Web ページ更新情報への自己組織化マップの適用

濱口佳孝† 池野篤司† 山本英子‡ 井佐原均‡
沖電気工業株式会社† 独立行政法人情報通信研究機構‡
hamaguti662@oki.com, ikeno546@oki.com, eiko@nict.go.jp, isahara@nict.go.jp

Web ページの記事の最近の傾向を把握する目的で、ある時点での Web ページの更新差分に自己組織化マップを適用した。しかし、広告、差分を得た時点に関する日時表現、記事に関係無いテンプレート的部分の語によるクラスタが約８割を占め、適当なクラスタが得られなかった。これについて、・固有表現抽出により得られた日付表現の排除、・<a>タグの特徴や特徴的な単語のパターンによる広告の一部の排除よる解決を試みた。また、使われる語が多様なテンプレート的な部分に関しては、それぞれの Web ページごとに、時間軸方向での出現頻度の高い語の重みを低くすることにより、その抑制を試みた。この結果、これらによるクラスタが作られることを抑制し、記事中の語によるクラスタを得ることができた。

1. はじめに

沖電気では、設定した URL を定期的にチェックし、更新部分の差分のみをユーザに通知するサービス、MAILPIA® [1](http://www.mailpia.jp)を提供している。これにより、興味を持つ Web ページの情報を効率良く得ることができる環境を提供することを目的としている。

このサービスでの更新情報は“興味を持ってウォッチングしている人がいるサイト”の物である。これらのある時点での傾向を把握できれば、最近の話題・興味の傾向の視点で情報をまとめて扱えることが期待でき、ユーザが興味を持つ情報の提供の高度化に活用できると考えている。

このために、ある時点で得られた Web の更新部分を、差分を得ることで取得して自動クラスタリングすることを試みた。

クラスタリング手法としては、各文書間の相互の距離によりクラスタリングする手法の他に、T.Kohonen による自己組織化マップ[2]と呼ばれる手法が知られている。

今回の実験の目的では、定期的に更新情報が取得され新しいデータ群が与えられることから、処理にある程度のスケーラビリティが求められる。この点で、各文書間の相互の距離を計算する必要がある手法は単純には更新情報数の自乗の処理が必要となり、好ましくない。

これに対して自己組織化マップは、各参照ベクトルの学習に必要な回数がほぼ一定であるとすれば、参照ベクトルの数、すなわちマップのサイズに比例した処理量が必要となる。更新情報中の話題の多様性は、最大でも情報の数に比例程度で、その時点での話題のバリエーションには限界があることを考えれば、マップの大きさはデータ数に比例するより小さくなることも期待される。このため、情報量に対する処理量の増加率は有利であることが期待できる。

また、短期間では話題の傾向は変わらず作られるマップに大きな変化は無いとすると、前回の更新情報から生成されたマップを初期状態として学習を開始することで、以後の学習回数を減らすことも考えられる。

このため、今回の実験では自己組織化マップを用いたクラスタリングを行うこととした。

2. 実験の概要

実験には、ユーザが設定している約 5000URL から、2005 年 3 月 31 日に 4 回取得した Web ページの差分である、732 データを用いた。

この実験では、Web ページの差分データの特徴ベクトルは、固有表現抽出・属性抽出[3]により得られた名詞系の語を各成分とした。また、この特徴ベクトルと、自己組織化マップの参照ベクトルとの類似度は内積とした。

自己組織化マップの形状は、マップの端の条件が変わることを避けるため、20×20 の矩形のマップの上下端と左右端が繋がったトーラス状とした。

学習のパラメータは、4.1 節で提案する手法を適用し、さらにテンプレートなどで用いられる語を手作業でストップワードとして登録することで抑制した実験（結果は図３）で、主観評価で最も良いクラスタが得られた値を用いている。

3. 差分全体の tf・idf を用いた実験

まず、固有表現の各差分データ中での出現頻度 tf と、その固有表現を含む差分データの数 df で全データ数を除算したものの対数である idf を用いて、その固有表現の成分の大きさを tf•idf とした。これをベクトル長が 1 になるように正規化したものを、各差分データについての特徴ベクトルとして、自己組織化マップを作成した。（図 1）

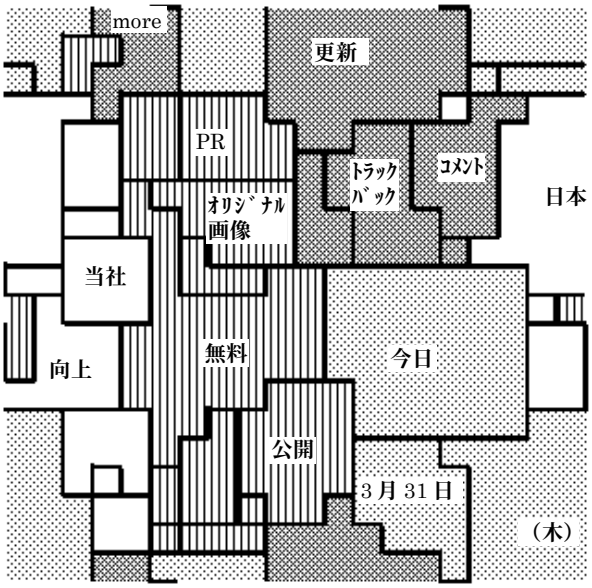
その結果、「3 月 31 日」など差分を取得した日時に関した固有表現や、「今日」という当日に関係した相対表現がもっとも強い参照ベクトルによるクラスタが 6 つ作られた。

また、「無料」「CLICK」等、単語がテキスト広告によく用いられることに起因する 12 クラスタが作られた。「ケータイ」「モバイル」や商品名によるクラスタなど話題に関連して形成されたように見えるクラスタの中にも、そのクラスタに集まる差分データの大半ではテキスト広告部分にそれらの語が含まれている物があった。これらは実際は記事ではなく広告により作られたクラスタであるため、広告によるクラスタとして分類している。

その他、「more」「トラックバック」「更新」などは、ブログや掲示板で全文を読むためのリンクや更新日時を示す等のために記事毎に使われるテンプレート的な部分で使われる語である。これらのテンプレート部分によるクラスタが 6 つ作られた。

以上は Web ページの更新部分に含まれる話題の傾向を把握する目的に沿わないが、できあがったマップの面積の約 8 割を占めている。

残りの部分は記事中の固有表現を元にクラスタが形成されている。この中には、「日本」「向上」等広い概念や「当社」など代名詞を最も大きい成分とする参照ベクトルを持つクラスタも含まれており、それに属する文書には互いに記事内容の関連は見られなかった。記事内容に関連が見られる文書が集まるクラスタは「JAVA」「ショップ」「ビジュアル」「シングル」「collection」を参照ベクトルの大きな成分とする 5 クラスタであった。



記事	JAVA、collection、ビジュアル、シングル、日本、言葉、ショップ、当社、解決、残り、情報
日付表現	3 月 31 日、MARCH 31,2005、31 日、(木)、今日、午後
広告	PR、オリジナル画像、動画、無料、CLICK、お金、応募、公開、ケータイ、モバイル、[商品名：2 件]
テンプレート	more、更新、トラックバック、TRACKBACK、コメント、ブログ

図 1：更新情報の自己組織化マップ

3.1. 課題

同じ時点で取得した差分データ群中では、その時点の日付がブログやニュースサイトなどに含まれていることが多い。また、日時の表現方法はバリエーションが少ないため、差分取得日時を示す同じ日時表現が適当な割合で複数の差分データに現れ、日時表現に関する固有表現によるクラスタが多く作られていた。

広告については、「無料」等、広告に好んで用いられる単語が存在し、異なる広告間でも適当なクラスタを形成する程度に複数の差分データに同じ語が含まれていた。また、商品名によるクラスタは、同じ広告が複数のサイトにたまたま表示されているためであった。これは、ランダム表示の広告の場合は毎回広告部分が差分データとして得られることによりデータ数そのものが相対的に多くなり、偶然同じ広告が載った差分データが適当な数に達しているためと思われる。

「トラックバック」等のテンプレート部に使われる語も、近年のブログの増加により常にある程度の割合で更新情報に含まれている。

これらの語は、以上述べたような各々の理由により、語の出現頻度がクラスタを形成するのに適当な値になっていると考えられる。このため、処理対象となる差分データ集合中での tf や df 等の統計量で、話題の分類に使用したい語と分離する試

みは、良い結果は得られなかった。

4. 提案手法

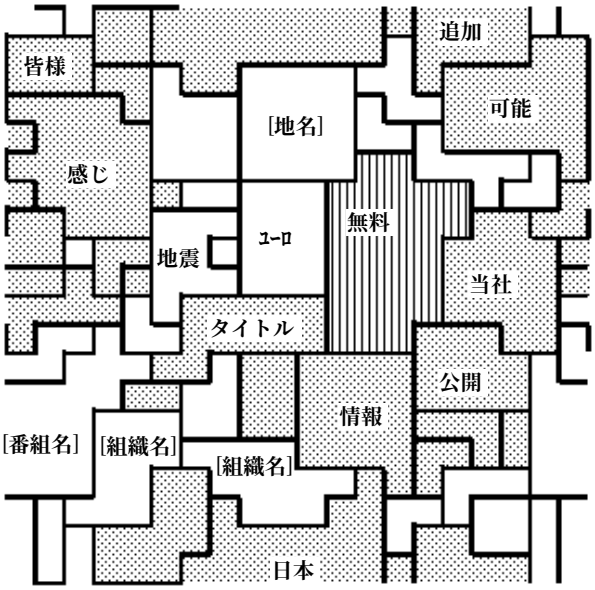
4.1 パターンによる語の削除

問題となる語のうち、日付の表現はバリエーションに限りがある。このため、固有表現抽出を行い、実験データを取得した時点に近い日付表現と「今日」などの現在に近い相対表現を特徴ベクトルの要素から削除した。

広告はリンク先を示す HTML の<a>タグに特徴があるものや、特定の語が使われる物が多い。この実験では、

- <a>タグのリンク先の cgi 等に、パラメータとして URL が与えられている場合
- 行頭に「【PR】」等、「PR」またはそれに括弧が付与されている文字列がある場合。

に、そのアンカー内や行を処理対象から外した。



記事相関有	JAVA、collection、ショップ、シングル、生活、体験、地震、ユーロ、逮捕、解消、エロイカ、ペット、[番組名]、[組織名：5 件]、[地名：3 件]
記事相関無	重要、皆様、掲載、ところ、感じ、コメント、関係、誕生、製品、メール、海外、日本、タイトル、充実、終了、情報、更新、最大、追加、可能、当社、公開、対応、発表、受付
広告	無料

図２：提案手法の結果

4.2. 時間方向の語の出現頻度による補正

テンプレート部に含まれる語についてはブログや掲示板以外にも存在すると思われ、そのバリエーションは無数にあり、ストップワードとして手で登録をすることが困難である。また、例えば「トラックバック」という語であっても、トラックバックに関して議論する記事であれば、特徴ベクトルの要素として必要であり、ストップワードとして全て排除することにも問題がある。

一方、ある Web ページの毎回更新される部分にテンプレート的に使われる語であれば、その Web ページの更新差分にほとんどの場合に含まれていることが期待できる。すなわち、あるサイトについて、過去ある程度の期間に得られた複数の差分データに、その語が含まれる頻度で判定できると期待できる。

今回の実験では処理対象とする差分データ d に現れる固有表現 t について、d が得られた Web ページ p からさらに過去に得られた n 回分の差分データ中に、t が含まれる頻度 tdf(p,t)を用いた。これらから得られる $1-\left(tdf\left(p,t\right)/n\right)^{\alpha}$ を tf・idf に乗ずることによって、過去に同じ固有表現が使われている頻度が高いほどその成分が小さくなるように補正した。α は経験的に求め、0.4 とした。

この処理は、同じページから得られる複数の差分データに、時間が経過しても繰り返し現れる固有表現を抑制することを目的としている。そのため、広告が無作為に含まれる場合も、そのバリエーションが少ない場合は同じ広告が差分データに出現する頻度が高くなり、この処理によりある程度抑制できることも期待できる。

4.3. 実験結果

4.1 節の処理の後に 4.2 節の処理を行う構成で実験を行った結果を図２に示す。図１と比較すると、固有表現レベルで削除した日時によるクラスタは無くなり、テンプレート部分によるクラスタも現れていない。広告に関しては「無料」を大きな成分とするクラスタが作られているのみである。

「コメント」など一見テンプレート部分の語に起因したように思われるクラスタもあるが、そこに集まる文書を観察すると、記事中の文章で「コメント」という語が使われている物であった。このため、図２ではテンプレートに起因するものとはしていない。

なお、図２では、記事中の単語で作られたクラ

スタではあるが、集まった記事の内容に互いに関連がないものは「記事相関無」と分類して図示している。このようなクラスタには、「日本」「情報」など広い概念の語の成分が強い参照ベクトルを持つものが多かった。

5. 考察

4.2 節の処理の替わりに、ブログや掲示板で用いられるテンプレート的な 38 語をストップワードとして処理を行うと(図３)、ストップワードとしなかったテンプレート的な部分の語によるクラスタが残る。例えばここでテンプレート的な語とした組織名は、ニュースサイトでニュースソースとして付記されている新聞社名である。このような語は他の場所では話題語となることもあり、ストップワードとすることはできない。

また、「コメント」などはストップワードとしてしまったためクラスタが作られないが、4.3 節で述べたように記事中で使われる場合もあり、提案手法(図２)ではこれによるクラスタが形成されている。

このような面で、時間軸方向の語の出現頻度による評価値はストップワードによる解決より有効と思われる。

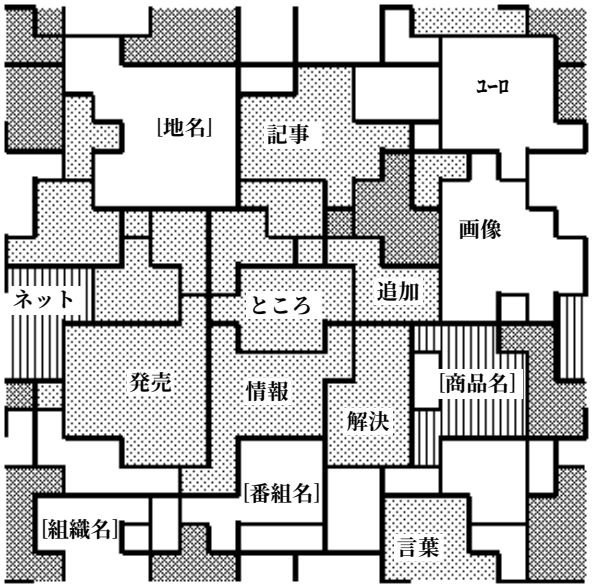
互いに関連の無い差分データの集合となったクラスタは、広い概念の語や一般的な語によるものが多い。マップのサイズが小さい場合にそのような現象が見られることが報告されているが[4]、今回の実験ではクラスタ中の記事の内容に互いに相関があまり無いことと、参照ベクトルとの類似度が小さいことから、クラスタを形成するような類似した記事が他に無い差分データの集合により作られたクラスタではないかと予想している。

今後、これらのカテゴリの判定方法の開発や、より大きなコーパスの df による一般語の抑制などの手法により解決したい。

6. まとめ

固有表現による日時表現の除去やパターンによる広告の除去、及び、時間軸方向での語の出現頻度による評価値により、Web ページの差分データを自動クラスタリングする上での課題を解決できることが確認できた。

また、テンプレート的に用いられる語に関しては、ストップワードによる抑制よりも、ページごとの時間軸方向での語の出現頻度を用いたほうが、より妥当なクラスタが作成された。



記事相関有	JAVA、ビジュアル、シングル、ショップ、野球、ゲーム、ユーロ、画像、桜、データ、引用、[地名：5 件]、[組織名：5 件]、[番組名]、[人名：2 件]
記事相関無	日本、残り、実現、提供、発売、対応、事情、記事、海外、可能、ところ、情報、追加、解決、感じ、言葉
広告	ネット、[商品名]
テンプレート	総数、VIDEO、日記、表示、20 件、出発予定、[組織名]

図３：ストップワードによる結果

参考文献

[1] 川北泰広, “MAILPIA® メールで受け取る情報収集支援サービス”, 沖テクニカルレビュー Vol.71, No.4, pp 38-41 (2004)

[2] T.Kohonen, “The Self-Organizing Map”, Proceedings of the IEEE Vol.78, No.9, pp.1464-1480(1990)

[3] 大沼宏行, 松平正樹, 渕上正睦, 森田幸伯, “コンテンツの分析に基づくオントロジ構築および属性抽出の試み”, 情報処理研究報告 2003-NL-157, 2003-FI-72 pp.49-52 (2003)

[4] 中村順一, 甲斐郷子, 村井幸一, “自己組織化マップによる WWW 日本語情報検索システムの評価結果”, 言語処理学会第３回大会予稿集 pp.365-368