

ソーシャルストリームの読み飛ばしを考慮した コンテキスト適合型インタフェース

土岐 真里奈[†] 牛尼 剛聡[‡]

[†]九州大学芸術工学部 〒815-8540 福岡県福岡市南区塩原 4-9-1

[‡]九州大学大学院芸術工学研究院 〒815-8540 福岡県福岡市南区塩原 4-9-1

E-mail: [†] mari7toki@gmail.com, [‡] ushiama@design.kyushu-u.ac.jp

あらまし

Twitter 等のソーシャルストリームを閲覧する際、ユーザは全てのコンテンツを読む訳ではなく、価値が低いと推定されるものを読み飛ばすことが多い。しかし、コンテンツの価値は常に同一ではなく、ユーザのコンテキストによって変化する可能性がある。例えば、忙しいときに読み飛ばされたコンテンツの中にも、時間的に余裕があるときにはユーザが価値を感じるものが含まれる可能性がある。本論文では、時間的に余裕がないコンテキスト「busy」とそれ以外の「free」を想定し、ユーザが「busy」のときに読み飛ばされたコンテンツから、ユーザにとって価値の高いコンテンツを抽出し、「free」のときに再提示することにより、ユーザがソーシャルストリームから効果的に情報を取得可能とするインタフェースを提案する。そして、コンテキストが free であるときの、ブラウザ上でのユーザの振る舞いによってユーザの興味プロファイルを生成し、価値の高いコンテンツを抽出する手法を提案する。そして、被験者実験に基づいて、提案手法の有効性を定量的に検証する。

キーワード ソーシャルコンテンツ, 読み飛ばし, コンテキストアウェアネス, コンテンツの価値, 個人化, ソーシャルストリーム, SNS, ユーザの振る舞い, tf-idf, インタフェース

A Context-Adaptive Interface for Social Stream Suitable for Skip Reading

Marina TOKI[†] and Taketoshi USHIAMA[‡]

[†] Graduate School of Design, Kyushu University 4-9-1 Minami-ku, Fukuoka-shi, Fukuoka, 815-8540 Japan

[‡] Faculty of Design, Kyushu University 4-9-1 Minami-ku, Fukuoka-shi, Fukuoka, 815-8540 Japan

E-mail: [†] mari7toki@gmail.com, [‡] ushiama@design.kyushu-u.ac.jp

Keyword Social Contents, skipping, Context Awareness, Value of Contents, Personalization, Social Stream, SNS, user's behavior, tf-idf, interface

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常的に利用されるようになった。ユーザが SNS に投稿した「ソーシャルコンテンツ」は、その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。ユーザは複数のユーザをフォローするために、情報が断続的に配信される。このように、多くの SNS では、断続的に、様々なユーザから様々な種類のコンテンツが配信され、システムは配信されたコンテンツを一系列に並べた形式でユーザに提示する。SNS 上でユーザに配信されたコンテンツ列は、一種のストリームデータと考えることができ、「ソーシャルストリーム」と呼ばれる。

一方、スマートフォンやタブレット型コンピュータの普及で、

場所と時間を問わず SNS を利用可能となった。ソーシャルストリームには次々に新しいコンテンツが追加されるため、多くのユーザは、日常生活の隙間時間を利用してソーシャルストリームを閲覧するようになった。

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、ユーザにとって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一つ一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あれば、有効に活用できるかもしれない。我々は、ユーザにとって価値が高いにも関わらず、ユーザが閲覧する際のコンテキストが適切でなかったために、活か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。

本研究では、ユーザのコンテキストを、時間的に余裕がない状況である「busy」と、時間的に余裕がある状況の「free」の2種類に分類する。このとき、見落としコンテンツは、「ユーザが内容を確認しなかったコンテンツの内、ユーザのコンテキストが busy のときに提示され、ユーザにとって価値の高いコンテンツ」と定義できる。見落としコンテンツの定義を図示したものを図 1 に示す。

本論文で提案するインタフェースでは、ソーシャルストリームを読むユーザの振る舞いから、ユーザのコンテキストが busy であるか、free であるかを自動的に区別する。そして busy のときに流れたコンテンツの中から、ユーザの興味プロフィールに適合する価値の高いコンテンツ(見落としコンテンツ)を抽出する。

本研究では、ユーザの興味プロフィールを、ソーシャルストリームの閲覧履歴から自動的に作成する。そのために、一般的な TF-IDF 法にコンテキストの要素を取り入れることで、ソーシャルコンテンツにより適した手法である TF-IDF-RT 法を開発した。

そして、コンテキストが free のときに、ソーシャルストリームの再構成により、抽出した価値の高いコンテンツをユーザに再提示する。これにより、ユーザがコンテキストに応じて適切なタイミングで適切な情報を受け取ることで、コンテンツの価値を最大限に活かすことを支援する。

本論文の構成は以下の通りである。第2章で関連研究を紹介する。第3章では、提案システム「Context Catcher」の概要を述べる。第4章ではシステムを実現するための以下の手法について述べる。ユーザの振る舞いデータの取得方法と、コンテキスト判定の手法、さらにコンテキストと滞留時間に着目した興味プロフィール作成手法と、コンテンツ価値判定手法について述べる。第5章で被験者実験について述べ、第6章で考察を行い、第7章でまとめを述べる。



図 1 見落としコンテンツの位置づけ

2. 関連研究

これまでに、膨大かつ多様な情報の中から、効率的に価値の高い情報を選別する研究がなされている[1]。例えば Facebook で利用されている、EdgeRank アルゴリズムでは、ユーザと投稿者との「親密度」という指標をもとに、表示するコンテンツをランク付けしている[2]。しかし EdgeRank では主にユーザの交友関係に重きを置いており、ユーザ自身の嗜好とコンテキストを考慮していない。そのため、ユーザの興味のある投稿が読み飛ばされ、充分に活用されない可能性がある。

松尾ら[3]は、web 上のユーザの振る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、ユーザが閲覧した文書の履歴から、ユーザにとって重要度の高い語を抽出し、ブラウジングでの読み飛ばしを防ぐ。この手法では、ユーザが閲覧する度にユーザにとって「身近な語」を収集し、文書を閲覧する際に「身近な語」との共起語を重要語とする。興味をユーザの過去の閲覧文書から抽出し、個人化に利用する点は、本研究と共通している。しかし、この手法では、一般的な web 文書を対象としているため、ユーザの閲覧時のコンテキストの違いは考慮していない。また、ソーシャルストリーム上のコンテンツは、ユーザの閲覧時のコンテキストに強く依存して、その価値が変化すると考えられるため、「読み飛ばし」が起こる要因が異なると考えられる。さらに、松尾らの手法では、コンテンツを「読み飛ばされないように」重要語をハイライトするのに対し、本研究では「読み飛ばされた」重要なコンテンツを、ユーザに時間的な余裕があるときに補助的に再提示することで、ユーザのコンテキストに応じてコンテンツの有効的な活用の支援を目的としている。

ユーザのコンテキストに着目した研究として、中野ら[4]は、発想支援において、オフタイム(休息のための時間)を有効に活用することに着目し、オンタイム(業務や勉強に充てる時間)で取得したユーザの暗黙的な興味情報に基づき抽出したテレビ番組をオフタイムに埋め込むことにより、ユーザの発想の広がりを促進するシステムを提案している。このシステムでは、オンタイムに於ける検索クエリ等からユーザの興味を推定し、ユーザの興味に関するテレビ番組を、オフタイム(発想しようと意識していないコンテキスト)に提示する。これにより、オンタイムで得られなかった発想が生まれ、同じコンテンツでもコンテキストに応じてユーザに与える影響が異なることを目指している。この研究では、具体的に有効な結果は出なかったものの、意外性のある興味深い番組抽出に成功している。

林ら[5]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。時間とともに発生するソーシャルコンテンツには、通常の TF-IDF 法では最新の状況を的確に反映することができな

いとし、文書数の増加に応じて IDF 項の値が変化するように、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。本手法では「変化」のあったコンテンツを重要だと捉えているが、我々の手法では、ユーザが興味のあるコンテンツが対象であり、時間経過によっても変化のない情報も、重要なものだと考える。

3. システムの概要

我々は「同一のコンテンツでも、ユーザのコンテキストによって情報の価値は異なる」という仮説をたてた。本論文で提案する SNS クライアント「Context Catcher」は、この仮定に基づきユーザが忙しい時間に閲覧した見落としコンテンツを抽出し、時間的な余裕がある時に再提示することにより、コンテンツが本来有する価値を生かすというアプローチをとる。なお、本研究では SNS の中でも、投稿数が多く、読み飛ばされやすいと考えられる Twitter を対象とする。

本システムの処理の流れを図2に示す。

- 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る舞いからコンテキスト(busy/free)を判定する。
- ② ユーザが free のときは、ユーザの振る舞いからユーザの興味を表す2つの興味プロファイルを生成する。興味プロファイルは、興味のある事項を表す「興味単語プロファイル」と、興味のあるユーザを表す「興味ユーザプロファイル」から構成される。
- 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、興味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、見落としとしていた価値の高い情報を活用可能となる。

本システムは、PHP と JavaScript を利用して記述され、JavaScript が動作する Web ブラウザ上で動作する。また、データベース管理システムとしては MySQL を利用している。

図 3 に Context Catcher のインタフェースのスナップショットを示す。本研究では、ユーザの自然な SNS の利用法のまま、コンテンツ価値を向上させることを理想としている。その

ため、特殊な設備を使用せず、インタフェースデザインも最も一般的であると考えられる公式の Twitter クライアントのデザインを踏襲している。



図 3 Context Catcher のインタフェース

4. 提案手法

4.1. コンテキスト判定

4.1.1. ユーザの振る舞いデータ

本研究では、ソーシャルストリーム閲覧時のユーザの振る舞いを抽出し、以下の目的のために利用する。

1. ユーザのコンテキスト(busy/free)の判別 (図 2①)
2. ユーザの興味の抽出 (図 2②)
3. 見落としコンテンツの抽出 (図 2③)

ユーザの振る舞いを抽出目的として、これまでに多くの手法が提案されている。例えば、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ために、アイトラッキングが用いられることが多い[6]。これは、ユーザが今注視しているスクリーンの位置と時間を、眼球の動きをアイトラッカー等の特別な装置を利用して取得し、分析する手法である。アイトラッキングを利用することにより、ユーザの興味を抽出可能である。しかし、全てのユーザがアイトラッカーを利用することを前提とすることは現実的ではない。そこで、本研究では特別な装置を利用せずにユーザの振る舞いを抽出することを目指す。このため、ブラウザ操作を元にユーザの振る舞いを取得する。今回対象とするソーシャルストリームでは、一次的にストリーム化されたコンテンツをスクロールによって閲覧することが多い。そこで、特にスクロール操作に注目する。他に、ソーシャルストリームを構成する個々のコンテンツ(ツイート)に対して、「表示領域の外に出た時刻」、「外部リンクのアクセス」、「リツイート」、「お気に入り登録」等をユーザの振る舞いに関するデータとして取得する。振る舞いデータはそれぞれイベントとして管理され、発生時刻とその識別番号(id)をもとに RDB 上に記録される。

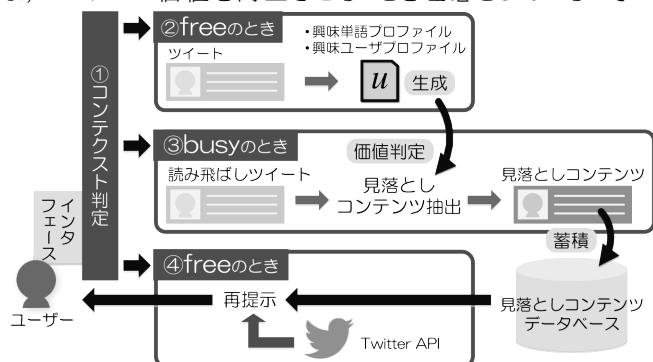


図 2 システムの処理

4.1.2. 滞留時間の推定

一般的に、ユーザは画面中央のツイートを読むと考えられるため、個々のツイートが表示領域の外に出た時刻に基づいて、中央に位置するツイートの滞留時間を推定する(図4). 具体的には、振る舞いデータのうち、ソーシャルストリームを構成するツイートがスクロールによりブラウザ画面の外部に移動したことを表す `displayEnd` イベントを用いる. いま、ツイート T_{w_j} が、スクロールによってブラウザの表示領域の外に出た時刻を `displayEndj` と表すことにする. また、表示領域上に n 個のツイートが表示されているとする. このとき、ユーザは表示領域内に表示されているツイートのうち、上から $n/2$ 番目に表示されているツイートを注目していると考え. そこで、 T_{w_j} を読むのにかった時間 rt_j を、ツイート $T_{w_{j-(n/2)-1}}$ が表示領域から出た時刻 `displayEndj-(n/2)-1` と、ツイート $T_{w_{j-(n/2)}}$ が画面から出た時刻 `displayEndj-(n/2)` の差として推定する.

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る舞いを行うことがある. 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、上記のような振る舞いに要する時間が含まれる可能性がある. また、ツイート長によっても、閲覧に要する時間は異なると考えられる. そこで、上記の要因を考慮して、滞留時間を正規化する. また、ユーザの注目点が中心からずれることがあるため、窓関数を適応し平滑化を行う. 本手法によって計測された滞留時間の例を図5に示す.

4.1.3. コンテキストの判別

システムを利用した際のユーザのコンテキスト(`busy` または `free`)は、システムを起動してソーシャルストリームを閲覧するセッション毎に判別する. 具体的には、そのセッションにおけるユーザの平均的な滞留時間および平均的なスクロール速度を利用して閾値に基づいて判別する.

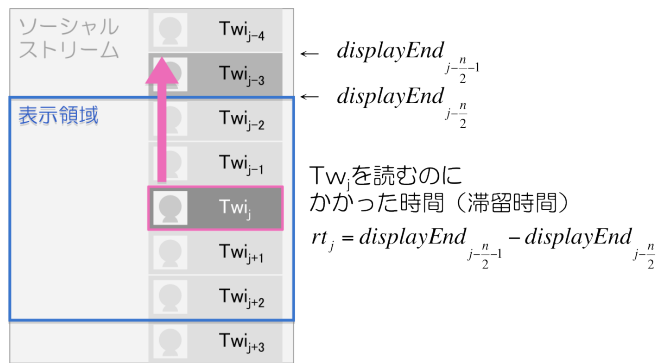


図4 ツイート単位の滞留時間の計算方法

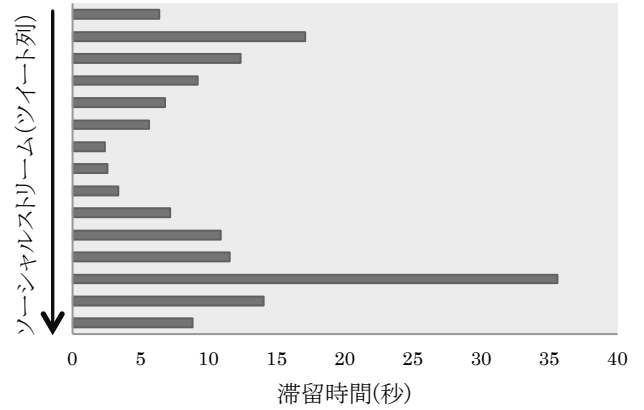


図5 推定された滞留時間の例

4.2. コンテンツ価値判定

見落としコンテンツの抽出のために、コンテンツの価値判定が必要である. コンテンツの価値判定のために、本システムでは、ユーザのコンテキストが `free` のときのソーシャルストリームに対するユーザの振る舞いを利用して、ユーザの興味を表す興味プロフィールを生成する. 次に、ユーザのコンテキストが `busy` のときに提示されたツイートのうち、見落としコンテンツとして再提示するべきものを、興味プロフィールを利用して抽出する.

本研究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ィールで表し、後者を興味ユーザプロフィールで表現する

4.2.1. 興味単語プロフィールの生成

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. 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 に於ける単語 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(1)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 に於ける単語 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.

$$tfidf_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right) \quad \dots (1)$$

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、個々のツイートを一つの文書としてとらえ、個々の文書に出現する単語の重要度をTF-IDF法をそのまま適用して求めることは適切ではない. ツイートは140字以内であるため、一つの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、ほとんどのツイートに於いて、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と考えられる。

この問題を解決するために、ユーザが興味を示しているだろうお気に入りツイートやリツイートしたツイートに含まれる単語を、「ユーザが興味のある単語」とする方法が考えられる。しかし、全てのユーザが興味のあるコンテンツ全てにお気に入りやリツイートといった操作をする訳ではない。また、ソーシャルコンテンツには「重要だが一度読めば十分である」ものも多く含まれている。そのため、お気に入りツイートに於ける重みは、ユーザにより大きく異なる要素となる。そこで、我々はユーザのツイートの閲覧時間を考慮する。

一般的に、ユーザがツイート群を閲覧するとき、興味のあるツイートをより時間をかけて読むと考えられる。また、freeの時に読むツイート群の方が、busyの時に読むツイート群よりも、ユーザはツイートの好き嫌いを選別していると考えられる。このことから我々は、freeのときに読んだツイート群の中から、より時間をかけて読んだツイートを「興味のあるツイート」と考える。そして、TF-IDF法を拡張したTF-IDF-RT法を提案する。単語*i*のツイート*j*に於ける重み $w_{i,j}$ の定義を式2に示す。

$$w_{i,j} = tfidf_{i,j} \times rt_j \quad \dots (2)$$

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時間(滞留時間) rt_j を掛け合わせている。これにより、お気に入りやリツイートといった明示的な操作がなくても、ツイートとそこに含まれる単語への着目度を、時間という観点から考慮可能になる。

なお、今回対象としたTwitterでは、一つの投稿は140字以内であり、1文書の文書長が短いため、滞留時間が閾値を超えた場合にツイート中の単語 $w_{i,j}$ に関しては $tf_{i,j}=1$ とし、閾値未満未満のツイート中の単語に関しては $tf_{i,j}=0$ とした。今回は、滞留時間の閾値は経験的に2秒と設定した。

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ィール $\mathbf{u}_{word}$ は、それぞれの単語*i*について、TLに含まれる全てのツイート*j*の重み $w_{i,j}$ の和として定義する(式3)。 $\mathbf{u}_{word}$ は、ユーザがどの単語にどの程度価値を感じるかというプロフィールである。この際、数字・漢数字のみやひらがな一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でよく見られる意味のない語はストップワードとした。

$$\mathbf{u}_{word} = \begin{pmatrix} \sum_{j \in TL} w_{1j} \\ \sum_{j \in TL} w_{2j} \\ \vdots \end{pmatrix} \quad \dots (3)$$

例として、著者が10日間システムを利用し作成された興味単語プロフィールの中で、重みが大きい上位15件の単語

を図6に示す。なお、ユーザが意識的に興味があると思っている単語と、無意識的に興味を持っている単語との相違がある。図6に示されているのは、ユーザが「意識的に選択している単語」(趣味やトレンドの単語)ではなく、ユーザが「どういった単語を用いるツイートを好むのか」という傾向である。

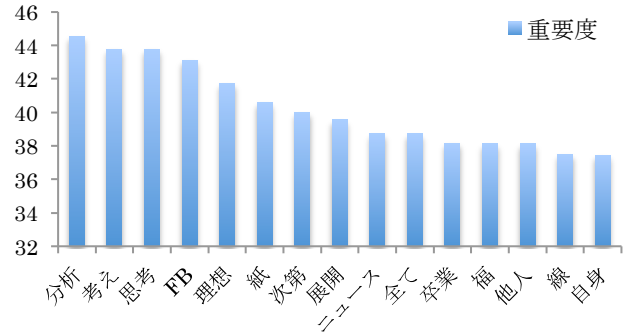


図6 興味単語プロフィールの例

4.2.2. 興味ユーザプロフィールの作成

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、発信者の情報がある。多くのTwitterクライアントのインターフェースデザインでは、ツイートの左側に発言者のプロフィールアイコンが表示されるため、ユーザは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。ユーザが発信者の重みを抽出するために、発信者*k*のツイートを、ユーザがどの程度時間をかけて読んでいるかという度合い $uval_k$ を用いる(式4)。ここで、ユーザがソーシャルストリームに表示した*k*のツイートの集合を $TW(k)$ 、ツイート*j*の滞留時間を rt_j 、ツイート*j*の総単語数を $word(j)$ とする。

$$uval_k = \sum_{j \in TW(k)} \frac{rt_j}{word(j)} \quad \dots (4)$$

興味ユーザプロフィール $\mathbf{u}_{user}$ は、式(4)で求められた $uval_k$ を全てのユーザに関してベクトルとして表現したものである。

$$\mathbf{u}_{user} = \begin{pmatrix} uval_1 \\ uval_2 \\ \vdots \end{pmatrix} \quad \dots (5)$$

4.2.3. コンテンツ価値判定

興味単語プロフィール $\mathbf{u}_{word}$ と、興味ユーザプロフィール $\mathbf{u}_{user}$ を利用して、ユーザのコンテキストがbusyのときに読み飛ばされたツイートの中から、見落としコンテンツを抽出する。

興味単語プロフィール $\mathbf{u}_{word}$ と、興味ユーザプロフィール $\mathbf{u}_{user}$ が与えられたとき、ツイート*tw*に対するコンテンツ価値 $value(tw)$ を以下の式で計算する。なお、simはツイート*tw*と興味プロフィールとの類似度を表す関数であり、 α はパラメータである。

$$value(tw, \mathbf{u}_{word}, \mathbf{u}_{user}) = \text{sim}(tw, \mathbf{u}_{word}) + \alpha \text{sim}(tw, \mathbf{u}_{user}) \quad \dots (6)$$

読み飛ばされたツイートそれぞれに対してコンテンツ価値を計算し、その値が高いものを、見落としコンテンツと判定する。

例として、著者が 10 日間システムを利用し 2 つの興味プロフィールを作成し、後日機械的に収集した対象ツイート 773 件に、次の 3 パターンでプロフィールを適用し推薦された上位ツイートを表 1～3 に示す。なお、下線を引いた語はプロフィールに適合した単語である。

- ① 興味単語プロフィールと興味ユーザプロフィールを併用して適用 (表 1)
- ② 興味単語プロフィールのみ適用 (表 2)
- ③ 興味ユーザプロフィールのみ適用 (表 3)

表 1 の 1 位のユーザは興味ユーザプロフィールの値は高くないが、興味単語プロフィールに該当する単語を多数含むツイートであった。2・3 位は著者の興味ユーザプロフィールの値が比較的高い友人である。ツイート自体にも著者の最近の関心事が反映されており、「理想」、「部屋」、「卒」、「研」といった比較的上位の興味単語が含まれている。また、表 1 の 3 位のツイートは表 2 では上位にランクインしなかったものだが、興味ユーザプロフィールを併用することで上位に上がっている。実際に、著者は以前表 1 の 3 位のツイートにリプライを返しており、十分に興味ツイートであったと考えることができる。また、表 3 の興味ユーザプロフィールのみの推薦では、興味ユーザプロフィールで上位にあたるユーザの発言がランダムに入るため、必ずしも興味が引かれる内容ではなかった。

以上より、表 3 のように興味ユーザプロフィールのみを利用すると、「特定ユーザの発言は全て」が「良い」ツイートとして扱われてしまう。一方、表 2 のように興味単語プロフィールのみでは、文脈によらず興味単語が含まれてさえいえば上位にランクインしてしまうため、これも適切とは言えない。また、単語数の多いツイートが比較的上位に入っているという特徴が見られた。

表 1 ①興味単語・興味ユーザプロフィールを併用して適用し推薦されたツイート

順位	ユーザ	コンテンツ 価値	ツイート
1	F	0.492	一握りのエリートの後ろには何千何万という俺達のような人間がいるんだ！そして社会の歯車、使い捨て、操り人形などと言われながらも、 <u>自分の夢</u> だったり家族や恋人の幸せだったり、…
2	G	0.483	【軽量鉄骨造】には防音性求めちゃ駄目なのか。【RC 造】がいいのか。【RC 造防音】でググっても、 <u>安心</u> できん。監禁犯が、分厚い壁の角部屋&隣空き部屋に住んで、 <u>誰も</u> 気づかなかった。てのがあった気が…
3	D	0.449	この時期にやることではございませんが、 <u>明日</u> こそ <u>そりわたくし</u> の卒研作品見に来てやってほしいぞっていう方いらっしやいますかね。

表 2 ②興味単語プロフィールのみを適用し推薦されたツイート

順位	ユーザ	コンテンツ 価値	ツイート
1	F	0.157	一握りのエリートの後ろには何千何万という俺達のような人間がいるんだ！そして社会の歯車、使い捨て、操り人形などと言われながらも、 <u>自分の夢</u> だったり家族や恋人の幸せだったり、…
2	G	0.155	【軽量鉄骨造】には防音性求めちゃ駄目なのか。【RC 造】がいいのか。【RC 造防音】でググっても、 <u>安心</u> できん。監禁犯が、分厚い壁の角部屋&隣空き部屋に住んで、 <u>誰も</u> 気づかなかった。てのがあった気が…
3	H	0.142	RT @KometsubuKomeo: 『髑髏城の七人』テレビではまず中々出会えない熱量だったと思います。魂、ってやつですかね？生の人間(スクリーンやけど)が今そこできさに演じて造ってるって感じがしました。…

表 3 ③興味ユーザプロフィールを適用し推薦されたツイート

順位	ユーザ	コンテンツ 価値	ツイート
1	A	1	“マンチェスターユナイテッド対リヴァプールのレビュー 並びに香川のプレススタイルの <u>話</u> - pal-9999 の日記” http://t.co/Mp1sg7Rm
2	B	0.777	汚れた壁のフリーテクスチャ素材 20 http://t.co/tFTn7AUv
3	B	0.777	レスポンス・ウェブデザインの流行パターンまとめ http://t.co/HWAKD3Zb

5. 実験

5.1. 実験内容

提案手法のなかで、ツイートの滞留時間に基づいたユーザの興味プロフィールの生成手法及び、生成した興味プロフィールを用いたツイートの価値判定手法の有効性を評価するための実験を行った。被験者は、19～25 歳の 9 人である。被験者には 5 日間、パソコンで Twitter を閲覧する際に、実験用システムを用いソーシャルストリーム(タイムライン)を閲覧してもらった。なお、被験者のコンテキストを free のときに限定するために、システムの利用は時間のある時のみに限定し、その他(数秒しか開かないとき、携帯端末での利用時等)では、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。その他、ツイート、お気に入り、リツイート、リプライなどの操作もシステムから行ってもらった。

次に、実験で得られたデータを用い各被験者の興味単語プロフィールと興味ユーザプロフィールの作成を行った。実験最終日より 7 日経過した後に、新たにアンケート用に被験者それぞれのソーシャルストリームから 700～800 件のツイート(対象ツイート)を取得した。提案手法と既存手法を含めた 5 つの手法によって対象ツイート全てのコンテンツ価値を計算し、推薦ツ

イトを選出した．今回対象とした 5 つの手法は以下の通りである．

- ① 興味単語プロフィールと興味ユーザプロフィールを併用した推薦手法
- ② 興味単語プロフィールのみを用いた推薦手法
- ③ 興味ユーザプロフィールのみを用いた推薦手法
- ④ TF-IDF 法を用いる推薦手法
- ⑤ TF 法を用いる推薦手法

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．

5.2. 実験結果

被験者 9 名の内、2 名はデータ不足のため検証不可能とし除外した．以下、実験結果の検証は有効なデータ量が得られた 7 名のデータを元に進める．

7 名の被験者の評価を元に、各手法の有効性を比較するため、推薦ツイートの面白さについて有意な差があるかを t 検定を用いて調べた結果を図 7 および表 4 に示す．提案手法①と②の推薦結果への評価が最も値が高くなっている．提案手法①と既存手法④、提案手法①と既存手法⑤に於いて有意水準 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に於いて、有意水準 5%で有意差が見られた．

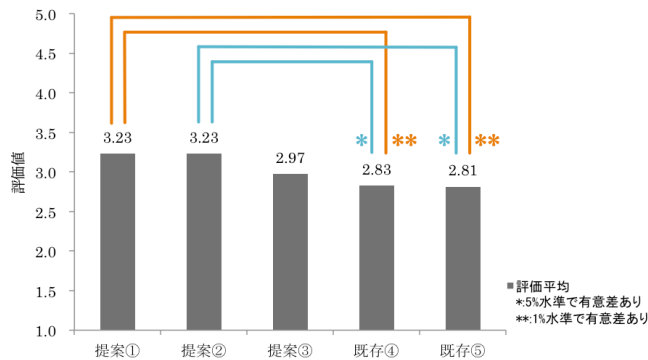


図 7 全被験者の各手法の評価

表 4 t 検定における p 値

	提案① と 既存④	提案① と 既存⑤	提案② と 既存④	提案② と 既存⑤	提案③ と 既存④	提案③ と 既存⑤
p 値	0.0088	0.0089	0.0101	0.011	0.2372	0.2047

6. 考察

6.1. 提案手法の有効性

上記の実験結果に於いて、「興味単語プロフィールと興味ユーザプロフィールを併用」した提案手法①、「興味単語プロフィールのみを用いた」提案手法②と

もに、既存手法と有意な差が見られた．しかし「興味ユーザプロフィールのみを用いた」提案手法③と既存手法には有意差は見られなかった．これは、たとえユーザが普段注目する発信者であっても、発信者が常にユーザにとって興味のあるツイートをする訳ではないことを示している．一方、提案手法①の場合、普段注目するユーザの発信する、ユーザの興味に合致した内容のツイートの推薦ができたと言える．今回提案手法①と既存手法の有意差、提案手法②と既存手法の有意差に大きな違いが出なかったが、これはコンテンツ価値 value を求める（第 5 章 式 6）際の、興味ユーザプロフィールへのパラメータの値が小さかった可能性がある．ユーザ毎にパラメータの値が変化することも考えられるため、実験により最適なパラメータを検討することが必要である．

以上より、検討すべき点はあるものの、情報収集を目的とするユーザに対し、滞留時間を用いた興味単語の抽出とコンテンツ推薦は既存手法に対し有効であると考えられる．

6.2. 滞留時間の推定に於ける課題と改善案

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）を元に興味単語の抽出を行った．しかしこの滞留時間は、ユーザが「画面中心のツイートを読む」という前提であり、その誤差による影響は無視できない．実験後、被験者数名に聞き取り調査を行ったところ、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」などの操作時に、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「目線の位置」を最適化することが可能だと考える．振る舞いデータを学習させることで、ユーザごとの目線の位置を最適化し、滞留時間の精度向上を目指したい．

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、ユーザが興味をもったツイート」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．この点に関しては、興味プロフィール作成時に滞留時間の閾値を設けることで、「偶然読んだ」ツイートへの誤差を減らすことで改善が期待できる．また、システムの継続的な利用により、長期的に出現する興味単語が、「偶然読んだ」ツイートに含まれる単語の影響を低減させることが可能と考えられる．

6.3. 抽出された重要単語

既存手法の TF-IDF 法の問題点として、4 章で「出現頻度の高い語ほど重要度が低く評価される恐れがある」と述べた。そこで、提案手法である興味単語プロファイルによる抽出単語と、TF-IDF 法による抽出単語の上位 10 件を表 5 に示す。例として被験者 B と F の結果を用いた。

被験者によらず、既存手法では「年」「日」「さん」「今日」などの、ツイート内のデータを表すだけでツイート自体の重要単語とは言えない単語が抽出されているのに対し、提案手法ではツイートの内容をふまえたものが抽出された。被験者 B はアップル社の新製品や、IT 系の新技術に着目しているユーザであり、抽出単語は適当である。なお、被験者 B の閲覧ツイートと照らし合わせたところ、「時代」「型」「便利」などの単語は、新製品や事件のニュースツイートに頻繁に含まれる単語であり、被験者 B の好むツイートの「傾向」を表す単語と言える。

表 5 提案手法と既存手法による抽出単語の比較

手法	ユーザ B				ユーザ F			
	興味単語 プロファイル		tfidf 法		興味単語 プロファイル		tfidf 法	
	単語	重要度	単語	重要度	単語	重要度	単語	重要度
1	more	37.08	年	10.49	ひとり	40.77	笑	12.15
2	技術	37.08	Brand	7.71	it	40.77	人	11.99
3	時代	35.65	円	7.63	便利	40.77	今日	10.40
4	型	34.03	デイリー	7.26	とこ	39.34	日	10.07
5	便利	33.24	日	7.11	精度	34.33	中	10.02
6	目	31.59	さん	7.04	種類	34.08	年	8.33
7	CES	31.35	新	7.03	games	34.00	私	8.28
8	警戒	31.07	月	6.71	金	32.41	分	8.04
9	ダウ	31.07	株	6.65	プレゼン	32.34	何	7.98
10	アップル	30.26	米	6.61	Harden	31.13	ノミスギタ	7.75

7. まとめ

従来、蓄積された膨大な情報の中から、ユーザに対して価値が高い情報を検索・推薦する手法は活発に研究されてきた。しかし、ソーシャルメディアの爆発的な普及のために、飽和状態の供給され続ける情報の中から、ユーザに適切なコンテンツをいかに提示するかが重要な課題となっている。本研究では、「同じコンテンツでも、ユーザのコンテキストによって情報の価値は異なる」という仮説をもとに、読み飛ばしが日常的に行われているソーシャルストリームにおける、有用なコンテンツを有効的に活用する新しいアプローチを提案した。時間に余裕のない busy のときにソーシャルストリー

ム上で読み飛ばされたコンテンツの中から、ユーザにとって価値の高い「見落としコンテンツ」を抽出し、ユーザの適切なコンテキスト(free)において再提示することで、コンテンツ価値の向上を目指した。

本研究では、アイトラッカー等の特別な装置を用いずに、スクロールなどのユーザのブラウザ操作のみを元に、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」という指標で捉える手法を開発した。また、滞留時間を用いユーザの興味を表す興味単語プロファイル、興味ユーザプロファイルの作成を行った。それに基づいて、読み飛ばされたソーシャルコンテンツの中から価値の高いコンテンツを推薦する手法を開発した。被験者実験により、この推薦手法は既存手法に対し有効であると分かった。滞留時間による興味単語抽出手法は、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 コマースサイトに於いても活用できると考えられる。

コンテンツが流れて行くことが前提であるソーシャルストリームでは、重要度の高いコンテンツの推薦精度の向上よりもむしろ、ユーザにとって無意識的に(自然な形で)受け取ることが可能なコンテンツ提示のデザインが必要であると考えている。今後は、ユーザにより受け入れやすい形でコンテンツを提供するきっかけをつくるべく、開発手法に基づきインタフェースの完成を目指す。

参 考 文 献

- ¹ Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, Ed H. Chi, “Eddi: Interactive Topic-based Browsing of Social Status Streams”, UIST '10 Proceedings of the 23rd annual ACM symposium on User interface software and technology, Pages 303-312, 2010
- ² <http://whatisedgerank.com/>
- ³ 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”, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 101(712), 85-92, 2002-03-07
- ⁴ 中野利彦, 西本一志, “オンタイムの興味情報をオフタイムに埋め込むアイディア・インキュベーション支援システム”, 情報処理学会研究報告. GN, [グループウェアとネットワークサービス] 2007(32), 1-6, 2007-03-22
- ⁵ 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 54(1), 2-12, 2011
- ⁶ 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, 一般論文 (Research Papers), 61-66, 2011-06