究を進めるにあたり，熱心に議論していただいた，九州大学牛尼研究室の皆様に感謝いたします．また，本論文を執筆するにあたり，本論文を大変丁寧に査読していただき，有用なコメントをいただいた査読者の方々に感謝いたします．本研究は JSPS 科研費 24500119 の助成を受けたものです．

参考文献

[1] Shneiderman, B., Preece, J. and Pirolli, P.: Realizing the value of social media requires innovative computing research, *Comm. ACM*, Vol.9, No.9, pp.34–37 (2011).
[2] Chen, J., Nairn, R. and Chi, E.: Speak little and well: Recommending conversations in online social streams, *Proc. SIGCHI Conference on Human Factors in Computing Systems* (*CHI '11*), pp.217–226 (2011).
[3] 大野健彦：視線を用いたインタフェース, 情報処理, Vol.44, No.7, pp.726–732 (2003).
[4] Leiva, L.A.: MouseHints: Easing Task Switching in Parallel Browsing, *Proc. 2011 annual conference extended abstracts on Human factors in computing systems* (*CHI 2011*), pp.1957–1964 (2011).
[5] What is edgerank? (online), available from ⟨http://whatisedgerank.com/⟩ (accessed 2013-07-22).
[6] 土方嘉徳：情報推薦・情報フィルタリングのためのユーザプロファイリング技術，人工知能学会論文誌，Vol.19, No.3, pp.365–372 (2004).
[7] 梅本和俊, 山本岳洋, 中村聡史, 田中克己：ユーザの視線を利用した検索意図推定とそれに基づく情報探索支援，日本データベース学会論文誌，Vol.10, No.1, pp.61–66 (2011).
[8] Umemoto, K., Yamamoto, T., Nakamura, S. and Tanaka, K.: Search Intent Estimation from User's Eye Movements for Supporting Information Seeking, *Proc. International Working Conference on Advanced Visual Interfaces*, pp.349–356 (2012).
[9] 松尾　豊，福田隼人，石塚　満：ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援，人工知能学会論文誌，Vol.18, No.4, pp.203–211 (2003).
[10] Morita, M. and Shinoda, Y.: Information filtering based on user behavior analysis and best match text retrieval, *Proc. 17th annual international ACM SIGIR conference on Research and development in information retrieval* (*SIGIR '94*), pp.67–74 (1994).
[11] Buscher, G., van Elst, L. and Dengel, A.: Segment-level display time as implicit feedback: A comparison to eye tracking, *Proc. 32nd international ACM SIGIR conference on Research and development in information retrieval* (*SIGIR '09*), pp.67–74 (2009).
[12] 林　春男，佐藤翔輔：膨大な情報から必要とされる情報を報せるビジネスツールとしての TRENDREADER，情報管理，Vol.54, No.1, pp.2–12 (2011).
[13] 徳永健伸，辻井潤一：情報検索と言語処理，東京大学出版会 (1999).
[14] Jarvelin, K. and Kekalainen, J.: Cumulated gain-based evaluation of IR techniques, *ACM Trans. Inf. Syst.*, Vol.20, No.4, pp.422–446 (2002).

土岐 真里奈

1990 年生．2013 年九州大学芸術工学部芸術情報設計学科卒業．在学中，SNS のインタフェースに関する研究に従事．現在，株式会社サイバーエージェント勤務．

牛尼 剛聡　（正会員）

1970 年生．1999 年名古屋大学大学院工学研究科情報工学専攻博士課程後期課程満了．1999 年九州芸術工科大学芸術工学部助手．2011 年九州大学大学院芸術工学研究院准教授．日本データベース学会，電子情報通信学会（シニア会員），ヒューマンインタフェース学会，ACM，IEEE-CS 各会員．

（担当編集委員　宇田川 佳久）