

既読文章におけるユーザの視線情報と興味を用いた 未読情報の再評価手法の提案

折原 レオナルド賢† 横井 健† 橋山 智訓‡

東京都立産業技術高等専門学校 創造工学専攻†

電気通信大学大学院 情報システム学研究科‡

1. はじめに

近年、多くの文章情報がインターネットを通じて容易に入手できるようになった。しかし、見るべき文章が多いことから、一つ一つを精査できず、興味のある文章だとしても、文章の冒頭や一部分のみを閲覧し、その文章の評価を行ってしまうことがある。このような閲覧方法では、文章内に読み飛ばした部分が残る、その読み飛ばした部分にはユーザの興味に沿った未読情報が存在する可能性があると考えられる。

本研究では、一度閲覧した文章において、ユーザが過去に興味があった文章と内容的に類似し、未読情報が多く含まれている場合、そこにはユーザにとって興味のある有益な情報が含まれている可能性があると考え、文章を閲覧しているユーザの視線情報を用いることで、ユーザが文章内にどの程度未読情報を残しているかが推定できる。そこで、本研究ではこの視線情報と興味を用いて、ユーザが一度閲覧した文章の中から未読情報を含む文章の再提示を行い、再評価する手法を提案する。

2. 関連研究

ユーザが情報閲覧を行っている際の行動を分析する手法としては Web 情報探索行動の分析が挙げられる。Web 情報探索行動の分析には、プロキシサーバを用いる分析[1]、ブラウザ操作に着目した分析[2]、眼球運動測定装置を用いた分析[3]といった手法が提案されている。その中でも、ユーザが見ている情報をより正確に取得するために、眼球運動測定装置を用いた分析に着目する。この手法では、実際にユーザが見た情報そのものの分析を行うことができる。

このため、本研究では眼球運動測定装置を用い、ユーザが冒頭や一部分のみを閲覧して読み飛ばした情報の判別を行う。

Re-evaluation of an unread part of an already read article using gaze tracking.

† Leonardo Ken Orihara

† Takeru Yokoi

‡ Tomonori Hashiyama

† Tokyo Metropolitan College of Industrial Technology

‡ The University of Electro-Communications

3. 提案手法

3.1. 記事の類似度

記事同士の内容的な類似度を求めるために、まず、n-gram インデキシングで記事 j の文章をパターンで分割し、出現頻度をベクトル $\mathbf{d}_j = (w_{j,1}, w_{j,2}, \dots, w_{j,V_j})$ の要素として構成する。ここで w_{j,V_j} は n-gram で分割した記事 j に含まれるパターンの出現頻度であり、 V_j はベクトル化された記事 j に含まれる全 n-gram パターン数である。

本システムでは、ユーザが興味のある記事を i とし、 i 以外の閲覧済み記事とそれぞれ類似度を式(1)で算出する。

$$Sim_{i,j} = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{\|\mathbf{d}_i\| \|\mathbf{d}_j\|} \quad (1)$$

3.2. 未読情報の検知

本システムでは記事ごとに、その記事の文章領域中に視点が存在する時間を視線滞在時間 τ_j として計測する。これに加え、予め実験的に求めた一秒あたりに読める文字数 v を用いて、式(2)により未読率 u_j を求める。なお、 len_j は記事 j の文字数である。

これに加え、記事を縦に2分割し、上部分と下部分の既読比率を求める。この分割線より視点が上部にある場合、視点は集合 At_j に属し、下部にある場合、視点は集合 Ab_j に属する。そして、上下の既読比率をそれぞれ rt_j と rb_j とし、式(3)で表す。式中の $|\cdot|$ は集合に含まれる要素数を表す。また、 $rt_j + rb_j = 1$ である。そして、冒頭に偏って閲覧しているかどうかを示す値として、偏り率 r_j を $0 < r_j \leq 1$ に収めるため、式(4)を用いてスムージングする。本研究ではこの r_j を閲覧の偏りとして用いる。

$$u_j = \begin{cases} 1 - v\tau_j/len_j & v\tau_j/len_j \leq 1 \\ 0 & v\tau_j/len_j > 1 \end{cases} \quad (2)$$

$$rt_j = \frac{|At_j|}{|At_j| + |Ab_j|}, \quad rb_j = \frac{|Ab_j|}{|At_j| + |Ab_j|} \quad (3)$$

$$r_j = \frac{rt_j + 2^{-32}}{rb_j + 2^{-32}} \div (2^{32} + 1) \quad (4)$$

3.3. スコア指標

本研究では興味のある記事と似ている記事において未読率の高い興味あり記事が冒頭に偏っ

て閲覧している場合、その記事を再提示する。式(5)は類似度と未読率、そして偏り率を用いて指標を出力する。この指標が大きいほど未読情報を含む可能性が高いと考える。

$$s_j = \text{Sim}_{i,j} u_j r_j \quad (5)$$

4. 検証実験

まず、1秒あたりに読める文字数 v を計測する実験を行う。1秒あたりに読める文字数の算出には式(6)を用いる。 M は記事数であり、今回は20件とした。 τ_{pre} は閲覧時間であり、今回は $\tau_{pre} = 20$ 秒として固定した。

次に、50件の記事をユーザ4人に閲覧してもらった。閲覧中は眼球運動測定装置での測定を行い続けた。記事を閲覧後、興味の度合いを0%から100%まで10%きざみで評価を行ってもらった。ここでの評価はユーザによって評価の尺度が異なると考えられることから、評価を正規化した。また、ユーザの興味のある記事を判別するため、分割点としてパラメータ prm_s を用いる。これは後に精度が高くなるよう推定を行う。

次に、一度閲覧済みの記事50件を再び閲覧してもらった。この時、記事は限無く閲覧してもらい、再び評価をしてもらった。2度目の閲覧時の評価も正規化を行う。加えて、それぞれの記事に目新しい情報があったかどうかを「Yes」、「No」で判定してもらった。そして、式(5)の指標により、ランキング R_i を作成した。このランキング R_i の評価はHijikataら[4]の論文にて示される新規性の検索精度を式(7)のように本研究に合わせ、以下のように一部変更し、評価を行った。集合 C は「評価が $prm_e\%$ 以上で2度目の閲覧後の評価が上昇している記事」であり、かつ「目新しい情報がある(Yes)とした記事」と定義する。 prm_e は prm_s と同様、後に推定を行う。また、再現率を0.0から1.0まで0.1きざみで11段階で変更し、各再現率において補完適合率を算出した。さらに、この補完適合率から、その平均である、11点平均適合率 S_i を求め、新規性の検索精度の評価とした。

$$v = \frac{1}{M} \sum_{j=1}^M \frac{\text{len}_j}{\tau_{pre}} \quad (6)$$

$$\text{適合率} = \frac{|C \cap R_i|}{|R_i|}, \quad \text{再現率} = \frac{|C \cap R_i|}{|C|} \quad (7)$$

5. 実験結果と考察

prm_s と prm_e はそれぞれ0~100の間を10きざみで変え、それぞれの11点平均適合率を求めた。この結果はユーザごとに最大値を出力した箇所を抽出し、これを表1に示す。ユーザ1では

prm_s を90、 prm_e を60に設定した際、高い精度が出力された。特徴的な記事に着目して見ると、 rt_j と rb_j の比率がおおよそ0.9:0.1であり、未読率はおおよそ0.6であった。これはつまり、閲覧時間の9割の記事の上部のみ閲覧し、記事全体の6割を閲覧していないということである。しかし、87.5%という高い評価をつけたことから、記事に対して興味はあるが、最後まで精査して閲覧するに至らなかったと考えられる。

ユーザ2とユーザ3は興味がある記事に対しては未読情報を残さずに閲覧しており、興味のない記事に対しては全てを閲覧していなかった。また、ユーザ4は未読率の平均が0.0であり、記事をはほぼ全て閲覧していたことがわかった。このように未読情報が少ないユーザに対して、本研究の手法では適合率の改善は見られなかった。

表1 ユーザごと最大の11点平均適合率

ユーザ	S_i	prm_s	prm_e
1	0.838	90	60
2	0.292	60	0
3	0.408	60	0
4	0.379	80	0

6. おわりに

本研究ではWebニュース記事を対象として、記事間の類似度と視線情報を用いて、ユーザが記事を読んだ際に残した未読情報を発見し、再提示するシステムを提案した。興味のある記事に対して冒頭や一部分のみの閲覧を行い、全てを閲覧するに至らなかった場合に限り、その記事中に残した未読情報を高い精度で発見し、記事を再提示することができた。

今後は記事中に未読情報が含まれているといった指標ではなく、記事中のどの部分が未読情報であるかの判別も行っていきたい。

参考文献

- [1] R. Atterer, M. Wnuk, and A. Schmidt. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. ACM WWW'06, pp.203-212 (2006).
- [2] 土方 嘉徳, 青木 義則, 古井 陽之助, 中島 周. マウス挙動に基づくテキスト部 分抽出方式と抽出キーワードの有効性に関する検証. 情報処理学会論文誌, Vol.43, No.2, pp.566-576 (2002).
- [3] 高久 雅生, 江草 由佳, 寺井 仁, 齋藤 ひとみ, 三輪 眞木子, 神門 典子. タスク種 別とユーザ特性の違いが Web 情報探索行動に与える影響: 眼球運動データおよび閲覧行動ログを用いた分析. 情報知識学会誌, Vol.20, No.3, pp.249-276 (2010).
- [4] Y. Hijikata, T. Shimizu, and S. Nishida. Discovery-oriented collaborative filtering for improving user satisfaction. IUI'09, pp.67-76 (2009).