新聞記事と blog からの動向情報の抽出と可視化

奥田 奈央[1]　難波 英嗣[1]　奥村 学[2]
1. 広島市立大学 情報科学部
2. 東京工業大学 精密工学研究所

1. はじめに

電子化された情報が膨大に存在する現在，ユーザが必要とする情報に効率的にアクセスするための技術が求められている．このような技術のひとつとして，日経平均株価や内閣支持率などに関する複数の新聞記事から，動向情報を抽出し，グラフとして提示する手法が提案されている．ここで，時間と共に変動するような情報のことを動向情報と呼ぶ．複数文書の内容をグラフ化するアプローチは，従来の複数文書要約のように複数文書の内容をひとつの文書としてまとめるアプローチと比べ，直感的にわかりやすいものであるが，グラフを見るだけでは，「なぜ数値が上がったり下がったりしているのか(要因分析)」，「数値の推移が社会にどのような影響を与えているのか」，「世の中の人はどのように受け止めているのか」といったことはわからない．しかし，これらの情報は，動向分析を行う上で，非常に重要な情報であると考えられる．

本研究では，新聞記事や blog から動向情報を抽出し，可視化を行う．新聞記事を対象にした動向情報の抽出は MuST ワークショップ[加藤 2004]ですでに行われているが，本研究では新聞記事だけでなく blog にも拡張する．新聞記事中の数値情報の周囲に書かれている情報には，動向分析に有益なものが少なくない．パソコンの販売台数に関する新聞記事を例に挙げると，記事中に販売台数の具体的な数値と一緒に書かれた“販売台数の増減の要因についての記述”がそれに当たる．しかし，新聞記事，特に報道記事には，客観的な事実しか書かれていない場合が多い．一方，blog には，客観的な事実だけではなく，blog の著者の意見や考えが多く含まれており，有用な情報源であると考えられるので，本研究では新聞記事だけでなく blog も対象にする．

本論文の構成は以下のとおりである．まず 2 節では blog における動向情報の記述について述べる．3 節では動向情報の抽出について説明する．4 節では提案手法の有効性を調べるために行った実験について述べ，結果を報告する．5 節では本稿をまとめ，6 節で今後の課題について述べる．

2. blog における動向情報の記述

以下に動向情報を含んだ blog の例を挙げる．

> 日本ではレギュラーガソリンの平均価格が136 円/1 リッターと過去最高水準を記録して話題になっていますが，アメリカ本国でも1 ガロン 2.88 ドル（約 84 円/1 リッター）と，これまた最高記録を更新しそうな勢い．2 年と 4 ヶ月でほぼ 2 倍になったアメリカのガソリン価格が，このハマーをはじめとした大排気量大型モデルの販売不振に直結しているのは間違いないみたいですね．

図において，下線部が動向情報を，波線部が主観的な情報を，それぞれ示す．このように，blog には，事実だけではなく，主観的な情報も含まれている場合が多いので，動向情報と主観的な情報を結びつけるときに，blog は有効であると考えられる．

動向情報が書かれている blog の情報源が独自のコンテンツであるか，また，これまでに我々が新聞記事を対象に構築してきたシステムが blog にどのくらい適用できそうかを調べるため，動向情報を含むいくつかの blog を調べてみた．前者の調査について表 1 に結果を示す．

表 1　新聞記事の引用方法別の blog の分類

	割合〔%〕
新聞記事全体を引用している blog 数	25.8 (80/310)
新聞記事の一部を引用している blog 数	4.8 (15/310)
新聞記事を引用していない blog 数	50.0 (155/310)
情報源を特定できない blog 数	19.4 (60/310)

表 1 より，blog に含まれる動向情報は，新聞記事から引用されたものよりも，blog 独自のものの方が多い結果となった．

後者の調査に関して，動向情報に関する blog は，新聞記事の引用の有無にかかわらず，文体が比較的堅いので，今までの新聞記事用のシステム

をそれほど大幅に変更することなく適用できるものの, ある程度は blog 用に拡張する必要がある. 次節では，その改良方法について説明する.

3. 動向情報の抽出

本節では, 新聞記事と blog からの動向情報の抽出について説明する．動向情報の抽出には時間情報の抽出と数値情報の抽出という 2 つの処理が必要となる．3.1 節では時間情報の抽出，3.2 節では数値情報の抽出について，それぞれ説明する.

3.1 時間情報の抽出

ある統計をグラフ化するとき，数値の情報とその数値がいつのものなのかという情報が必要になる．ここでは，後者を時間情報と呼ぶ.

本研究では, 時間情報の抽出に cabocha[1]を用いる．なお,「～以来」や「～ぶり」といった時間の相対的な差異を表す表現(相対表現)は，本研究では処理対象としないため除去する．以下に本研究で除去する相対表現を示す.

除去する相対表現（時間）

～以降，～後，～前，～連続，～ぶり， ～以上，～越し，～以来，～末，～目， ～続，～より，～比，～ごと，～毎

次に, 抽出した日付情報の補完を行う. 例えば,「昨日」が時間情報として抽出されている場合, テキストの書かれた日付から,

<DATE>2007-1-25</DATE>

といった形式に変換する．また,「昨年 12 月」のように日付が明記されていない場合には

<DATE>2006-12-??</DATE>

と変換する.

新聞記事，特に報道記事では，記事の冒頭部で重要事項を書くのが一般的であるため，これまでに我々が構築してきたシステムでは，新聞記事の冒頭部のみを処理対象としていた[難波 2005]. しかし, blog では必ずしもそういった書き方をするとは限らないので，動向情報を記事全体から抽出する必要がある．この場合，記事の冒頭部には出てこない「同年」や「同月」といった表現も補完しなければならない.「同年」といったような時間情報は,「2007 年の～は～，同年の～は～」のような形式で出現することが多い．このような時間情報は，1 つ前の日付タグに注目すれば補完できると考えられるため，1 つ前の日付タグの YYYY，MM，DD をそれぞれ記憶しておく．なお，初期値は，新聞記事あるいは blog が記述された年月日をそれぞれ記憶しておくこととする.

[1] http://chasen.org/~taku/software/cabocha/

3.2 数値情報の抽出

本研究では，数値の直後に単位が出現する個所(<NUM>**</NUM><UNIT>**</UNIT> 形式の個所）を数値情報の候補と考える．<NUM>タグや<UNIT>タグも cabocha の解析結果をもとに付与していく．<NUM>タグは，cabocha の解析結果で，品詞が「名詞，数」の形態素に付与する．一方，<UNIT>タグは品詞が「名詞，接尾，助数詞」の形態素に付与するが，例えば「本塁打」のように一般名詞であっても「三本塁打」のように助数詞として使われるものもある．そこで，数字(<NUM>タグ)の直後に名詞(<NP>タグ)が出現した場合，その名詞を単位と考え，<NP>タグを<UNIT>タグに置き換える．置き換えた<UNIT>タグの直後に「～増」や「～減」といった相対表現が出現している場合，それらの数値情報を本研究では処理対象外にしているため，除去しておく.

数値情報も時間情報と同様に，記事全体から抽出する必要がある．また，一文中に複数の数値情報がある場合，それらを 1 つずつ抽出していかなければならない．この問題を解決するために，数値情報の抽出では cabocha の係り受けを用い，文節が隣接しておりなおかつそれらが直接係り受け関係にあるものに限り文節を統合する．以下に文節統合の例を示す．図において，“id”は文節の ID を，link は文節の係り先の文節の ID を示す.

```
<chunk id="0" link="1">今日の</chunk>
<chunk id="1" link="2">レギュラーガソリンは</chunk>
<chunk id="2" link="4">130 円，</chunk>
<chunk id="3" link="4">ハイオクは</chunk>
<chunk id="4" link="-1">150 円だった. </chunk>
```

↓

```
<chunk>今日のレギュラーガソリンは 130 円，</chunk>
<chunk>ハイオクは 150 円だった．</chunk>
```

本研究の動向情報抽出プログラムでは，数値情報の抽出をキーワードの文字列一致で行う．例えば，内閣支持率を知りたい場合「内閣支持率」といったものをキーワードとしてシステムに入力する．最終的に，統合された文節の中で，キーワードと数値情報を含んでいるものを全て出力する．上の例で，キーワードを「レギュラー」とす

ると，「130 円」の方だけが抽出される．

また，出力時にはあらかじめ指定した単位を含む数値情報だけを抽出する．以下の例では，<DATE>タグ，<NUM>タグ，<UNIT>タグ以外は省略している．

例（毎日新聞 1999 年 6 月 24 日）

石油情報センターが２３日発表した給油所石油製品市況調査によると，６月のガソリン価格は全国平均でレギュラー１リットル当たり９２円となり，前月比で２円上昇した．

↓

石油情報センターが<DATE>1999-6-23</DATE>発表した給油所石油製品市況調査によると，<DATE>1999-6-??</DATE>のガソリン価格は全国平均でレギュラー<NUM>1</NUM><UNIT>リットル</UNIT>当たり<NUM>92</NUM><UNIT>円</UNIT>となり，<DATE>1999-5-??</DATE>比で<NUM>2</NUM><UNIT>円</UNIT>上昇した．

この例で，キーワードを「レギュラー」，単位を「円」で実行すると，次の結果が出力される．

<DATE>1999-6-??</DATE><NUM>92</NUM><UNIT>円</UNIT>

4. 実験

3 節で述べた手法の有効性を調べるために実験を行った．

4.1 実験に用いるデータ

本研究では，新聞記事は MuST コーパスの 1998 年～1999 年のデータを，blog は blog 検索システム blogWatcher[2][南野 2004]の検索 API で集めた 2006 年のデータを用いる．

MuST コーパスには 27 のトピックに関する新聞記事集合があり，その中から人手で，blog と統合できそうな 8 トピックを選んだ．選択したトピックとそれらのトピックに関して検索クエリを人手で与えて収集した blog データを表 2 にまとめる．この 2 つが実験で用いるデータである．

4.2 実験方法

本システムの入力は，データセットとキーワードの 2 つである．各トピックのデータセットに適する任意のキーワードを入力し，時間情報と数値情報を抽出する．このときのキーワードは，例えば PoliticsTrend のトピックであれば，「内閣支持率」や「政党支持率」といったものである．

[2] http://blogwatcher.pi.titech.ac.jp/

表 2　MuST コーパスのトピックと blog データ

トピック名	新聞 記事数	blog エントリ数
BeerIndustry (ビール業界)	22	68
CarProduction (自動車生産)	16	265
CommunicationDevice (通信機器)	26	225
Gasoline (ガソリン)	20	578
Movie (映画)	6	192
NikkeiStockAverage (日経平均株価)	37	195
PersonalComputer (パソコン)	20	57
PoliticsTrend (政治動向)	17	108

抽出結果から，以下の尺度で再現率と精度を求める．

$$\text{再現率} = \frac{\text{システムが抽出した正解数}}{\text{人手で抽出した正解数}}$$

$$\text{精度} = \frac{\text{システムが抽出した正解数}}{\text{システムが抽出した全件数}}$$

4.3 実験結果

新聞記事と blog から得られた結果から，それぞれ時間情報のみ，数値情報のみ，時間情報と数値情報の 3 つにわけて，再現率と精度を求める．以下に実験結果を示す．

表 3　新聞記事からの抽出結果

	再現率〔%〕	精度〔%〕
時間	6.3 (31/491)	31.3 (31/99)
数値	19.6 (96/491)	97.0 (96/99)
時間 数値	6.3 (31/491)	31.3 (31/99)

表 4　blog からの抽出結果

	再現率〔%〕	精度〔%〕
時間	55.5 (239/431)	74.7 (239/320)
数値	59.9 (258/431)	80.6 (258/320)
時間 数値	44.8 (193/431)	60.3 (193/320)

表 3 からもわかるように，新聞記事からの時間情報抽出の再現率と精度が極端に低いので，新聞記事に対する日付タグ付与について調べた結果を表 5 に示す．ここで，補完率とは，日付タグの中の「YYYY-MM-DD」が正しいかどうかの割合であり，次の尺度で求めることとする．

$$\text{補完率} = \frac{YYYY-MM-DD\text{が正しい付与数}}{\text{システムが付与した正解付与数}}$$

表 5　新聞記事に対する日付タグの付与結果

再現率〔%〕	精度〔%〕	補完率〔%〕
94.1 (985/1047)	85.7 (985/1149)	52.2 (514/985)

4.4 考察

実験結果から，新聞記事からの動向情報の抽出よりも，blog からの動向情報の抽出の方がうまくいっていることがわかった．時間情報の抽出に関しては，blog では「同月」や「同年」といった補完を必要とする表現があまり出てこず，ほとんどの場合 blog の書かれた日付が正解となっていたため，新聞記事よりも精度がよかったと考えられる．また，数値情報の抽出に関しては，新聞記事でも blog でも精度が高いこともわかったが，システムの抽出数自体が少ないので，再現率は低くなっている．

新聞記事に対する日付タグの付与に関して見ると，再現率と精度は高かったが，補完率は低かったので，このことが新聞記事からの動向情報の抽出に影響していると考えられる．日付タグ中の補完の失敗例の 1 つを以下に示す．

失敗例（毎日新聞 1999 年 3 月 12 日）

総市場のシェアは，キリンビールが３８・５％で首位を守り，前月はキリンと１・３ポイント差に迫ったアサヒは３４・７％．

↓

総市場のシェアは，キリンビールが<NUM>38.5</NUM><UNIT>%</UNIT>で首位を守り，<DATE>1999-2-??</DATE>はキリンと<NUM>1.3</NUM><UNIT>ポイント</UNIT>差に迫ったアサヒは<NUM>34.7</NUM><UNIT>%</UNIT>.

この例は，下線部の補完を間違っている．これは 1999 年 3 月 12 日の記事であるが，書かれている内容は 1999 年 2 月のビール・発泡酒の課税出荷数量のことなので，「前月」は「1999 年 1 月」を指す．

このような失敗とは別に，「○年度」，「四半期」，「○，○の両日」といったような部分でも補完に失敗していた．しかし，このような補完の失敗は，時間情報の抽出における例外処理として，新たなルールを加えることで解決できると思われる．

5. おわりに

本研究では，新聞記事と blog からの動向情報の抽出を試みた．新聞記事では全体的な再現率及び精度は低く，blog の方が高かった．新聞記事の場合，特に時間情報の抽出に関して，抽出すべき個所自体は特定できているものの，「同月」や「同年」といった補完を必要とする表現が多く出現し，その補完に失敗するケースが多かった．これに対し，blog ではこのような表現がほとんど出現していなかった．一方，数値情報の抽出精度に関しては，再現率は十分ではないものの，精度に関しては新聞記事で 97%，blog で 80%と，高い値が得られた．

6. 今後の課題

今後の課題として，時間情報の抽出に関する例外処理のルールを付け加えることで，日付タグの補完率を上げ，新聞記事からの動向情報の抽出の再現率及び精度を上げることが考えられる．

また，現在のシステムでは抽出数が少ないので，その点も改善していかなければならないと思われる．

謝辞

本研究では，動向情報の要約と可視化に関するワークショップの MuST コーパスを使わせていただいたことに，深く感謝致します．

参考文献

[難波 2005] 難波 英嗣，国政 美伸，福島 志穂，相沢 輝昭，奥村 学，“文書横断文間関係を考慮した動向情報の抽出と可視化”，情報処理学会自然言語処理研究会，NL-168，pp.67-74，2005.

[加藤 2004] 加藤 恒昭，松下 光範，平尾 努，“動向情報の要約と可視化に関するワークショップの提案”，情報処理学会自然言語処理研究会，NL-164，pp.89-94，2004.

[南野 2004] 南野 朋之，鈴木 泰裕，藤木 稔明，奥村 学，“blog ページの自動収集と監視”，人工知能学会論文誌，Vol.19，No.6，pp.511-520，2004.