対話型ネットニュースグループにおける話題転換記事の推定

内元 清貴　小作 浩美　井佐原 均
郵政省通信総合研究所 関西先端研究センター

1　はじめに

我々は、対話型ネットニュースグループから、ユーザが指定した記事と関連する記事群を自動的に抽出して提示するシステムを開発中である [1] 。対話型ネットニュースグループでは、複数のユーザによる対話が記事の形式で行われている。その一連の流れは記事中の References という情報から比較的容易に、ほぼ自動的に復元可能であり、関連する記事を集めるには、Referances を用いるのが簡便であると考えられてきた。ところが、fj.life.health と fj.living のニュースグループについて調査した結果、References を用いて復元できる一連の流れは、少ないもので1記事、多いものでは 407 記事と多岐に渡っている [2]。したがって、適度な数の記事群をユーザに提示するためには、References から関連付けた記事が少ない場合には、関連する記事をさらに集め、多い場合には、話題の関連が弱い部分を削除する必要がある。

これまで、関連する文書の検索やテキストセグメンテーションに関する研究の多くは何らかの辞書情報を用いることを前提に行われてきた。しかし、これはある程度使われる単語が限られている場合、あるいは検索対象の文書の量が安定している場合にのみ有効である。一方、ネットニュース記事の投稿者及び話題は多種多様であり、使われる言葉も多岐に渡っている。その上、記事の投稿数は年々増加する傾向にある。そのため、辞書を用いることにすれば、大規模にならざるを得ず、また頻繁に更新する必要が生じる。そこで我々は、References から関連付けた一連の流れをもとに、辞書を用いず、関連記事の収集、話題が転換している記事の推定を試みた。

関連記事の収集には、現在、キーワード検索による手法を開発中である。その際、抽出したキーワードのノイズを少なくするためには、関連付けた一連の流れから話題が異なる記事を枝刈りしておくのが良い。そのためにも、話題が転換している記事の推定が不可欠である。話題が転換している記事の推定に関しては、単純には、ニュース記事中の Subject という、記事の題名に相当する情報を利用する方法が考えられる。つまり、Subject が変わったところが転換点であると推定する方法である。しかし、実際には余程のことがない限り、話題が変わっていても Subject は変更されない。あるいは話題が変わっていないのに勝手に Subject が書き換えられることもある。そのため、話題転換記事の推定には別の手法が要求される。

そこで本稿では、単語辞書などの情報を使わず、漢字文字列の頻度や表層的手がかりのみを用いて話題転換記事を推定する手法を提案する。実際のニュース記事を用いた実験から、本手法により、高い精度で話題転換記事を推定できることが示された。

2　話題転換記事

ネットニュースグループに投稿される各記事は、References という、記事間の参照関係を表すリストを情報として持っている。そして、この情報を利用することにより、記事間の参照関係を図1右のような木構造に容易に復元できる。我々は、この木構造を**リファレンスツリー (RT)** と呼んでいる。 ここで、下位の記事は上

記事	References
A	
B	A
C	A-B-C
D	A-B-D
E	A-B-D-E
F	A-B-D-E-F

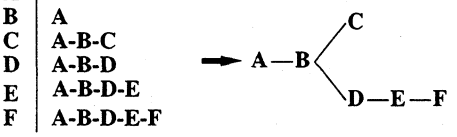

図 1: References と RT の関係

位の記事に対する回答またはコメントであるため、RT の根から葉に向かって、例えば図1の A→B→D→E→F のように一つのパスを辿れば、記事群中の一連の流れを追うことができる。また、枝分かれは、ある記事に対して複数の回答またはコメントがあったことを示している。

パスが長くなれば、あるいは枝分かれが多くなれば、異なる話題に転換する可能性が高くなる。実際、話題の転換点となる記事はこのような場合に現れることが多い。以降では、この話題の転換点になる記事を**話題転換記事**と呼ぶ。ただし、話題が転換しているかどうかは後続の記事がどのように内容をつないでいるかで決まることが多いため、RT の葉に相当する記事は話題転換記事とは考えない。

3 話題転換記事の推定法

同じ話題の記事群中では、使われる単語の種類が類似していることから、一連の記事群においては、使われる単語が前後で大きく異なるところで話題が転換していると予測される。本手法では主にこのような仮定のもとで、文字の出現頻度の変化から話題の転換している記事を推定する。

手法としては以下の二つの方法がある。これらはそれぞれ記事の流れを片方向あるいは双方向にとらえた方法であり、結果としてお互いに補い合うため、実際の話題転換記事推定は二つの手法を合成して行う。つまり、二つの方法の結果の積を推定結果とする。

本手法では辞書を用いないため、単語の厳密な分離抽出は不可能である。そのため、単語の代わりに**キーワード**を用いる。ここでキーワードは、一文字以上からなる全ての漢字列あるいは「n グラム」のようなカタカナ英数字列とする。

3.1 手法 1 の基本的な考え方

各記事中に出現するキーワードの頻度を RT の根から葉にかけて順次調べていく。すると話題転換記事には「**記事中のキーワードの中で新たに出現したキーワードの割合が、前記事のそれに比べて高くなる (特徴 1)。**」という特徴が見られると考えられる。したがって、この特徴を持つ記事を話題転換記事と推定する。

適当な大きさの記事からなる RT では、下位の記事へ行くにしたがって、根に相当する記事中で用いられていたキーワードと同じキーワードの占める割合が単調に減ると考えられる。ところが、実際には記事の大きさは多様であり、長い記事では前の記事と同じ話題について述べていても新出の語がたくさん現れる可能性が高いため、必ずしもそうはならない。そこで、長い記事の影響を軽減するために、根記事中のキーワードと同じキーワードの占める割合が単調に減少している場合のみ、話題転換記事と推定する。

実験では、対象とする対話型ネットニュースグループを fj に限定し、日本語の記事のみについて本手法の適用を試みた。日本語の記事の場合、「今日私」などの任意の漢字文字列がキーワードとなるので、新出のキーワードの割合は常に高くなってしまう。そこで、この影響を軽減するために、キーワードの代わりにキーワードを文字単位にばらした**キーワード要素 (KE)** を用いる。ただしこれは、漢字の場合にのみ有効であるので、カタカナ英数字列に関しては、キーワードを KE として用いる。

以上の制約のもとで、例えば図 2の記事 A_i のように、新出の KE の割合 (**新出率**) が高くなっていて、かつその記事の前後で、根の記事の KE と同じ KE の占める割合 (**重複率**) が単調に減少している記事を話題転換記事と推定する。

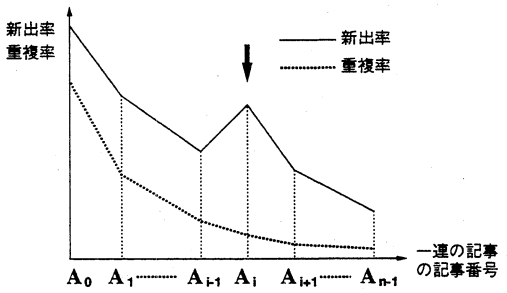

図 2: 記事と KE の新出率、重複率の関係

3.2 手法 1 のアルゴリズム

以下の手順で話題転換記事を推定する。まず、一つの RT を構成する各記事ごとに**キーワード要素 (KE)** の頻度を調べる。例えば、「東京特許許可局」ではそれぞれ「東 (1)、京 (1)、特 (1)、許 (2)、可 (1)、局 (1)」が KE となる。括弧内は頻度を表す。

次に RT の根から任意の葉までのそれぞれのパスについて、各記事における KE の**新出率**と、根の記事に対する KE の**重複率**を計算する。ここで、あるパス $\{A_0, \ldots, A_i, \ldots, A_{n-1}\}$ での記事 A_i における KE の新出率、重複率をそれぞれ次のように定義する。

$$\text{新出率} = \frac{\left(\begin{array}{c}A_i\text{のKEで}A_0 \sim A_{i-1}\text{に}\\ \text{一度も現れなかったもののA}_i\text{での頻度の和}\end{array}\right)}{A_i\text{の全KEの頻度の和}}$$

$$\text{重複率} = \frac{A_0\text{と}A_i\text{に共通して現れるKEのA}_i\text{での頻度の和}}{A_i\text{の全KEの頻度の和}}$$

最後に、パス $\{A_0, \ldots, A_i, \ldots, A_{n-1}\}$ において、以下の条件を全て満たす記事 A_i $(1 \leq i \leq n-2)$ を話題転換記事と推定する。推定は根の記事から葉の記事にかけて順に行う。

- (A_{i-1} の KE の新出率) < (A_i の KE の新出率)、かつ、A_{i-1} が話題転換記事と推定されていない。
- (A_{i-1} の KE の重複率) > (A_i の KE の重複率) > (A_{i+1} の KE の重複率)

3.3 手法 2 の基本的な考え方

記事群を話題転換記事より前と以後の二つに分けるとき、「**一方の記事群では高頻度、他方の記事群では**

低頻度で現れるキーワードの割合が高い(特徴2)。」という特徴が見られると考えられる。そこで、RT中の各記事を取り上げ、その記事の前後で特徴2が見られる場合にその記事を話題転換記事と推定する。特徴2を満たすかどうかは次のようにして判断する。まず、取り上げた記事の前後二つの記事群それぞれについて、キーワードを抽出し、各記事群における出現頻度を調べて図3の左表を作る。もし取り上げた記事が話題転換記事なら、キーワードの出現頻度は、同じ記事群での頻度 H_1、H_4 の方が他記事群での頻度 H_2、H_3 より十分大きくなる。したがって、例えば図3の右図のように、H_1/H_2、H_4/H_3 が十分に大きくなっている記事 A_i を話題転換記事と推定する。実際には、RT中での一般的な語の影響を除くため、キーワードに得点付けして、キーワードの選別を行っている。

実験では、手法1と同様、日本語の記事のみについて本手法の適用を試みた。ここでも3.1節で述べたのと同様の理由で、$H_1 \sim H_4$ として、キーワードの代わりにキーワード要素(KE)の頻度を用いる。さらに、前後両記事群間での記事数の違いによる影響を軽減するために、KEの頻度として、一記事あたりの頻度に換算したものを用いる。

	前記事群における頻度	後記事群における頻度
前記事群のKW	H_1	H_2
後記事群のKW	H_3	H_4

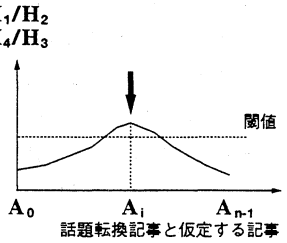

図3: 話題転換していると仮定する記事とスコアの関係

3.4 手法2のアルゴリズム

RT中のパス $\{A_0, \ldots, A_i, \ldots, A_{n-1}\}$ において、記事 A_i が話題転換記事であるかどうかを以下の手順で推定する。これは、各記事 A_i について行う。

1. 話題転換記事 A_i を仮定し、その記事を境に記事群を前記事群 G_{pre}=$\{A_j | 0 \leq j \leq i-1\}$、後記事群 G_{post}=$\{A_j | i \leq j \leq n-1\}$ に分ける。
2. 各記事 A_j $(0 \leq j \leq n-1)$ からキーワード $\{KW_{A_jw} | 1 \leq w \leq l_j\}$ (l_j は A_j 中のKWの数)を抽出する。

キーワードの得点 $S(KW_{A_jw})$ は、A_j の属する記事群 G での KW_{A_jw} の頻度をプラス、記事 A_j の属さない記事群 G' での KW_{A_jw} の頻度を重みつきのマイナスで反映して、一記事あたりの頻度に換算したものに、キーワードの前後の機能語を考慮した得点 $S_0(KW_{A_jw})$ を加え、

$$S(KW_{A_jw}) = \sum_{k=1}^{n} S_k(KW_{A_jw}) + S_0(KW_{A_jw})$$

のように計算する。ここで、

$$S_k(KW_{A_jw}) = \begin{cases} \frac{H_k(KW_{A_jw})}{|G|} & (A_k \in G) \\ -(\text{重み}) \times \frac{H_k(KW_{A_jw})}{|G'|} & (A_k \in G') \end{cases}$$

である。また、$H_k(KW_{A_jw})$ は KW_{A_jw} のキーワード要素 $KE_{A_jwc}(1 \leq c \leq l_{A_jw}$: l_{A_jw} は KW_{A_jw} の文字列長)の A_k における頻度 $H_k(KE_{A_jwc})$ から次のように換算できる。

$$H_k(KW_{A_jw}) = \frac{\sum_{c=1}^{l_{A_jw}} H_k(KE_{A_jwc})}{l_{A_jw}}$$

3. 前後各記事群ごとに得点がある閾値を越えるキーワードを集め、文字単位にばらしてキーワード要素 $\{KE(pre)_i | 1 \leq i \leq l_r\}$、$\{KE(post)_j | 1 \leq j \leq l_o\}$ を得る。そして、各記事群のキーワード要素 $KE(pre)_i$、$KE(post)_j$ の前記事群における頻度 $H_{pre}(KE(pre)_i)$、$H_{pre}(KE(post)_j)$、後記事群における頻度 $H_{post}(KE(pre)_i)$、$H_{post}(KE(post)_j)$ からそれぞれ一記事あたり、かつ一文字あたりの頻度 H_1、H_2、H_3、H_4 を求める。計算式は例えば、

$$H_3 = \frac{\sum_{j=1}^{l_o} H_{pre}(KE(post)_j)}{|G_{pre}| \times l_o}$$

となる。

4. H_1/H_2 及び H_4/H_3 が十分大きく、かつ一記事あたりの平均キーワード数が一定数以上あれば記事 A_i で話題が転換していると推定する。

4 実験と評価

対話型ネットニュースグループであるfj.life.healthとfj.livingから約10000記事を取り出し、RTを構成した。この中から話題転換記事を含むRT20個、合計約400記事に対して、本手法を適用した。キーワードの抽出は各ニュース記事からヘッダとフッタを切りとったメッセージの部分に対して行っている。

本手法の評価のために、本手法を適用した合計約400記事を対象とし、予め被験者3人によって話題転

換記事の認定を行った。話題が変わるとともに記事中のSubjectが変わっているものは少ないため、各記事を実際に読むことによって認定を行った。認定するしないには個人差があるため、正解としては、3人のうち3人とも認定した記事(正解1)、2人以上が認定した記事(正解2)の2段階用意した。また人間でも、認定の有無、認定箇所の揺れがあるので、評価には二つの基準を設け、基準1ではシステムの推定した記事が正解と一致したものを正答と認め、基準2ではシステムの推定した記事が正解記事の前後であっても正答と認めた。

実験結果は表1の通りである。基準2の再現率と適合率の計算で、分子の値にずれがあるのは、正解に広がりを持たせたことにより、正解とシステムの推定結果が一対二あるいは二対一に対応する場合があるためである。

表1: 実験結果

基準1	再現率	適合率
正解1	5/6 (**83%**)	5/18 (28%)
正解2	10/35 (29%)	10/18 (56%)

基準2	再現率	適合率
正解1	5/6 (**83%**)	6/18 (33%)
正解2	21/35 (60%)	17/18 (**94%**)

5 考察

表1において、正解1に対しては、再現率が良いのが望ましい。また正解2に対しては人間の認定結果にも揺れがあるため、基準2のように正解の前後の記事を含めて正答と認めたときに、推定を誤らない、つまり、適合率が良いのが望ましいと考えている。結果は、正解1に対して再現率83%、正解2に対して基準2で適合率94%と良かった。これは、十分実用に供し得る結果だと考えられる。

正解2に対して基準2で一箇所だけ推定を誤っているのは、記事の大部分を2記事前からの引用部分で占めており、推定箇所が2記事分ずれてしまったためである。このような場合に対処するため、今後引用部分の扱いを検討する必要がある。また、正解1に対して一箇所だけ推定が困難であったのは、これを含むRTそのものが記事数、テキスト量ともに少なかったためである。このようにテキスト量が少ない場合にはSubjectを利用する方法も考えている。

ちなみに、実験対象のRT全てにおいて、パスの途中でSubjectが変更されている所は14箇所あった。そのうち、被験者の2人以上が話題転換記事と認定した箇所と一致しているのは、7箇所、隣の記事まで許せば11箇所であった。この7箇所の内、推定が困難であった一箇所を除く6箇所は本手法でも推定できた。

本手法は、3節に示したアルゴリズムからも分かるように、RTの枝の途中で話題が転換している記事を推定する場合には、優れている。しかし実際には、RTの根に相当する記事が複数の話題を含んでおり、ここから各話題に分岐していることもある。この場合については話題が転換している訳ではないので、本手法では推定できない。この場合でも推定できるようにするため、枝ごとに話題の抽出を行って同じ話題かどうかを推定する方法を検討中である。

6 おわりに

対話型ネットニュースグループから、話題転換記事を自動的に推定する手法を提案した。推定には単語辞書などの情報を用いず、文字列の頻度や表層的な手がかりのみを用いる。システムの推定した記事が正解記事の前後の記事であっても正答と認めると、日本語の記事に対し、正解1に対して再現率83%、正解2に対して適合率94%と良い精度の結果が得られた。

本稿では、日本語の記事に対して推定を試みたが、3節で挙げた特徴1や特徴2は使用言語に依らず話題転換記事に見られる特徴であるため、他の言語の記事についても同様の方法で推定できると考えられる。ただ、複数の言語が混在する場合には言語をまたがっては単語間の対応が取れず、推定を誤ることになる。また、日本語の記事のみでも、同じ単語なのに異なった表記が使われている場合や、同じ概念なのに類義語が使われている場合には推定しにくい。したがって、より精度を挙げるためには、少なくとも基本的な語彙についての類義語辞書などの知識が必要となるだろう。

参考文献

[1] 小作浩美, 井佐原均, 話題関連性に着目した知的ニュースリーダの提案, 平成7年電気関係学会関西支部連合大会, (1995).

[2] 小作浩美, 内元清貴, 井佐原均, 知的ニュースリーダが対象とする対話型ネットニュースの特徴, 情報処理学会自然言語処理研究会, Vol. NL117-4, (1996).