時系列分析によるWeb文書の情報信頼性判断支援：時系列変化からの重要トピックの抽出

大西 貴士　岡嶋 穣　河合 剛巨　中澤 聡　安藤 真一
NEC 共通基盤ソフトウェア研究所
{t-onishi@bq, y-okajima@bu, t-kawai@bx, s-nakazawa@da, s-ando@cw}.jp.nec.com

1 はじめに

Web文書に書かれている内容の信頼性を判断するには、少数のWeb文書から得られる偏った内容だけで判断するのではなく、様々な観点でその内容を吟味し、それらを総合して判断を下す必要がある。我々は信頼性を判断する対象として単文レベルのテキスト情報（以下、言論とする）を扱い、その情報信頼性判断の支援を行うにあたって、時間的・論理的な観点で関連する言論やトピックを抽出し、抽出してきた言論やトピックを俯瞰・要約する技術の研究開発を行っている[1]。

本稿では、その中で時間的な観点で関連するトピックの抽出を取り扱う。情報信頼性を判断する上で、その言論に関して過去に盛り上がった際の要因となった言論やトピックを知ることは重要である。例えば、「温暖化の原因は CO2 である」という言論の情報信頼性を判断したい場合に、2007 年 2 月に発表された「IPCC 報告書」を知っておくべき重要なトピックとして提示することは、利用者が情報信頼性を判断するにあたって有益な手がかりとなると考えられる。

そこで本研究では、情報信頼性の判断をしたい言論（以下、着目言論とする）の出現数の時系列変化に注目し、着目言論が盛り上がった（出現数が大きく増加した）要因となるトピックを重要トピックとして抽出する手法を提案する。

2 課題とアプローチ

着目言論の出現数の時系列変化を利用して重要言論を抽出する手法としては、一定の期間ごとに、その期間に特徴的に出現する言論を抽出する手法が考えられる。だが、この手法では、その期間内で起きた大きな話題についての言論だけが抽出され、重要だが大きな話題に隠れている話題があっても抽出できないといった課題がある。着目言論に関連する大きな話題を抽出することももちろん重要であるが、大きな話題に関連する言論のみ抽出するのでは、情報信頼性判断を支援するための情報としては偏りが生じ適切ではない。利用者にとっても大きな話題は既知であることが多いと考えられるため、新規性のない情報の割合が多くなってしまう（課題1）。また、特定の期間に着目して重要言論を抽出する手法では、着目する期間内に偶然同時に起きた、着目言論とは関係ないニュースやイベントに関する言論も抽出されてしまう（課題2）。

そこで、上記の課題1を解決するために着目言論の出現数が大きく変化した区間（変化区間）に注目する。重要言論は、変化区間において着目言論と同期して出現すると考えられる。例えば、着目言論に関係の深いニュースやイベントが発生することによって着目言論の出現数が大きく増加したり、着目言論に対立する言論が現れた結果、それ以降での着目言論の出現数が減少したりすることがある。よって、そうした着目言論の出現数に大きな変化が起きた時点にフォーカスを当てて調べることでそうした変化の要因や同じ要因で出現数が変化した言論を浮かび上がらせることができる。さらに、複数の変化区間について別個に扱うことができるので大きな話題に関連する言論だけでなく小さな話題に関連する言論も抽出することができる。

また、重要言論であれば変化区間だけでなく変化区間以降でも同期して出現すると考えられるため、変化区間以降においても継続的に相関関係を持つ言論を優先的に抽出することで課題2に対応し、精度よく重要言論を抽出することができる。

3 重要トピック抽出手法

以上をふまえて、着目言論の変化区間に特有に現れる言語表現を重要トピック候補とし、変化区間以降の着目言論と重要トピック候補の相関性を用いて重要トピックを抽出する手法を提案する。具体

的な手順は以下の通りである。

3-1 着目言論の時系列データの取得

着目言論を含む文書を、文書が書かれた日付情報付きでデータベースから取得する。なお今回は着目言論を含む文書の取得にキーワード検索を使用し、着目言論を含む文書の文書集合と同等の文書集合を返す検索クエリを人手で作成して着目言論を含む文書を取得した。

取得した文書集合は時間情報を持っているため一定期間ごとの文書数から着目言論の出現数の時系列データが得られる。データベース中の期間ごとの総文書数自体も時系列変化があるため、期間ごとに着目言論の出現数を総文書数で割ることで正規化した時系列データを得る。図 1 の太線は、ある関連言論の時系列データを 3 日ごとにプロットしたものである。

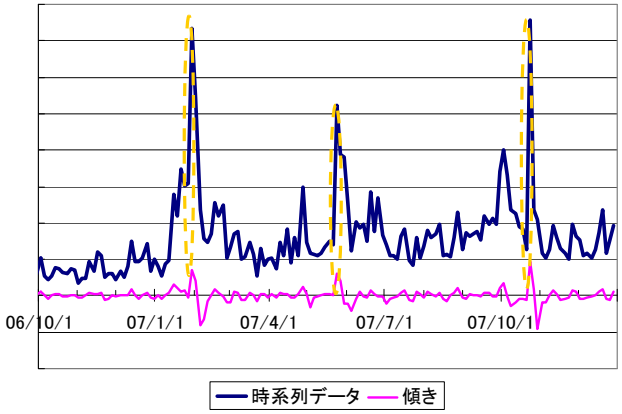

図 1 着目言論の時系列変化と変化区間

3-2 変化区間を抽出

3-1 で取得した着目言論の時系列データの各点で傾きを求め、そこから変化区間を抽出する。時系列データの傾きは、傾きを求める時点を含めた過去 3 点の値から得られる回帰直線の傾きとした。

変化区間は、時系列データの傾きが大きい点から順に抽出していき、傾きが正の区間とする。ただし、得られた変化区間が予め設定された最小区間幅に満たない場合は、後方(未来側)の傾きが負である区間の範囲内で最小区間幅に達するまで変化区間を拡大する。最小区間幅を設定しているのは、3-3 で重要トピック候補を抽出するのに十分な文書数を確保するためであり、予備実験の結果から最小区間幅は 15 日とした。図 1 の細線が時系列データの傾きで、点線で囲った部分が変化区間である。

3-3 重要トピック候補の抽出

3-1 で取得した文書集合について、3-2 で求めた変化区間に含まれる文書と変化区間以前の 1 年間に含まれる文書とを比較し、変化区間に含まれる文書に特徴的に出現する言語表現を重要トピック候補として抽出する。

重要トピック候補の抽出は、文書を構文木に変換し、構文木の全ての部分木について拡張型確率的コンプレキシティ(Extended Stochastic Complexity; ESC)を求め、ESCの値が大きいものから順に抽出する[2]。ESCは情報理論的な複雑さを表す尺度であり、意見分析における特徴語抽出の尺度としての優位性が示されている[3]。比較対象の文書集合を変化区間以前の 1 年間としているのは季節による変動を抑制するためである。

以降は 3-3 で抽出した重要トピック候補ごとに各手順を行う。

3-4 重要トピック候補を含む文書集合の取得

3-3 で抽出した重要トピック候補ごとに、3-1 で取得した文書集合に含まれる文書で、重要トピック候補も含む文書を取得し、時系列データを得る。

3-5 変化区間以降での相関性による絞込み

着目言論に関して重要なトピックは変化区間以降も着目言論と同期して継続的に話題に上っていると考えられる。逆に、変化区間以降で相関関係がないものは偶然変化区間に同時に話題に上っただけのトピックであると考えられる。

そこで、変化区間以降での着目言論の時系列データと重要トピック候補の時系列データとの相関係数を求め、相関係数が閾値以上の重要トピック候補を重要トピックとして出力する。

4 評価実験

重要トピック抽出手法の 3-3 までの段階で適切な重要トピックを含んだ重要トピック候補を抽出できているかどうかを確かめるために以下の実験を行った。着目言論は、表 2に示す 4 つを想定し、それぞれキーワード検索のクエリとして表現したものを使用した。対象データは、2006 年、2007 年のブログデータから各着目言論において出現数の変動が大きい四半期を最大 3 つとって使用した。

表 2 実験に使用した着目言論

想定する着目言論	キーワード検索のクエリ
裁判員制度は問題である	裁判員制度
二酸化炭素の排出規制が温暖化の抑制につながる	温暖化 (co2 二酸化炭素)
動画サイトは違法である	(動画 YouTube ニコニコ動画 ニコ動) (侵害 違反 抵触 違法)
メタボリックシンドロームの診断基準は不適切だ	(メタボ メタボリック メタボリックシンドローム) 基準

※ ( )で囲まれている部分は OR 条件で、それ以外は並列されている単語の AND 条件を示す。

4-1 提案手法

各四半期から変化区間を 3 つ抽出し、それぞれ重要トピック候補を 6 個抽出する。つまり、1 四半期から合計で 18 個の重要トピック候補を抽出する。

4-2 ベースライン

各四半期とそれよりも過去の期間に含まれる文書を比較し、その四半期に特徴的に出現する言語表現を ESC が大きい順に 18 個抽出する。

4-3 評価方法

与えられた着目言論の情報信頼性を判断するにあたって、抽出された重要トピック候補が有効であるかどうかの観点で 3 人の評価者による評価を行った。しかし、本研究では重要トピック候補は部分木の形で抽出されるため、それだけを見ても評価ができない。そこで、重要トピック候補の周辺に今後抽出可能になる重要言論があると考え、本実験では重要トピック候補の前後 100 字程度を評価者に読んでもらい着目言論の情報信頼性判断に有効な言論が含まれているかどうかで評価を行った。

また重要トピック候補が情報信頼性判断に有効であると評価されたものについては、さらに、その言論が評価者にとって既知の情報かどうかの判定も行った。これによって、抽出された重要トピック候補の中に既に知っている当たり前の情報がどの程度含まれているかをみることができる。情報信頼性の判断支援を行う上では、既知の割合が小さいほど利用者に新規の情報を提示できたということになる。

4-4 結果と考察

有効言論の含有率、既知率はそれぞれ図 3、図 4のようになった。ここで含有率、既知率は、3 人の評価者の評価結果をマージして算出した。

含有率については、提案手法とベースラインに有意な違いは見られなかったが、既知率については、両手法で 0%だったメタボ基準を除いてすべて提案手法がベースラインよりも小さくなった。このことから提案手法は、関連はしているが当たり前となっている情報のみ提示しているのではなく、利用者にとって新規の有用な情報を、精度を犠牲にすることなく抽出できていることが分かる。

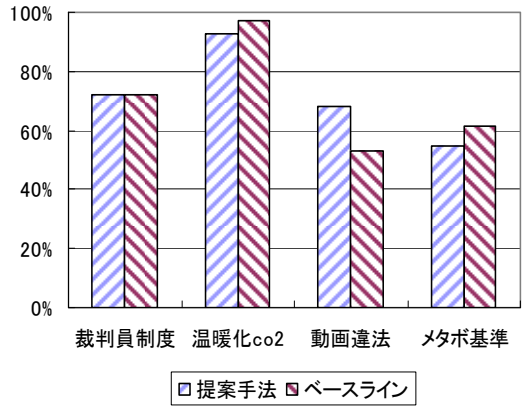

図 3 有効言論の含有率

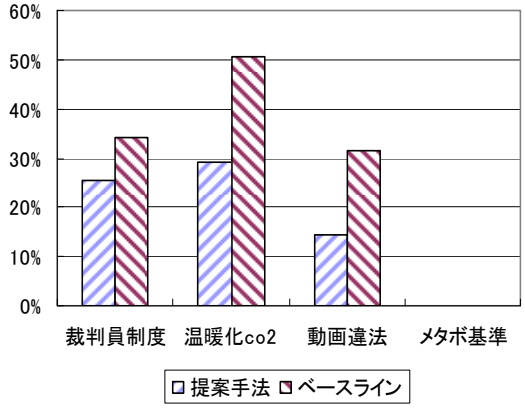

図 4 有効言論の既知率

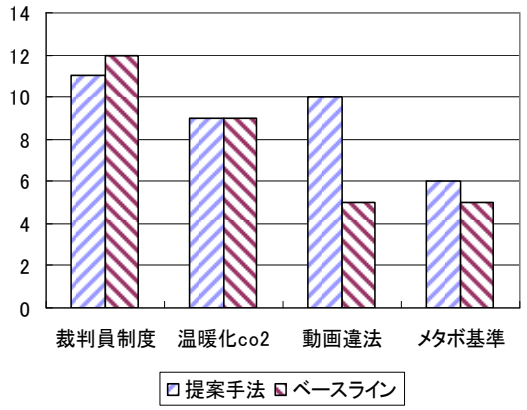

図 5 重要トピック候補中の話題数

また図 5は、同一文書に出現する重要トピック候補を同一視しユニークな話題だけをカウントしたものである。裁判員制度、温暖化についてはベースラインでも多くの話題が抽出できているため提案手法とベースラインとの差は見られないが、動画違法、メタボ基準では提案手法のほうがベースラインよりも多くの言論が得られた。これは、ベースラインでは同一の話題に関する言論が多数重なって抽出されるのに対して、提案手法では、3 つの変化区間でそれぞれ話題になった言論が抽出されるので、同じ基準で上位の重要トピックを同数個だけ抽出しても提案手法のほうが、抽出される重要トピックの重なりが少なくなったと考えられる。

以上の結果から、提案手法は、まず複数の変化区間を抽出し、ついで変化区間ごとに重要トピックを抽出することで、大きな話題に隠れているが着目言論の情報信頼性判断に有効な重要トピックを抽出できていることが分かる。

最後に、重要トピック抽出手法の 3-5 で閾値を0.3としたときの重要トピックの含有率を算出したところ、表 6のようになった。

これによると、温暖化 co2、動画違法の 2 つで含有率の向上が見られた。逆に、裁判員制度では大きく含有率が低下したが、有効でないと判定されたものをみてみると光市母子殺害事件に関連するものが多かった。内容をみると、判定に用いた前後 100 字の範囲には裁判員制度と関連する記述はなかったものの、光市母子殺害事件の裁判の経過をうけて、裁判員制度が始まった際の問題点や疑問点を述べているブログが多く、裁判員制度に関連する重要トピックとしてもよいと考えられる。

以上の結果から、変化区間以降の相関性を利用することが重要トピックの抽出に有効であると言える。

表 6 相関係数 0.3 以上での含有率

裁判員制度	温暖化 co2	動画違法	メタボ基準
39%	100%	80%	58%

5 まとめと今後の予定

本研究では、着目言論の時系列データから変化区間を抽出し、変化区間ごとに特徴的に出現する言語表現を着目言論の重要トピック候補とし、変化区間以降の着目言論と重要トピック候補の相関性を利用して重要トピックを抽出する手法を提案した。また、実際の Web データを用いた評価実験を行い、着目言論の情報信頼性判断に有効な重要トピックを精度よく抽出できることを確認した。

今後は、キーワードレベルで扱っていた言論表現を単文レベルへ拡張し、手法の改良を行う予定である。さらに、別途開発している情報信頼性検証技術と統合し、一般の利用者が着目する主題の情報信頼性を判断する手がかりとなる情報を総合的に提供する情報信頼性支援システムの開発に取り組む予定である。

謝辞

本研究は、独立行政法人情報通信研究機構(NiCT)の委託研究「電気通信サービスにおける情報信頼性検証技術に関する研究開発」の成果である。

参考文献

[1] 中澤聡,岡嶋穣,大西貴士,河合剛巨,安藤真一: 時系列分析によるWeb文書の情報信頼性判断支援:全体概要, 言語処理学会第 15 回年次大会, 2009.

[2] S. Morinaga, H. Arimura, T. Ikeda, Y. Sakao, and S. Akamine: “Key Semantics Extraction by Dependency Tree Mining,” Proc. of KDD2005, p.666-671, 2005.

[3] H. Li and K. Yamanishi: "Mining from Open Answers in Questionnaire Data," Proc. of KDD2001, p.443-449, 2001.