キーワードによるネットワークニュース記事群の構造化

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、討論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抽出して提示するシステムを開発中である [1]。平成8年度にはそのプロトタイプが作成された [2]。本研究の目的は、ユーザに記事群を提示する際、ユーザが議論の流れを把握しやすいようにすることにある。議論のまとめとしては記事の要約などの手段が有効であるが、議論の流れは要約では伝わりにくい。そこで、議論の流れを要約するのではなく、記事群を構造化することによって、ユーザがスムーズに読み進められるよう支援する。

話題の移り変わりに着目すると、議論の流れには、最後まで一貫して同じ話題について議論される場合、途中で話題が変わる場合、ある記事で複数の話題が取り上げられ、その後、それぞれ別々に議論される場合がある。さらにそれぞれの流れの中には、提案、賛成、反対、補足など投稿者の態度が反映される。本稿では、投稿者の態度については保留することとし、話題に着目することによって、議論の流れをユーザに分かりやすい形で提示する手法について述べる。

ユーザに分かりやすい形として、我々は次のような構造を考えている。

- 話題が転換あるいは分岐している部分にはタグが振られ、話題の違いは代表的なキーワードによって示されている。

こうすれば、ユーザは記事を読み進めていくにしたがって、次に話題が転換しているのか分岐しているのか、議論されている話題がどのように異なるのかを知ることができ、議論の流れを把握しやすい。我々の手法の特徴は、ユーザがある記事を読んだとき、次に読むべき記事の内容が大まかに分かるという点にある。

2 記事群の構造化

構造化した記事群の概念図を図1 に示す。ここで、木は抽出した一連の議論を表す。これは記事中の References という情報を用いて抽出できる。転換、分岐というタグは、その記事から話題が転換、分岐していることを表す。このような話題の転換点、分岐点に相当する記事は、記事中の主だったキーワードが記事間でいかに違うかを調べることによってよい精度で推定できる。ここで推定を決定づけるキーワードは話題の違いを表すものであり、これを抽出して図1 のように提示する。楕円で囲まれている記事は、話題が分岐した後、同じ話題について述べられている範囲を示す。

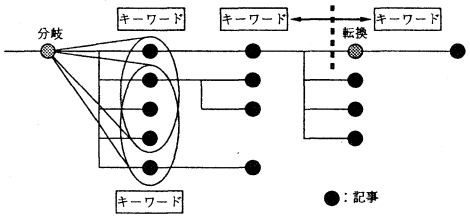

図 1: 概念図

以下、2.1 節で話題転換記事、話題分岐記事を定義し、2.2 節、2.3 節でそれぞれの記事を自動推定する手法について述べる。

2.1 話題転換記事、話題分岐記事

討論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。その一連の議論は記事中の References という記事間の参照関係を表すリスト情報から比較的容易に、ほぼ自動的に木の形で復元できる。我々は、この木を**リファレンスツリー (RT)** と呼んでいる。この木で、下位の記事は上位の記事に対する回答またはコメントであり、分岐は、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、あるいは分岐が多くなれば、異なる話題に転換、分岐する可能性が高くなる。以降では、この話題の転換点、分岐点になる記事をそれぞれ、**話題転換記事**、**話題分岐記事**と呼ぶ。

2.2 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。主にこのような仮定のもとで、文字の出現頻度の変化を調べることによって話題の転換している記事を高精度で推定することができる [3]。

本研究の一部は、情報処理振興事業協会「独創的情報技術育成事業」の一環として行われたものである。

推定には、次のような特徴を利用する。

1. 話題転換記事では、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。
2. 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が多い。

推定された各話題転換記事に対し、この二つ目の特徴を満たすキーワードを取り出し、ユーザに提示するキーワードとして用いる。

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」のような漢字カタカナ英数字列とする。これらキーワードはテキストから平仮名、句読点、記号を除くことによって得られる。一文字の場合、動詞や形容詞の語幹であることが多いため、原則としてキーワードとはしないが、後に手がかりとなる語「は」「が」「を」「として」のいずれかが付随していればキーワードと認める。

2.3 話題分岐記事の推定法

2.3.1 基本的な考え方

ある記事が複数の話題を含んでいる場合、その記事から分岐したそれぞれの枝では別々の話題について議論されていることが多い。しかし、はっきりと別々の話題に分かれる訳ではなく、多くの場合、図 1 の左の分岐部分のようにその境界部分が重なっている。そこで、ある記事から分岐した記事を、同じ記事が複数のクラスタに属することを許してクラスタリングする。クラスタリングは、任意の二記事を比較し、同じ話題について議論されている記事同士が同じクラスタに属するよう調節することによって行う。クラスタリング終了後、複数のクラスタが残った場合に話題が分岐していると推定する。例えば、ある記事 A_0 から 5 つの記事 $A_1 \sim A_5$ が分岐していたとする。任意の二記事を比較した時、話題の同じ記事が図 2 の表の ○ 印で示されるように判断されたとき、クラスタリングの結果は図 2 の右図のようになる。

話題が同じかどうかの判定には次の特徴を利用する。ある記事から分岐した枝の中で、同じ話題について議論されている枝同士では、

1. 分岐点に相当する記事から同じ部分を引用していることが多い。
2. 共通に使われる単語が多い。

具体的には、記事中の共通引用部分と共通するキーワードの割合を調べることによって判定する。キーワードとしては、記事中の全キーワードのうち、位置情報、前後の記事における頻度情報などを用いてスコアに重み付けし、閾値を設けて間引いたものを用いる。

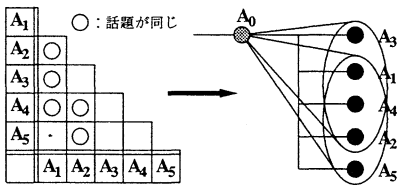

図 2: 記事のクラスタリング例

2.3.2 アルゴリズム

以下の手順で話題分岐記事を推定する。

1. 頻度、付随する手がかり語をもとに、各記事の各キーワードにスコアを与える。
2. 以下の条件にしたがって、RT を構成する各記事の各キーワードに重み付けする。重み付けの後、スコアが 0.5 以下のキーワードは削除する。以下で、注目している記事を子記事、子記事の一つ上位の記事を親記事、子記事の一つ下位の記事を孫記事と呼ぶ。

[位置情報の利用]

引用部分よりも非引用部分で述べられている内容の方が重要である。特に引用部の直後では投稿者の言いたい内容が書かれていることが多い。したがって、引用部分の直後の非引用行に現れるキーワードはスコアを 10 倍、その次の非引用行では 5 倍、それに続く非引用行では 2 倍とする。

[親記事、孫記事での頻度情報の利用]

親記事や孫記事にも使われているキーワードは話題の中心であることが多く重要であると考える。そこで、

- 親記事に使われていないキーワードはスコアを 0.1 倍とする。
- 孫記事があり、かつ孫記事にも使われているキーワードには、孫記事におけるキーワードのスコアを加える。
- 子記事には現れないが、親記事と孫記事に現れるキーワードを孫記事から取り出し、子記事のキーワードとして加える。

3. 各RTから根が複数の枝を持つ部分木を全て抽出する。そして、親記事A_rから分岐した子記事A_{c_k} ($1<k<n$: nは記事A_rの持つ枝の数) を引用行及びキーワードの一致する割合に基づいて分類する。各記事A_{c_k}を分類先の核と考え、核となる記事と他の任意の一記事が以下の条件を満たす場合、同じ分類先に分類する。

[両記事とも引用行がある場合]

(a) 親記事にも使われているキーワードのうち、比較している二記事に共通するキーワードの割合がいずれかの記事で0.8を越える。

(b) 比較している二記事に共通する引用行の割合がいずれかの記事で0.7を越え、かつ、非引用行に現れるキーワードで共通するものの割合が0.3を越える。

(c) 比較している二記事に共通する引用行の割合がいずれかの記事で0.1を越え、かつ、非引用行に現れるキーワードでスコア5を越えるものの中に共通するものがある。

[どちらかの記事に引用行がない場合]

(a) それぞれの記事に親記事にも使われているキーワードが三つ以上あり、かつ、そのキーワードのうち二つの記事に共通するキーワードの割合がいずれかの記事で0.7を越える。

(b) 各記事の全キーワードをベクトルの要素としてcos値を計算したとき値が0.1を越える。

4. 各分類先をそれぞれクラスタとみなし、部分集合となっているクラスタを排除する。

5. クラスタを構成する記事が持つキーワードをクラスタのキーワードとし、これをユーザに提示するキーワードとして用いる。提示の際、一つのクラスタにしか現れないキーワードを優先する。

6. 残ったクラスタが複数あり、かつ、それぞれのクラスタが少なくとも一つ以上のキーワードを持つとき、分岐しているものとみなす。

2.3.3 引用行の抽出

引用行かどうかは上位下位の関係にある記事の各行の対応関係を調べることによって判断する。まず、子記事の各行と親記事の各行を比較する。親記事の各行の先頭より数文字分を部分文字列として含む子記事の行があれば、それらを引用・被引用関係の候補とする。このとき子記事の行のうち、親記事の行頭に対応する文字より前の文字列を引用記号の候補とする。例えば、図3のような二記事があったとすると、引用・被引用関係の候補として図3の真中の対応関係が得られる。引用記号の候補は「>」、「> す」である。

次に、子記事の各行について、引用行かどうか、引用行だとすれば親記事のどの行を引用しているかを判断する。次の条件の順に優先して引用・被引用関係にあるとみなす。

1. 親記事の連続する複数の行からその順に子記事の連続する行が候補に挙がっている。
2. 子記事の行から見たとき、候補となる親記事の行が1行だけある。ただし、引用記号の候補の最後が記号あるいは空白の場合に限る。
3. 上の二つの条件で引用行とみなされた行から引用記号の候補を得、共通するものをその記事での引用記号とする。その引用記号から始まる行はすべて引用行とみなす。対応関係は一つ上あるいは下の引用行の引用・被引用関係を参考に決定する。

例えば、図3では親の１２３と子の１２３が条件1を親4と子6が条件2を親3と子5が条件3を満たす。

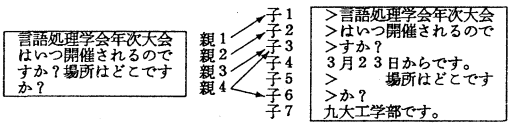

図3: 引用・被引用関係の候補の例

2.4 実験と評価

討論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。この中から無作為抽出した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。

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、各記事を実際に読むことによって話題転換、分岐記事の認定を行った。これを正解とし、システムの出力と比較した。実験結果は話題分岐記事の場合、再現率が78%(正解22記事中18記事を正しく推定)、適合率が82%(システムが推定した23記事中18記事が正解) であった。システムが正しく推定した18記事のうち、枝を3つ以上持つものは9記事あり、そのうちクラスタも正しく抽出できていたものは6記事(67%) であった。話題転換記事の場合、システムが2名以上の被験者が話題転換記事と認定した記事の前後の記事を話題転換記事と推定しても正解とした場合、再現率57% 適合率94% であった [3]。

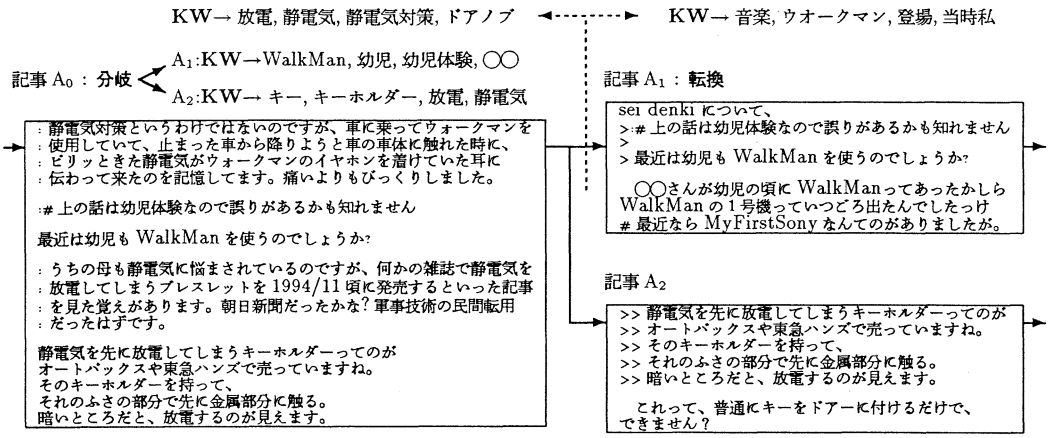

図 4: 実際の例 (KW→ : ユーザが記事 A_0 を読んだとき、次の記事を読む前に見るべきキーワード)

3 構造化の例と考察

2.4 節の実験で、話題転換記事、分岐記事はそれぞれ 35 記事、22 記事あり、そのうち、それぞれ 17 記事、18 記事がシステムにより正しく推定された。因みに話題が転換かつ分岐している記事は 3 記事あった。

この実験で本手法を適用した記事群をシステムの推定結果とその際抽出したキーワードを基に構造化した。実際の記事の一部を例として図 4 に挙げる。ここで、キーワードはスコアの高いものから上位 4 番目までを提示している。提示したキーワードのうち、図の 3 記事中に現れていないものは、図に挙げた記事の前後の記事で使われていたキーワードである。記事 A_0 までは主に静電気の話題であったが、その後話題が分岐して、記事 A_1 ではウォークマンの話題に変わり、記事 A_2 では引続き静電気の話題が議論されている。

提示されたキーワードが実際、議論の流れの把握に役立ったか、キーワードやその数は適切か、については今後実際のシステムに組み込んで心理実験を行うことによって評価したい。

4 関連・先行研究

キーワードを用いて文書を可視化する手法として、これまでに討論型ニュースグループ、WWW を対象にしたものが提案されている [4, 5]。いずれも、同じ話題について議論されている記事を近くに配置し、代表するキーワードを表示することによってどのような話題について議論されているかが視覚的に分かりやすいようにしている。ユーザが文書群を見たときに、全体としてどの部分でどのような話題が中心的に議論されているかを把握しやすいという利点があるが、議論の流れそのものは扱っていない。我々のプロトタイプシステム [2] では、議論の流れを RT として取り出し、さらに話題の転換点を検出することによって同じ話題が議論されている記事の範囲を提示できる。つまり、ユーザがある話題に興味を持ったときに読むべき記事の範囲を指定できる。本手法ではさらに話題が分岐している部分も自動推定し、推定に用いたキーワードを用いて、話題の違いが分かりやすい形でユーザに議論の流れを提示することができるようになった。

5 おわりに

討論型ネットワークニュースグループにおける議論の流れをユーザに分かりやすい形で提示する方法について述べた。本手法では、話題の転換点、分岐点を検出し、それぞれにタグを付けるとともに、話題の違いをキーワードを用いて提示する。特徴は、ユーザがある記事を読んだときに、次に読むべき記事の内容が大まかに分かるという点にある。本稿では、投稿者の態度の扱いについては保留したが、モダリティ表現などを検出して、提案、賛成、反対、補足など投稿者の態度を推定できるようにすることが今後の課題である。

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] 井佐原均, 小作浩美, 内元清貴 : 討論型ニュースグループを対象とする知的ニュースリーダの開発, 情報処理学会自然言語処理研究会, NL119-3,1997.

[3] 内元清貴, 小作浩美, 井佐原均 : 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, 言語処理学会第 3 回年次大会,1997.

[4] 矢部純, 高橋伸, 柴山悦哉 : ネットニュースの議論の意味的な分岐を反映した可視化, 日本ソフトウェア科学会第 14 回大会,1997.

[5] 有田英一, 安井照昌, 津高新一郎 : 単語集合の自動構造化機能を持つ「情報散策」方式, 信学技報, NLC95-17,1995.