テキストの重要箇所推定のための読み手のモデル

岡崎 直観* 松尾 豊† 石塚 満*
東京大学大学院情報理工学系研究科

〒 113-8656 東京都文京区本郷 7-3-1
Tel: 03-5841-6755
Fax: 03-5841-8570
e-mail: okazaki@miv.t.u-tokyo.ac.jp
URL: http://www.miv.t.u-tokyo.ac.jp/~okazaki/

1 はじめに

テキスト自動要約の根幹を担う要素技術として，適当に切り出したテキスト断片の重要度を計算する重要箇所抽出法がある．今までの重要箇所抽出の多くは，語の頻度などの尺度を元にして統計的に優位なテキスト断片を割り出し，それらに高い（もしくは低い）スコアを与えるというものであった [1]．これは，著者が苦労して物事を説明したり論を展開するときに，特定の語を繰り返し用いるという現象が観察されるため，文章中の語の出現頻度を重要度の基準として重み付けを行えば，計算機が文章の内容を理解しなくても重要箇所を推定できるという考えに基づいている [2].

既存手法の大部分は，書き手の言わんとすることを抽出することが目的であり,「要約を読んでいる人が過去にどのような文章を読んできたのか」ということに関して考慮していない．人間の目は二つあっても，複数の文書を同時に読むことはできないので，ある文書を読み，そこで得られた知識と興味を以って次の文書に臨むはずである．計算機が作成した要約は読者が直接読むのであって，読む人がどのような人であり，与えられたテキストをどのように処理するのかを要約というタスクの中に組み込むことが望ましい．このように要約タスクをユーザに応じてパーソナライズすることは，古くからテキスト自動要約の課題とされている [3]．本発表は，情報の読者の視点に立ったパーソナライズを行うモデルと，そのモデルを用いた重要文抽出について提案するものである．

2 文章の読み手モデルの構築

人間は情報・知識を得るために文章を読み，そこで得られた情報・知識を所有し，時にそこで得られた情報が新たな興味をもたらす.「テキスト自動要約」についての論文を読んでいると,「頻度」や「リード文」,「修辞構造」,「機械学習」など，新たな考え方・手法と出会い，さらにそれらの手法を詳細に紹介している論文を漁り，各々の理解を深めていく．一方で，これら「テキスト自動要約」の論文の序論部はたいてい現状の説明や問題提起であり，どれもほとんど同じような内容である．後から読む論文では序論を読み飛ばし，過去に読んだ論文との差や，その論文独自の箇所を重点的に読む．このように，人間はある文書から情報を得たとき，さらに別の情報を渇望したり，

*東京大学大学院 情報理工学系研究科 電子情報学専攻
†産業技術総合研究所 サイバーアシスト研究センター

いったん獲得した情報を無視するなど，情報に対する嗜好の傾向が変化する．このことに着目して，本研究では人間の情報に対する嗜好性を以下のようにモデル化する．

(1) 今までに受け取った 情報 と関連のある 情報 を嗜好する．

(2) いったん受け取った 情報 は記憶され，その 情報 に対する嗜好は鈍る．

(3) 情報 に対する嗜好性は時間と共に薄らいでいく．

これでは「情報」をどのように計測するのか曖昧であるが,「情報」はタスクに応じて計測方法が異なるので，これについては次節において定義する．

このモデルを定式化するために，すべての情報の数を n とし，時間と共に嗜好性が変化するので時刻 t を導入する．情報 c_i（ただし $(0 \leq i < n)$）に対する読み手の嗜好性の大きさを $\mathrm{A}^{(t)}(c_i), (0 \leq \mathrm{A}^{(t)}(c_i) \leq 1)$ で表す．初期状態（時刻 $t = 0$）においては，すべての情報に対する嗜好性は等しいと考えて，嗜好性 $\mathrm{A}^{(0)}(c_i) = 0.5$ とする．次に，ユーザが時刻 t において情報を得て，時刻 $t+1$ に遷移する場合，嗜好性 $\mathrm{A}^{(t+1)}(c_i)$ は以下のように計算する．

$$\mathrm{A}^{(t+1)}(c_i) = (1-\alpha)\left\{0.5 + \frac{\mathrm{A}^{(t)}(c_i) - 0.5}{2}\right\} + \alpha\left\{\mathrm{P}^{(t)}(c_i) - \mathrm{O}^{(t)}(c_i)\right\} \tag{1}$$

ここで，$\mathrm{P}^{(t)}(c_i)$ は時刻 t において得た情報群の中で c_i と関連があるものの量，$\mathrm{O}^{(t)}(c_i)$ は時刻 t において情報 c_i が直接的に伝達された量である．α は $0 \leq \alpha \leq 1$ を満たすパラメータである．

つまり式 1 は，時刻 $t \rightarrow (t+1)$ への遷移において，$(1-\alpha)$ で与えられる減衰定数を用いて情報に対する嗜好性を鈍化させると同時に，時刻 t で受け取った情報の中に，情報 c_i と関連するものがあれば情報 c_i に対する嗜好性を増加させ，情報 c_i そのものが存在すれば嗜好性を減少させるものである．このままでは，時間 t の定義や，$\mathrm{P}^{(t)}(c_i)$ や $\mathrm{O}^{(t)}(c_i)$ の定義が抽象的でよく分からないが，重要文抽出タスクにおける定義について次節で述べる．

3 重要文抽出への応用

前節での考え方を重要文抽出に向けて具体化する．本発表では，ユーザが関連する複数の文書を一つずつ閲覧する代わりに，重要文抽出で作成される要約を一つずつ読んでいくというタスクを想定する．文書を一つ読む毎に前節で説明した時間 t は増加すると考える．

「情報」という言葉を何となく使ってきたが，本発表では語と語のペアを「情報」と考える．語と語のペアの作り方はいろいろあるが，本発表では文を単位とした共起関係[1]を用いる．重要文抽出は，文書中に含まれる共起関係 c_i を重み付けし，決められた文字数の中でその重み和が最大になるように文を選ぶ．この要約問題は，コスト和を最小にしながら，それぞれの文を要約に含めるべきか，含めないべきかという組み合わせ最適化問題で記述でき，第 3 回 MYCOM で筆者が発表した重要文抽出法 [4, 5] をそのまま適用できる．

共起関係 c_i の重み $W(c_i)$ は，前節で説明した読者の嗜好 $A(c_i)$ と，従来どおり統計的情報から求めた重み $D(c_i)$ の平均を用いて以下のように定義する．

$$W^{(t)}(c_i) = \frac{D^{(t)}(c_i) + A^{(t)}(c_i)}{2} \tag{2}$$

重み $D(c_i)$ はテキスト現象から観察される重みであり，時刻 t において，

$$D^{(t)}(c_i) = c_i \text{ の共起回数} \tag{3}$$

と定義する．式 1 における $\mathrm{O}^{(t)}(c_i)$ と $\mathrm{P}^{(t)}(c_i)$ の定義は，

$$\mathrm{O}^{(t)}(c_i) = \begin{cases} \mathrm{W}^{(t)}(c_i) & \text{時刻 } t \text{ で読んだ文書に共起関係 } c_i \text{ が含まれていた場合} \\ 0 & \text{時刻 } t \text{ で読んだ文書に共起関係 } c_i \text{ が含まれていない場合} \end{cases} \tag{4}$$

[1]異なる 2 つの語がある一つの文の中で共に出現しているとき，この 2 つの語には共起関係があると考える

$\mathrm{P}^{(t)}(c_i)$ = 時刻 t において共起関係 c_i を構成する 2 つの語が作るすべての共起関係の重みの平均 (5)

である．

4　結果と考察

提案手法を用いて，複数の新聞記事からユーザの嗜好性がどのように変わるのか，実験をした．実験に利用した記事は毎日新聞 98 年–99 年中に含まれる「２０００年問題の対応」に関する 10 記事である．記事の日付順にシステムに記事を渡して要約を作成してもらい，その過程でユーザの嗜好性を計算する．元の記事の内容をすべて載せることはできないので，それぞれの記事の見出しを表 1 に掲載する．

1	2000 年問題対応　大阪のホテルがコンピュータ関係会社などに宿泊プラン売り込み
2	2000 年問題診断、ソフト提供を開始–マイクロソフト
3	2000 年問題　未対応船を拘留も–英海上沿岸警備庁が方針
4	ロス在住の男性、ＮＥＣ米子会社を提訴　購入のパソコン「2000 年問題に非対応」
5	ＪＲ西日本、列車を止めず　年始対応、東日本と分かれる–2000 年問題
6	「カーナビ版 2000 年問題」旧型機種、きょうＸデー–休日返上、4 社が補修対応
7	問い合わせ続々、事故報告なし–カーナビ誤作動「2000 年問題」
8	問い合わせ 2200 件–カーナビ誤作動「2000 年問題」
9	「まだ不安」4.5 %も–金融機関の 2000 年問題、日銀が助言へ
10	2000 年元日、ＪＲ四国は運行

表 1: 記事番号と記事の見出し

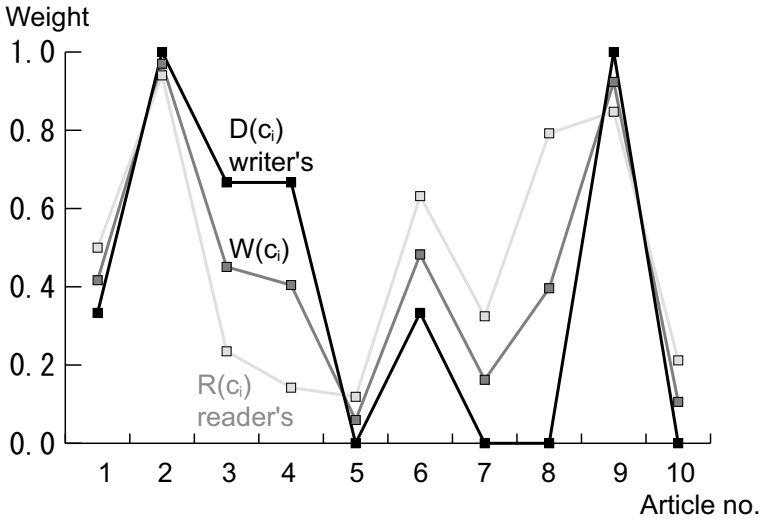

図 1: 「問題–対応」という単語ペアに対する $D(c_i), R(c_i), W(c_i)$

実際には重要文抽出が行われ，要約が作成されているのであるが，紙面の都合もあるので，ここでは読み手の嗜好性がどの程度妥当なのか調べるために，図 1 に「問題–対応」という単語ペアに着目した $D(c_i), R(c_i), W(c_i)$ の時間遷移を示す．

読み手の嗜好性 $R(c_i)$ に着目すると，記事 1 から 2 に遷移するときに嗜好性が高くなっている．

これは，記事 1 の要約に「問題」や「対応」という語が含まれたが，記事 1 で作成された要約には「問題–対応」という共起関係が存在しなかったことを示している．このため，記事 2 における嗜好性は上昇したが，記事 2 には「問題–対応」という共起関係が多く含まれており（$D(c_i)$ 参照），記事 2 の要約文中にこの共起関係が選ばれた．したがって，記事 3 においては，読者に伝達した情報への嗜好の鈍化が起こり，嗜好性が激減している．

ここに挙げた例は比較的納得のいく例であり，読者に伝達した情報への嗜好の鈍化をさせる部分に関しては，過去に含んだ情報を除外するという単純なメカニズムのため，比較的うまくいっている．しかしながら，読者の興味の推定に関しては，要約に含まれた語とたまたま共起してしまった語にも高いスコアを与えてしまう現象が見受けられ，こちらの精度に関しては改良の余地が残されていると思われる．

5　今後の課題

本発表では，文書を読んでいるユーザの嗜好をモデル化することで，ユーザーにとって有益な情報を選択するための枠組みを提案した．重要文抽出に的を絞って簡単な評価実験を行ったが，この実験をさらに進めて本枠組みの有効性について検証するとともに，ユーザの過去の文書閲覧履歴からユーザの興味があると思われる情報を推定して提示するシステムなど，別のタスクへの応用も検討していきたい．

謝辞

我々は国立情報学研究所情報学資源研究センターの支援により開催されているワークショップ NTCIR-3, NTCIR-4 のテキスト自動要約タスク TSC-2, TSC-3 に参加し，本研究にあたっては毎日新聞記事データ，要約課題データを利用させていただきました．

参考文献

[1] Salton, G. *Automatic Text Processing*. Addison-Wesley, 1989.

[2] Luhn, H. P. The automatic creation of literature abstracts. *IBM journal of Research and Development*, Vol. 2, No. 2, pp. 159–165, 1958.

[3] 奥村 学，難波 英嗣．文書自動要約に関する研究動向．自然言語処理,「テキスト要約のための言語処理」特集号，Vol.6, No.6, pp.1-26, 1999

[4] Naoaki Okazaki, Yutaka Matsuo, Naohiro Matsumura, Hironori Tomobe, and Mitsuru Ishizuka: Two Different Methods at NTCIR3-TSC2: Coverage Oriented and Focus Oriented, Working Notes of the Third NTCIR Workshop Meeting, Part V: Text Summarization Challenge2 (TSC2), pp.39–46, 2002.

[5] 岡崎直観，松尾豊，石塚満．関連する複数新聞記事からの重要文抽出法．第 3 回 AI 若手の集い MYCOM2002，pp.80–86, 2002.