F-020

Web 文章における閲覧時間と未読情報に着目した情報推薦

Recommendation of the Web Articles Including an Unread Part Based on User's Browsing Time

折原 レオナルド賢 †　　横井 健 †
Leonardo Ken Orihara　　Takeru Yokoi

1．はじめに

近年インターネットが急速に普及し，多くの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．一方で，個人が全ての興味のある情報を網羅することは困難になっている．また，興味のある情報にたどり着いたとしても，見るべき情報が多いため 1 つ 1 つを精査することが出来ず，有益な情報を見落としてしまう可能性がある．しかし，その見落としには有益な意外性のある情報が含まれている可能性があると考えられるため，それらを推薦することは有益である．

従来，多くの情報推薦手法[1]では，ユーザの嗜好を学習し，類似性の高いものを推薦している．そのため，既知のコンテンツに推薦対象が制限されてしまうことが多い．推薦システムはユーザの関心があるコンテンツを推薦することが目的であるが，関心があることに加えて，推薦結果の目新しさや多様性といった視点にも昨今注目が集まっている[2]．そこで，ユーザがすでに知っているとして読み飛ばした情報に目新しさを含む意外性のある情報が含まれているのではないかと考える．

本研究では類似性に加えて，閲覧時間を用いることで，見落としている情報を含む文章を推薦する手法を提案する．閲覧時間を計測することで，ユーザが文章をどの程度読んだかが特定できると考えられる．また，閲覧時間の短い文章が興味のある文章と類似している場合，これら 2 つの文章が全く同じ文章でない限り，読まれていない情報が存在すると考えられる．この読まれていない情報の中に意外性のある読み落とし情報がある可能性が高いため，このような情報を含む文章を推薦する．

2．提案手法

本研究では，すでに閲覧済みの文章内における見落とし情報を，Web 文章中の読まれていない情報（未読情報）と定義し，それらを含む記事を選別するために類似度と閲覧時間に着目する．提案するシステムでは，あらかじめ閲覧済みの文章をユーザに評価してもらうことで，興味のある文章のプロファイルを作成し，システムに収集させておく．この興味のある文章と似た文章を短時間で読み終えていた場合，そこには意外性のある見落とし情報が存在すると考え，その興味のある文章に似ている文章を文章間の内容の類似度と未読率から算出するスコアに基づき 1 件推薦する．

2.1　文章の類似度

本研究における類似度とは，文章同士の内容がどの程度似ているかを測る指標とし，その指標として cos 類似度[3]を用いる．cos 類似度を求めるためには，各文章をベクトル空間法[4]を用いてベクトル化する必要がある．そのために，まず文章を n-gram インデキシングで分割する．n-gram インデキシングとは，文字列の先頭から 1 文字ずつずらしながらn文字単位で区切る手法のことである．ベクトル化された文章$\mathbf{d}_i$は式(1)のように表現される．

$$\mathbf{d}_i = \{w_{i1}, \cdots, w_{iV}\} \quad \cdots (1)$$

なお，w_{iV}は 4-gram で分割した文章$\mathbf{d}_i$に含まれる単語の出現頻度，また，Vは文章集合中の 4-gram で分割した全文字列数である．

本システムでは，ユーザが興味のある文章を$\mathbf{d}_i$，それ以外の文章を$\mathbf{d}_j$とし，$\mathbf{d}_i$の中から推薦を行う．cos 類似度はベクトル化した文章$\mathbf{d}_i$，$\mathbf{d}_j$に対して，式（2）を用いて算出する．

$$Sim_{ij} = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i||\mathbf{d}_j|} \quad \cdots (2)$$

$|\cdot|$は L2 ノルムとし，推薦を行う際は$i \neq j$とする．また，式（2）を用いる時，文章$\mathbf{d}_i$，$\mathbf{d}_j$に共起する単語がない場合は互いに存在しない単語の出現頻度を 0 としてベクトルに挿入することで，文章$\mathbf{d}_i$，$\mathbf{d}_j$の次元数を一致させる．

2.2　未読率の計算

本研究では見落としている情報を含む記事を推薦するため，文章中における未読情報の量を判別する．本システムでは，この未読情報の量を未読率という割合で求める．未読率は文章の文字量にかかわらず，1 つの文章全体のうち何割読むことができなかったかを表す値である．これを求めるためには各ユーザ個人の 1 秒あたりに読める文字数Cを求める必要がある．この 1 秒あたりに読める文字数Cは予め式（3）で求めておく．

$$C = \frac{1}{n}\sum_{j=1}^{n}\frac{L_j}{T_j} \quad \cdots (3)$$

式（3）におけるnは文章の総数，L_iは文章$\mathbf{d}_i$の文字数，T_iは$\mathbf{d}_i$の閲覧時間をそれぞれ表す．

本システムでは各文章の閲覧時間を計測し，一秒あたりに読める文字数Cとの積によって既読文字数を求め，この既読文字数を全体の文字数に対する割合である既読率として算出する．なお，既読率が 1 を超える場合は 1 とする．この既読率を用いて，式（4）により未読率U_iを求める．

$$U_j = 1 - \frac{T_j C}{L_j} \quad \cdots (4)$$

この未読率U_iは 0 から 1 の値を取り，値が大きいほど未読情報の量が多いことを表す．

2.3　推薦指標

本研究では，あまり読まれていない記事が他の興味の高い記事と似ている場合，その記事を推薦するため，式（5）に従い，式（2）で求まった類似度Sim_{ij}と式（4）で求まった未読率U_iの積を最大とする文章rを推薦する．

$$r = \arg\max_j (Sim_{ij} U_j) \quad \cdots (5)$$

† 東京都立産業技術高等専門学校

本システムでは，興味のある文章 1 件に対して，類似度Sim_{ii}と未読率U_iの積が最大となる文章 1 件を推薦するため，興味のある文章を 1 件のみ推薦を行う．

3. 実験

3.1 予備実験

提案手法の有効性の検証実験を行うにあたって，個人の閲覧時間に関する予備実験を 3 つ行った．本提案手法では閲覧時間から未読率を算出するとき，1 秒あたりに読める文字数が必要となる．そのため，1 つめの予備実験では，1 秒あたりに読める文字数を計測し，これが固定することができるのかを検証する．これは Web 上から記事を無作為に 20 件用意し，それを一人で読み，式（3）から 1 秒あたりの読める文字数の平均Cと標準偏差を算出した．今後の予備実験においてはこの値を基準値として実験を行う．

2 つめの予備実験では，1 秒あたりに読める文字数を固定することができるかの検証を行うため，既読文章と未読文章では閲覧時間が変動するのかを調べた．また，文章を構成する文字種（漢字・英字・ひらがな・カタカナ）の含有率の違いによって閲覧時間は変動するのかも調べた．40 件の文書の閲覧時間を 10 秒で固定して何文字読めたかの測定を行った．41 件の文章の内訳は既読の記事 21 件，未読の記事 20 件とし，読む速度が何%増加したかで評価を行った．ここでは，記事の順序が読む速度に影響されないようランダムで記事を表示させた．図 1 が実験に用いたシステムである．

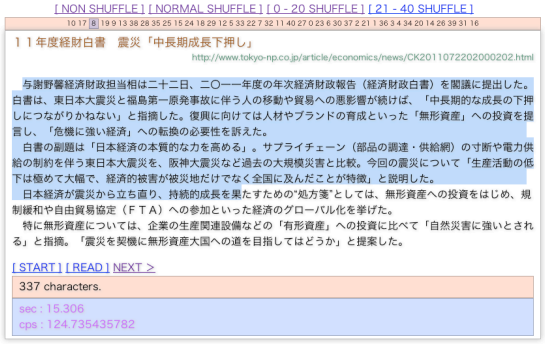

図 1 閲覧時間の測定システム

図 1 のシステムではページを読み込むと同時に JavaScript のタイマーが作動し，予め設定された時間（10 秒）になると文字を白地にすることで視覚的に読めないようにする．その後，読み終えた部分をマウスでドラッグすることで，システムが自動的に文字数をカウントし，閲覧時間から 1 秒あたりに読める文字数の基準値との差を算出する．システムは自動で始まるが，もう一度処理をはじめから行いたい場合，[START]を押すことで手動でやり直すことができる．また，文字数のカウントは範囲を指定した時点で始まるが、範囲を再指定した場合，[READ]を押して手動で再度文字数をカウントすることができる．このシステムの処理は JavaScript で行い，算出結果は Ajax でデータベースに挿入する．画面上部の[NON SHUFFLE]を選択すると記事番号 0 から 40 までの記事を順番に表示させるようにし，[NORMAL SHUFFLE]では記事番号 0 から 40 の記事をランダムで表示させるようにできる．また，[0 - 20 SHUFFLE]では記事番号 0 から 20 を，[21 - 40 SHUFFLE]では記事番号 21 から 40 の記事をランダムで表示させることができるようになっている．

3 つめの予備実験では，ユーザの興味の度合いを閲覧時間から推定することができるのかを調べるため，ユーザの興味の度合いを閲覧時間の関係を調べた．まずはじめに，ユーザ 10 人の Web 記事の閲覧時間と興味の度合いの相関関係を調べた．ここでは，100 件の Web ページを閲覧し，その興味の度合いを 0%から 100%まで 10%きざみで評価を行ってもらった．次に，記事$\mathbf{d}_i$の閲覧時間T_iを式（6）を用いて記事内に含まれる文字種の含有率に重みを付けてスコアS_iを算出した．これは，記事を構成する文字種によって閲覧時間が変わると仮定し，元の閲覧時間との差を検証するためである．式（6）におけるEn_iは記事$\mathbf{d}_i$の英字の含有率を，Ka_iは記事$\mathbf{d}_i$のひらがな・カタカナの含有率を，Kn_iは記事$\mathbf{d}_i$の漢字の含有率をそれぞれ表す．ここでは，個人で 42 件の Web ページを閲覧し，その興味の度合いを 0%から 100%まで 10%きざみで評価を行った．

$$S_j = T_j(En_j + Ka_j - Kn_j) \quad \cdots (6)$$

3.2 予備実験の結果と考察

Web 上の記事 20 件の閲覧時間を測定した結果が表 1 である．

表 1 20 件の記事の平均と標準偏差

平均 [chars/sec]	標準偏差 [chars/sec]
8.149	2.264

この結果から，1 秒あたりに読める文字数Cは標準偏差が小さいことからこれを基準値としても問題ないと考えられる．

1 つめの予備実験で計測した 1 秒あたりに読める文字数Cを基準値と仮定し，既読の記事 20 件，未読の記事 20 件を読み，読む速度の変化を調べた．その結果を表 2 に示す．

表 2 既読記事の速度変化

	時間	平均増加率[%]
既読	10sec	7.1
未読		−1.8

表中の平均増加率とは，1 秒あたりの読める文字数Cを理想とした時，どの程度読む速度が増加したかを示す数値であり，プラスの場合は読む速度が早くなり，マイナスの場合は読む速度が遅くなったことを表す．表 2 は既読記事と未読記事を再び読み，その速度を計測した結果である．既読の文章では平均増加率が 7.1%上昇し，読む速度が上がったことがわかる．また，未読記事では平均増加率が-1.8%となり，読む速度が下降している事がわかる．このことから，1 秒あたりに読める速度（文字数）は未読記事と既読記事によって変化することがわかった．

また，読む速度が早い Web 文章と遅い Web 文章について調べたところ，以下の表 3 に示すように含まれる文字種に傾向が確認できた．

表 3 文字属性別の速度比較

読む速度	早い	遅い
多い文字種	ひらがな、英語	漢字

ひらがなが多く含まれる文章は読む速度が早く，逆に漢字が多い文章では読む速度が遅くなっている傾向が見られた．また，英字が多く含まれている文章は読む速度が極端に上昇している傾向があった．文章を読む際，日本語にお

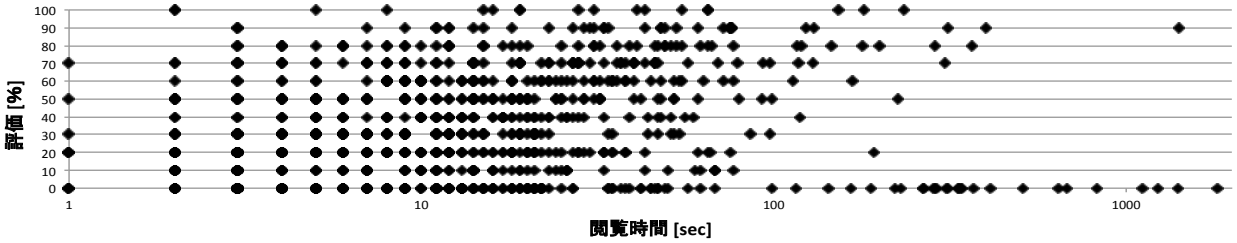

図 2 10 人の閲覧時間と興味のグラフ

ける 1 文字と英字における 1 文字の扱いが異なるため，英字を多く含む場合，文字数に比べて閲覧時間が短くなる傾向が見られた．このため，文章の属性によっても読む速度が異なることがわかった．

3 つめの予備実験では，閲覧時間と興味の度合いの関係を調べた．100 件の閲覧時間を 10 人分計測し，合計で 1,000 件の閲覧データが得られた．この結果を図 2 に示す．

このグラフの相関係数は-0.01 となり，閲覧時間と興味の間には相関関係が無いことがわかった．

また，1,000 件の閲覧データをヒストグラムにした結果が図 3 である．

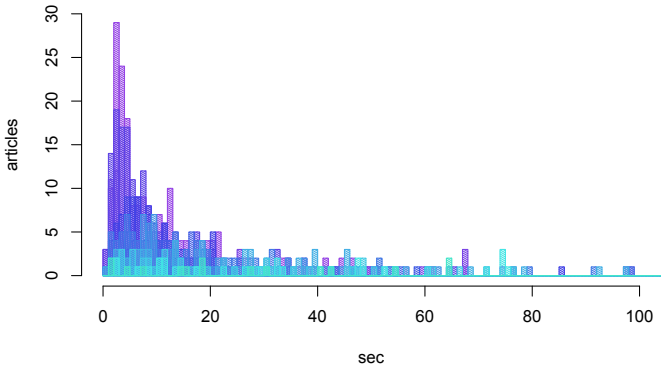

図 3 時間のヒストグラム

図 3 は横軸を閲覧時間[sec]，縦軸を記事数としたヒストグラムである．この結果から，閲覧した記事のほとんどにおいて，閲覧時間が短いことがわかる．図 3 のグラフでは閲覧時間が 100 秒以上のデータは表示していないが，全データのうち，100 秒以上閲覧されている記事数は全体の 5%程しかないことから外れ値とした．

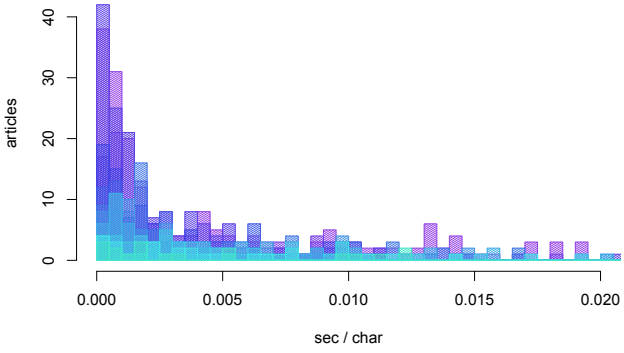

図 4 文字数で正規化したヒストグラム

図 3 の結果を受けて，記事の文字数が少ない場合，閲覧時間は短くなるため，図 4 では閲覧時間を各記事の文字数で割り，正規化を行った．図 4 では，横軸が 1 文字あたりの閲覧時間[sec/char]，縦軸を記事数としヒストグラムである．この結果（図 4）から，最頻値が 0 に収束していることが見て取れる．このため，閲覧時間が短いことは文字数が少ないためでは無いことが言える．

次に，個人で閲覧した 42 件の閲覧データを評価別にグラフ化した．その結果が図 5 である．図 5 では x 軸を閲覧時間[sec]とし，対数表示にしている．y 軸は興味の度合いの評価[%]である．文字数で正規化した閲覧時間の相関係数は 0.51 だった．このため閲覧時間と評価の間には相関関係が存在しないと考えられる．

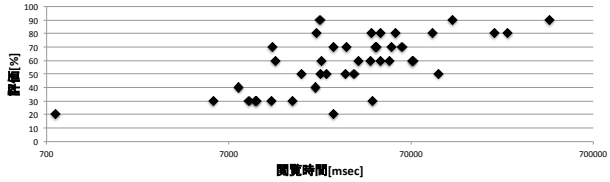

図 5 閲覧時間と評価のグラフ

また，記事$\mathbf{d}_i$の閲覧時間T_iを式（6）を用いて記事内に含まれる文字種に重みを付けてスコアS_iを算出し，そのスコアと興味の度合いの評価の関係を図 6 のように示した．

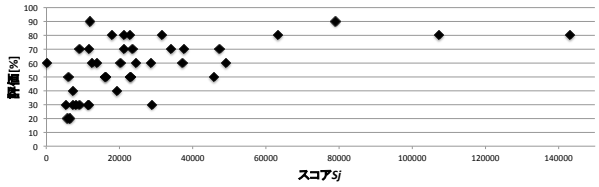

図 6 スコアS_iと評価のグラフ

図 6 では x 軸をスコアS_iとし，対数表示にしている．y 軸は図 5 と同じく評価[%]である．このグラフにおける相関係数は閲覧時間と評価の相関係数と同じく 0.51 だった．このため閲覧時間と評価の間には相関関係が存在しないと考えられる．

なお，この結果から閲覧時間にばらつきはあるが，記事を読む際にこのばらつきは考慮する必要がないと考えられる．そのため，提案手法の検証実験において 1 秒あたりに読める文字数は固定することは妥当であると考えられる．

3.3 提案手法の検証実験

実験では，Web 上から無作為に取得した 100 件の Web ニュース記事を用いた．それらの記事に対して，10 人のユーザの閲覧時間を計測すると共に，各ユーザが興味の度合いを 0%から 100%まで 10%きざみで評価した．そのうち 80%以上の評価がついた記事を推薦時に用いる類似度の基準となる興味あり記事として設定した．これらの記事それぞれに対して，類似度が高く，閲覧時間が短い記事を 1 件推薦した．その後，推薦した記事に対して 10 人のユーザ

が再び興味の度合いを 0%から 100%でまで 10%きざみで再評価を行った．

3.4 実験結果と考察

検証実験の結果を表 4 に示す．表 4 では興味のある記事のうち，推薦前の評価が 80%未満だった記事と 80%以上だった記事に分類し，それぞれの興味の度合いの遷移を調べた．「 - 」はこれらのデータが存在しないことを示す．まず，推薦前の評価が 80%未満の記事集合における未読率が高い記事，全 47 件の推薦前と推薦後のそれぞれの評価の平均を求めたところ，推薦後の方が推薦前より評価が上がっていることがわかった．また，t 検定を行った結果，有意水準 1%で有意差があった．このため，興味がなく未読率が高い記事を推薦すると評価が上がることがわかった．この傾向は各ユーザ個人においても確認できた．

次に，評価が 80%以上の記事集合における未読率が高い記事，全 13 件の推薦前と推薦後のそれぞれの評価の平均である．このデータからはもともと評価が高かった記事を再度推薦しても意外性はなく，評価は下がる傾向が見られた．また t 検定を行った結果，有意水準 5%で有意差があった．

このため，評価が高く未読率が高い記事を推薦すると評価が下がることがわかった．この傾向は各ユーザ個人においても確認できた．一方，表 4 におけるユーザ 3 では，推薦前が 80%未満の記事で推薦後の平均が下がっている傾向が見られた．このようなユーザは本手法でプロファイリングをすることができないと考えられる．

以上の結果を受けて，閲覧時間の短い記事を推薦すると，記事の評価が上がることから，一般的に文章の一部やタイトルだけを軽く見渡しただけで評価をつけてしまう傾向があると考えられる．

表 4 検証実験結果

		推薦前が80%未満の記事			推薦前が80%以上の記事		
ユーザ	興味のある記事数	推薦された記事数	推薦前平均	推薦後平均	推薦された記事数	推薦前平均	推薦後平均
1	9	7	43	57	2	80	75
2	12	7	33	61	5	86	58
3	4	3	40	20	1	80	70
4	10	10	38	60	0	-	-
5	4	4	25	75	0	-	-
6	8	5	20	20	3	100	87
7	1	1	0	80	0	-	-
8	12	10	32	42	2	80	35
9	0	0	-	-	0	-	-
10	0	0	-	-	0	-	-
総数	60	47	-	-	13	-	-
平均	-	-	29	52	-	85	65

4. まとめ

本研究では情報閲覧時間を用いてユーザのまだ知らない情報を発見し，推薦するシステムを提案した．特に，閲覧時間を用いることでユーザの未読率を計測し，その未読率を用いて情報推薦を行った．この結果，興味の度合いは低いが興味の度合いが高い記事と似ている記事を推薦すると，興味の度合いが 23%上昇した．また，もともと興味の度合いが高く，未読率も高い記事を推薦すると，興味の度合いが 20%下降した．この結果から，一般的には軽く記事を見渡しただけで評価をしてしまう傾向が有ることが観測できた．

本研究では閲覧時間のみに着目して推薦を行なっているため，文章のどの部分を閲覧しているのか判定することができない．このため，文章量に対して閲覧時間が極端に長い場合や，極端に短い場合は本提案手法では良い結果が得られない．閲覧時間以外の着目点としては，ユーザの視点情報を識別するインターフェースを用いて，視点に着目し，どの部分を読んでいるかを解析することで，ユーザの興味の判別精度を向上させる方法を考えている．また，ブラウザのスクロール位置から Web 文章中の既読箇所・未読箇所の判定を取り入れることで，より詳細に情報の未読を測定し，推薦精度の向上を目指す．

参考文献

[1]土方 嘉徳，“嗜好抽出と情報推薦技術”， 情報処理学会論文誌，Vol.48，No.9，pp.957-965，2007.

[2] 野田陽平，清田陽司，中川裕志，“Wikipedia カテゴリネットワークからの意外性のある関係性の抽出”，人工知能学会研究会資料，SIG-SWO-A901-04，pp.1-4，2009.

[3] 徳永 健伸，“情報検索と言語処理”，東京大学出版会，第 4 版，2006，p.122.

[4] G Salton，“A vector space model for automatic indexing”，Communications of the ACM，Vol.18，pp.613-620，1975.