4M-9

Wikipediaのノートページにおける編集者の重要度算出手法

近藤 弘隆 †　　鈴木 優 ††　　石川 佳治 ††,‡
† 名古屋大学工学部電気電子・情報工学科　†† 名古屋大学大学院情報科学研究科
‡ 国立情報学研究所

1　はじめに

Wikipediaは誰でも編集が可能なインターネット上の百科事典である．複数の異なる意見を持つ編集者が対立した場合，編集者は議論や対話を行い，記事を編集していかなければならない．このとき，支持されている編集者が誰であるか分かれば，対立の内容を理解でき，議論や対話が容易になる．そこで，支持されている編集者が誰であるかを特定する必要がある．

そこで本研究では，Wikipediaのノートページにおける編集者の重要度を算出する手法を提案する．多くの支持を得ている重要な編集者がノートページで発言するとき，本文にその発言を反映した記述が行われ，その記述は他の編集者によって削除されないと考えられる．この仮定から，ノートページにおける発言とその発言の本文への反映度合いを測定することによって，編集者の支持されている度合い，つまり編集者の重要度を測定することができると考えた．

2　関連研究

Web上の会話を評価する研究としては，緒方らによる研究 [1] が存在する．この手法では会話内容から，人物の内面的特徴を評価している．

また，Web上のテキストの重要度に関する研究として，藤井の研究 [2] が存在する．藤井は論点の固有度と重要度を使い，意見の可視化を行うシステムを提案している．

これらの研究ではtf-idfを用いて単語のスコア付けを行っている点，Webテキストから重要度を求めている点が本研究と類似している．しかし，本稿では，Webテキストの論点の重要度ではなく，ノートページにおける編集者の重要度を求めている．また，tf-idfや割合を用いるのではなく，評価の対象がWikipediaの編集者であることを考慮しtf-idfを改良した，意見が反映されているかという情報を用いて算出している．

3　編集者の重要度算出手法

この章では編集者の重要度算出手法について述べる．

編集者がノートページへ意見を書き込む際は，本文の編集と同様に，ノートページを編集することで行う．そのため，本文の編集行為と区別するために，ノートページに意見を書き込むことを**発言**と表記する．

ある記事における編集者の重要度を以下の手法で算出する．まず，重要度を求めたい編集者のある記事での発言を全て取得する．発言に対応する本文の編集行為を求める．発言内容が重要であるとき，つまり他の編集者に支持された発言内容であるとき，発言内容が本文に反映されており，その反映された本文は他の編集者から削除されないと考えられる．よって，編集内容が残っていれば，その編集の基になった発言は支持されているといえる．したがって，重要度を求めるためには，発言を反映させる編集の編集内容が残っているか求める必要がある．そのため，発言に対応する編集を元に戻す編集行為を取得する．また，発言に対応する編集内容がその編集行為だけで扱われているような編集であった場合，よく編集される内容の編集であった場合と比較して，その編集の方が意見を反映させていると考えられる．よって，発言に対応する編集によって追加，削除された単語が特徴的か，残っているかという情報を用いて，発言の重みを求め，編集者の重要度を算出する．

3.1　発言に対応する編集の取得

発言に対応する編集行為の取得手法を説明する．

ある発言に対応する本文の追加や削除は一回であるとは限らず，複数回存在する可能性がある．ある編集者がノートページで発言した後に，期間を空けず，その編集者が本文を編集した場合，その編集内容は発言内容に関するものであると思われる．編集者は発言内容が同意されるのを待ち，編集を行うと思われる．

そのため，発言に対応する本文の編集を求めるために，発言が行われた時刻から T 時間後，もしくはその発言が行った編集者が次に発言を行うまでの期間に本文で行われた編集を，その発言に付随する編集であると定める．

3.2　打ち消す編集の取得

発言に付随する編集を元に戻すような編集行為を取得する手法を本節で説明する．

編集行為には内容を追加する行為と削除する行為が

A Method for Assessing Importance of Editors in Wikipedia using Discussion in Talk Page
Hirotaka Kondo †, Yu Suzuki ††, Yoshiharu Ishikawa ††,‡
† Department of Information Engineering, Nagoya University
†† Graduate School of Information Science, Nagoya University
‡ National Institute of Informatics

存在する．編集によって内容を追加した場合について考える．ある編集行為における追加された内容が元に戻される際，追加された内容と同様の内容を削除する編集行為が行われる．よって，発言に付随する編集ごとに，発言による追加に一番似た削除を一定の期間内に行っている編集を tf-idf を用いて求め，それを打ち消す編集とする．削除についても同様のことが言える．

打ち消す編集を求める具体的な手法を説明する．一度の編集行為を一つの文書とみなす．まず，記事に対する全ての編集において追加された単語 i の逆文書頻度 idf_i^a，削除された単語の逆文書頻度 idf_i^d を次のように求める．

$$idf_i^a = \log_{|E_a|} \frac{|E_a|}{df_i^a}, \qquad idf_i^d = \log_{|E^d|} \frac{|E_d|}{df_i^d} \tag{1}$$

$|E_a|$ は内容の追加があった編集行為の数，df_i^a は単語 i が追加されている編集行為の数である．$|E_d|$ は内容の削除があった編集行為の数，df_i^d は単語 i が追加された編集行為の数である．この idf_i^a，idf_i^d は記事で共通のものである．

ある発言 j に対応する編集で追加された単語 i の出現頻度 $tf_{i,j}^a$，編集で削除された単語 i の出現頻度 $tf_{i,j}^d$ を以下の式で求める．

$$tf_{i,j}^a = \frac{n_{i,j}^a}{\sum_k n_{k,j}^a}, \qquad tf_{i,j}^d = \frac{n_{i,j}^d}{\sum_k n_{k,j}^d} \tag{2}$$

発言 j における単語 i の追加，削除された回数をそれぞれ $n_{i,j}^a$，$n_{i,j}^d$ としている．

発言 j における追加された単語 i の重みを $tf_{i,j}^a \times idf_i^a$，削除された単語 i の重みを $tf_{i,j}^d \times idf_i^d$ とする．追加，削除した各単語の重みをベクトルの要素とした特徴ベクトル $\boldsymbol{v}_a$，$\boldsymbol{v}_d$ をそれぞれ作成する．

ある編集行為における内容を元に戻すような編集行為は元の編集から一定の時間内に行われると考えられる．そこで，発言に付随するある編集から，次のその発言者の編集まで，もしくは期間 P までになされた編集を取得する．取得した編集 e ごとに，追加，削除された単語 i の出現頻度 $tf_{i,e}^{a'}$，$tf_{i,e}^{d'}$ を求める．編集 e ごとに各単語の追加，削除それぞれの出現頻度を要素とした追加と削除の特徴ベクトル $\boldsymbol{v}'_a$，$\boldsymbol{v}'_d$ を作成する．そして，$\boldsymbol{v}_a \cdot \boldsymbol{v}'_d$，$\boldsymbol{v}_d \cdot \boldsymbol{v}'_a$ を求める．この値は，編集 e が発言に付随する編集の編集内容を元に戻す内容であるほど高くなる．一番値の大きい編集を求めることで，追加，削除それぞれの打ち消す編集を取得できる．

3.3 編集者の重要度算出

3.1 節で求めた発言に対応する編集，3.2 節で求めた打ち消す編集を利用して，発言の重要度を算出する．逆文書頻度を用いて，意見を反映させる単語の重みを大きくできる．打ち消す編集を用いて，編集の反映度合いがわかる．これらを用いて重要度を算出する．

発言 j における追加された単語 i の重み $t_{i,j}^a$，削除された単語 i の重み $t_{i,j}^d$ を以下の数式で算出する．

$$t_{i,j}^a = \frac{n_{i,j}^a - 2n_{i,j}^{d'}}{\sum_k n_{k,j}^a} \times idf_i^a, \; t_{i,j}^d = \frac{n_{i,j}^d - 2n_{i,j}^{a'}}{\sum_k n_{k,j}^d} \times idf_i^d \tag{3}$$

$n_{i,j}^a$，$n_{i,j}^d$ はそれぞれ発言 j における単語 i の追加数，削除数である．$n_{i,j}^{a'}$，$n_{i,j}^{d'}$ はそれぞれ打ち消す編集での単語 i の追加数，削除数である．

発言 j の追加の重み C_j^a，削除の重み C_j^d を求める．そして，求める編集者の発言の重みの平均を取ることで編集者の重要度 W を算出する．

$$C_j^a = \sum_i t_{i,j}^a, \qquad C_j^d = \sum_i t_{i,j}^d \tag{4}$$

$$W = \frac{\sum_j C_j^a}{n_a^c} + \frac{\sum_j C_j^d}{n_d^c} \tag{5}$$

ここで n_a^c と n_d^c はそれぞれ追加のあった発言の数，削除のあった発言の数である．

この手法で求めた編集者の重要度 W の高い編集者が重要な編集者となる．

4 おわりに

本研究では Wikipedia の編集行動とノートページでの発言を基に，発言に対応する編集が残った度合いを算出し，重要人物の特定に用いた．

本稿では，支持されている人物の特定を意見の本文への反映度合いのみを用いて求めた．しかし，この手法では編集がまだ行われていない議論中の話題に関する情報はわからない．そこで，重要人物の特定に，ノートページにおける議論の構造も用いることで，そのような話題で支持されている人物を特定することができると考えられる．

謝辞

本研究の一部は，科研費（25280039，23700113）の助成を受けたものです．

参考文献

[1] 緒方進, 池田真司, 牟田高信, 木本勝敏. Web 上のテキスト情報を用いた人物評価手法 (辞書構築). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2005, No. 1, pp. 9–14, jan 2005.

[2] 藤井敦. Opinionreader : 意思決定支援を目的とした主観情報の集約・可視化システム (データマイニング). 電子情報通信学会論文誌. D, 情報・システム, Vol. 91, No. 2, pp. 459–470, feb 2008.