E-013

数値による新聞記事テキストマイニングシステムの提案

Mining News Articles by Numbers

杉浦 隆博 † 吉田 稔 ‡ 山田 剛一 † 増田 英孝 † 中川 裕志 ‡
Takahiro Sugiura Minoru Yoshida Kouichi Yamada Hidetaka Masuda Hiroshi Nakagawa

1 はじめに

計算機の処理能力の向上，またネットワーク環境の普及に伴い，ユーザが利用可能な情報は増化の一途を辿っている．これに伴いユーザの関心や興味と合致する情報を，より直観的かつ簡易に提示するための技術が求められている．本稿では，そのような技術の一つとして，数値によるテキストマイニングシステムを提案する．

数値を対象としたテキストマイニングとしては，動向情報を対象とした研究が挙げられる．動向情報とは，商品の価格や売上高，内閣支持率などのように，時系列変化に伴って変動する統計量のことである．近年，これらの動向情報を対象とした複数文書要約や可視化に関する研究が活発化している [1][2].

動向情報のほとんどは数量表現によって表すことが可能であり，数値情報を文書中から取り出すことで，文書検索のランキングにおける新しい軸の一つとして数値を用いる事が可能になる．例えば，検索結果を「電機メーカーの売上高」の順に提示することができ，各記事から電機メーカーの売上高に関する詳細な情報を得ることができる．数量表現抽出の研究では，係り受け構造と優先規則による抽出規則に基づく抽出方法 [3] が存在し，数量表現と対応する事柄の抽出に関して高い再現率と適合率を得ている．

しかし，新聞記事などの文章中に出現する数値について知識を得ようとする場合，その数値単独ではなく，他の数値との関係を知ることが重要となる．例えば，ビールメーカの出荷額について「キリンが３００億」「アサヒが５００億」だったという数値情報を，それぞれの単独の値として扱うのではなく，両者を同時に提示することにより，相対的にそれらがどの程度の値なのかの理解を得ることができる．また，例えば，「企業の営業利益」について，営業利益の額そのものに加え，それが過去の値と比べてどの程度「増加」あるいは「減少」しているのかの情報や，「経常利益」や「売り上げ」等，他の値と比べてどの程度の額なのか，といった情報を同時に提示することにより，より俯瞰的な視点が得られる．

本研究ではこのような問題意識に基き，従来の数値情報とその統計量名の抽出に加え，「相対表現の抽出」と，「複数の数値表現の比較」を可能とするシステムを提案する．前者は，「１０％増」「(全体の) ３０％」等，言語によって他の数値との相対的な関係を表現した部分を提示するものであり，後者は抽出した数値と統計量を用い，特定の統計量に基く記事ランキングや，統計量同士の関係性を視覚的に表現したグラフの自動作成を行うものである．

† 東京電機大学大学院工学研究科, Graduate School of Engineering, Tokyo Denki University Graduate School
‡ 東京大学情報基盤センター, Information Technology Center, University of Tokyo

2 動向情報コーパス

本システムでは「動向情報の要約と可視化に関するワークショップ (略称 MuST) における研究用データセット [4][5]」にある動向情報コーパス (通称：MuST コーパス) を参考にし，数値情報と関連する情報の抽出を行っている．

このコーパスは，各記事に対して，統計量の名前や値，日付などの要素を抜き出し，値に関してはどの統計量のものか，日付に関してはその絶対表現はいつかを記述したものである．以下は毎日新聞の PC 出荷シェアの記事にタグを付与したものである．

```
<unit stat="メーカー毎の PC 出荷シェア">
    <par> NEC など昨年の上位 5 社 </par> の
    <name> シェア </name> は
    <pro ref="前年比" id="9801220"> 同 </pro>
    <rel type="prop">3.1 ポイント </rel> 低い
    <val>82.7%</val>
    となった
</unit>.
```

タグの詳細な仕様に関しては表 1 に記す．本稿では，数量表現に関係する情報として，統計量の名前，値，値の相対表現の自動抽出を行う．

表 1 コーパスで使用するタグの意味

タグ	意味
<unit>	動向情報の統計量や出来事に言及している部分を示す.
<name>	統計量の名前を示す.
<par>	出来事をおこした主体，出来事の一部となる事物など統計量名のパラメータを示す.
<val>	統計量の値を示す.
<rel>	統計量の値の差や順位，比などの相対値を示す.
<pro>	参照表現を示す.

3 数値情報の自動抽出

本研究では，毎日新聞 98 年版 [6] と毎日新聞 99 年版 [7] の新聞記事を対象に数値情報の抽出を試みる．ここでいう数値情報とは，数値と数値に関連する統計量名と相対値のことである．本手法では，まず Cabocha[8] による係り受け解析を行い，その結果を依存構造木の形に変換する．図 1 は，電機メーカの決算に関する文を依存構造木に変換した結果である．

本研究では，この依存構造木を用いて数値と関連する統計量名と相対値の特定を行うことになる．

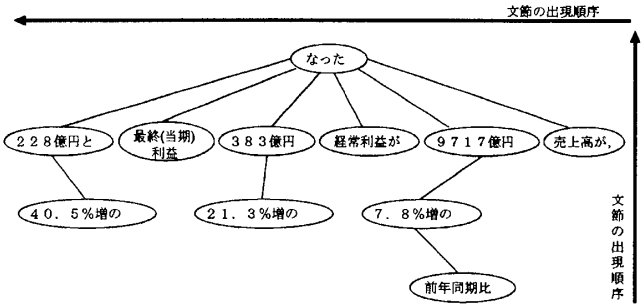

図１　依存構造木

3.1　数値の抽出

新聞記事を構文解析後に，数値とその数値の単位の抽出を行う．ここでいう数値とは統計量の値の候補となるものである．数値の特定には，Cabocha による係り受け解析を行う際に取得する品詞情報を利用する．品詞情報が「名詞－数」である形態素を持つ文節を，数値ノードとする．また，数値の単位に関しては，品詞情報が「名詞－接尾－助数詞」となる形態素の抽出を行う．

3.2　統計量名の特定

統計量名の特定では，統計量名は数値と関係する主語や数値を修飾する文節等になりやすいという仮定に基き，依存構造木中にある全ての数値ノードを対象とし，統計量名の探索と特定を行う．図１に関しては，「９７１７億円」「３８３億円」「２２８億円」「７．８%増」「２１．３%増」「４０．５%増」が対象の数値ノードとなる．まず，数値ノードの子要素が統計量名であるか判定を行う．この子ノードの探索では「売上高は５０００億円」等，直接数値に係る文節を抽出することができる．子ノードが統計量名ではない場合，数値ノードの親ノード方向への統計量名の探索を行う．例として，図１の「３８３億円」の統計量名を探すことを考える．「３８３億円」の子ノード「２１．３%増」は統計量名ではないため，親ノード方向への探索を開始する．

親ノード方向の探索では，親ノードが統計量名であるか判定を行い，統計量名でなければ動詞または文末となるまで親ノードを辿る．親ノードの探索では「５００億円の損益」等，数値が直接係る文節の抽出を行うことができる．動詞または文末となる親ノードが見つかったとき，そのノードに辿り着く直前のノードを C とし，C と同じ深さのノードの探索を行う．「３８３億円」の親ノードは「なった」という動詞であるため，「３８３億円」が C となり，このノードと同じ深さのノードの探索を行う．

C と同じ深さのノードの探索では，C と同一の親ノードを持つノードを対象とし，文中で C より前（図中では右方向）に出現するノードに統計量名があるかの探索を行う．この探索では数値より前に出現する文の主体を抽出することができる．「３８３億円」と同じ親ノードを持ち，「３８３億円」より前に出現するノードは「経常利益が」「９７１７億円」「売上高が」の三つである．このノードの中で最も「３８３億円」に近いノードから順に統計量名の判定を行い，この例では「経常利益が」という主体が「３８３億円」と対応する統計量名となる．

どの数値ノードに対しても，統計量名を特定できた段階で統計量名の探索を終了する．数値ノードに対する探索を最後まで行った時点で，統計量名を特定できていないとき，各ノードの格助詞に着目した探索を再度行う．主格や連体格，対象格となるノードの探索を同じ手順で行い，発見できた時点でそのノードを統計量名とする．どちらの探索でも統計量名を特定できなかった場合は，その数値ノードに関しては「統計量名なし」となる．

統計量名の判定では，数値の単位と対応する統計量名が対象ノードに含まれるか判定し，対象ノードが統計量名を含んでいる場合そのノードを統計量名とする．単位と統計量名の対応情報は毎日新聞の 98 年版と 99 年版から人手で抽出した情報であり，表２が一覧である．

3.3　統計量の相対値の特定

基本的な処理の流れは統計量名の特定と同様である．相対値の特定に関しては，文節中に「%」，「割」，「前期比」などの比率を表す語を含むものに着目し，これを統計量名の相対値とする．

表２　単位と統計量名の対応表

単位	対応する表現
円	売上高，売り上げ，販売額，販売高，利益，損失，消費支出，赤字，連結決算，費用，累損，負債市場，歳入
台	出荷台数，販売台数，生産台数，　自動車生産，国内生産，海外生産，海外販売，出荷実績，累計，加入電話，加入数，規約当事者数，増加数
%	シェア，市場，国内物価指数，状況判断 DI，一致指数，普及率，平均消費支出
単位なし	状況判断 DI，物価指数，卸売物価，輸出入物価
ケース	出荷数量 出荷量 総量 発泡酒 ビール
度	気温

4　抽出結果の評価

4.1　再現率の評価

再現率に関しては，MuST で配布している MuST コーパスを正解データとし，MuST コーパスに含まれる 1998 年から 1999 年までの毎日新聞の 581 記事を評価対象とする．MuST コーパス中にある統計量の値と組となる，統計量名，統計量名のパラメータ，そして統計量の相対値が正しく抽出できたものを正解とする．

MuST コーパス中にある数値情報に関連する抽出結果に対して，MuST コーパスと同様の統計量名，統計量の相対値が完全に抽出できた場合のみ○，不完全に抽出したものを△，抽出できなかったものを×とすることで再現率を評価する．統計量名と統計量の相対値のそれぞれを，「物価」や「内閣支持率」といった MuST コーパスの 27 トピック毎に評価している．

表３が抽出結果全体の再現率の評価結果である．表４は特に評価が良かった統計量名と相対値の上位３トピックの再現率であり，表５は評価結果下位３トピックの再現率である．

評価結果から，統計量名の再現率に関しては「ソニー」，「エアコン」，「商業販売統計」といったトピックが比較的高く，反対に「ガソリン」，「長野五輪」，「物価」といったトピックは低い再現率となった．

また，相対値の再現率に関しては「ソニー」，「商業販売統計」，「百貨店」とったトピックが高く，反対に「ガソリン」，「住宅」，「景気予測」，「総合電機３社」といったトピックは低い再現率になっている．

表 3　再現率の評価

	統計量名	統計量の相対値
再現率	35.1%	28.9%
再現率 (△を含む)	42.3%	29.2%

表 4　上位 3 トピック

順位	統計量名	再現率	統計量の相対値	再現率
1	ソニー	95.0%	ソニー	84.4%
2	商業販売統計	76.5%	百貨店の売上高	83.3%
3	エアコン	76.4%	商業販売統計	82.3%

表 5　下位 3 トピック

順位	統計量名	再現率	統計量の相対値	再現率
1	ガソリン	9.6%	景気予測	3.1%
2	長野五輪	10.0%	総合電機 3 社	11.3%
3	物価	16.3%	住宅	12.0%

4.2　適合率の評価

適合率の評価対象も，再現率と同様の 1998 年から 1999 年の毎日新聞の 581 記事である．適合率では，MuST コーパスでは取り扱っていない数値情報 (例：ビールメーカの各製品の出荷数量等) も評価対象とする．そのため，正解の判定は人手で行い，数値情報に関係する統計量名，統計量の相対値に対して評価する．適合率の評価は，再現率と同様に各トピック毎に評価している．

表 6 が抽出結果全体の適合率の評価結果である．表 7 は特に評価が良かった統計量名と相対値の上位 3 トピックの適合率であり，表 8 は評価結果下位 3 トピックの適合率である．

評価結果から，統計量名の適合率に関しては「ソニー」，「商業販売統計」，「百貨店の売上高」といったトピックが比較的高く，反対に「人口」，「長野五輪」，「為替レート」といったトピックは低い適合率となった．

また，相対値の適合率に関しては「ソニー」，「百貨店」といったトピックが比較的に高いものの，他のトピックに関しては全般的に低い適合率となっている．

表 6　適合率の評価

	統計量名	統計量の相対値
適合率	57.1%	37.8%

表 7　上位 3 トピック

順位	統計量名	適合率	統計量の相対値	適合率
1	ソニー	94.1%	ソニー	70.6%
2	商業販売統計	82.5%	商業販売統計	66.7%
3	百貨店の売上高	80.7%	百貨店の売上高	64.9%

表 8　下位 3 トピック（統計量名のみ）

順位	統計量名	適合率
1	長野五輪	24.7%
2	人口	31.7%
3	為替レート	33.3%

5　評価結果の考察

5.1　統計量名の抽出結果に関する考察

再現率の評価と適合率の評価の両方で，比較的良い評価結果が得られた「ソニー」，「商業販売統計」，「エアコン」といったトピックは，統計量の値に対応する統計量名が「売上高」，「出荷台数」，「販売額」など出現傾向が明らかであった．そのため，単位と対応する統計量名を決定しやすく，良い再現率と適合率を得ることができたと言える．

これに対し，「長野五輪」，「為替レート」，「人口」といったトピックは再現率と適合率が低くなっている．

評価結果が低かった理由はそれぞれのトピック毎に異っており，「人口」に関しては以下のような文構造において，統計量名を確定することができなかったためである．

> ６５歳以上は１９７９年に１０３１万人と１０００万人を上回り，１２年後の９１年には１５５８万人と１５００万人を超えた．

上記の文構造の場合，「１５５８万人」と「１５００万人」という数値情報に関しては正しい統計量名のパラメータである「６５歳以上 (の人口)」が抽出できるが，「１０３１万人」と「１０００万人」という数値情報に関しては，係り受け構造上，本手法では抽出することが困難であり，再現率と適合率が低下してしまった．

「長野五輪」に関しては，以下のような表形式の文構造において，数値情報に関係する統計量名を抽出することが不可能であるためである．

◇国別獲得メダル表◇

	金	銀	銅	計
ノルウェー	5	6	3	14
ドイツ	5	4	4	13
ロシア	5	3	1	9
.....				
	計29	29	29	87

また，上記の表形式の文構造は「長野五輪」以外のトピックでも出現し，その場合でも数値と関係する統計量名を抽出することは不可能である．

そして，「為替レート」に関しては「1 ドル＝１３６円台」といった，文の一文節中に数値情報とそれに対応する統計量名が出現しており，現在の抽出手法ではこのような場合を考慮していないため，評価結果に影響を与えている．

5.2　統計量の相対値の抽出結果に関する考察

統計量の相対値の抽出結果は，再現率と適合率の両方において，統計量名の抽出結果よりも低い値となっている．これは，統計量名の相対値が統計量名とは異り，必ず数値情報と組になるものではないからである．現段階では，数値情報と関係する相対値の有無の判定を行っていないため，統計量名よりも低い評価結果となっている．

6　数値による情報検索システム

本研究では，抽出した統計量の値と統計量名，統計量の相対値を利用して，新聞記事上の数値によるテキストマイニングシステムを実装した．検索対象の新聞記事は、数値情報の自動抽出を行った毎日新聞 98 年版 [6] と毎日新聞 99 年版 [7] である．本システムの機能として統計量名による数値検索機能と，数値による統計量名の出現傾向のグラフ化機能が存在する．

6.1 数値検索機能

数値検索機能では，検索結果を数値によって並び替えることが可能であり，数値の比較や数値の変動について知ることができる．

図 2 は「営業利益」という統計量名に対して，検索を行った結果である．ここでは，上位 3 件はいずれも「ソニー」の営業利益に関する記事であった．検索結果から，「ソニー」の過去最高の営業利益は５２０２億円である，といった情報や「ソニー」の営業利益がどのように変化したのか，といった情報を得ることが可能であり，特に各文中に存在する相対値に着目することで，数値の変動の流れや過去の数値との繋がりを知ることができる．

また，多数のトピックが混在する場合には，複数企業の数値が混在すると考えられるが，これに関してもキーワードによる文書の絞り込みを併用することで対処できる．

6.2 統計量名の出現傾向のグラフ化

統計量名の出現傾向のグラフ化は，数値に対してどのような統計量名が関係するのかを視覚的に提示することで，数値の比較や変動をより直観的に伝えることを目的とする機能である．統計量名の出現傾向のグラフの操作では，まず単位を選択し，その単位に対応する数値の範囲を定めることで，その範囲内の数値と関係を持つ統計量名をグラフ上に MuST コーパスの 27 トピック毎に提示する．このとき，表示するグラフは横軸は文中に含まれる数値を示し，縦軸が各トピックを表している．

図 3 は単位を「円」とし，数値の範囲を１０００億から５０００億にしたときの統計量名の出現傾向である．このグラフからは，「ソニー」（トピック２５）に関する「経常利益」「営業利益」「最終利益」等の額や，「エアコン」（トピック１）に関する総販売額等の様々な額についての相対的な大小関係について知ることができる．図 4 は単位を「円」とし，数値の範囲を５０００から２００００にしたときの統計量の出現傾向である．このグラフからは，「日経平均株価」（トピック２０）の変動の幅について傾向を知ることができる．

また，現在は MuST のカテゴリ分けを使用しているが，将来的にはキーワードによる文書の絞込みや，自動文書分類を併用することにより，一般の文書に適用可能とする予定である．

□ 4502：5202億1000万円
参照記事：　連結段階の売上高は6兆7554億円(19・3%増)、本業のもうけを示す営業利益も5202億1000万円[redacted]いずれも史上最高だった
統計量名：示す営業利益も
[redacted]：(40・5%増)と、

□ 2627：3386億円、
参照記事：テレビ、ビデオなどエレクトロニクス機器の価格競争激化や、昨秋以降の急激な円高が影響し、売上高は前期比0・6%増の6兆7946億円にとどまり、営業利益は[redacted]3386億円、税引き前利益は同18・9%減の3681億円の大幅減益だった
統計量名：営業利益は
[redacted]：同34・9%減の

□ 2355：1794億円で、
参照記事：営業利益も[redacted]1794億円で、減収減益になった
統計量名：営業利益も
[redacted]：同19・8%減の

図 2　統計量名「営業利益」による検索例

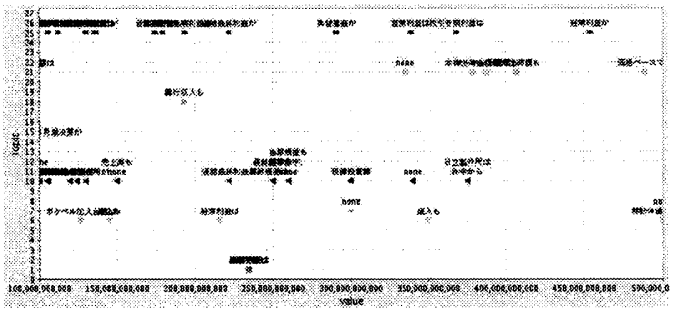

図 3　各トピックにおける経常利益や販売額の出現傾向

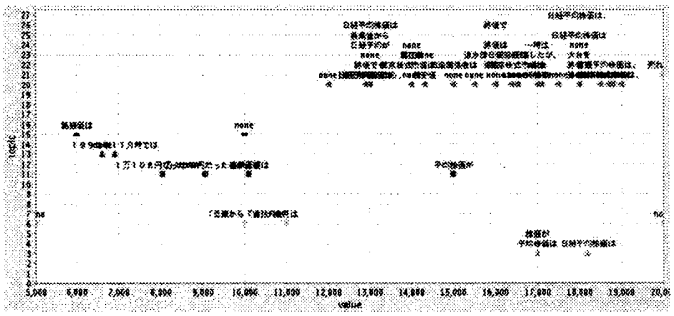

図 4　日経平均株価の出現傾向

7　おわりに

本研究では，数値情報の関係性の抽出を目的とした，統計量名と統計量名の相対値の抽出と新聞記事テキストマイニングシステムの試作を行った．統計量名と統計量の相対値の抽出結果は，一部のトピックのみしか良い再現率は得られなかった．統計量名の抽出に関しては，問題点が明確化されたため，抽出手法の改良を行う必要性があり，統計量の相対値は，統計量値の相対値の有無の判定の実装，そして相対表現の推定の精度の向上が必要である．今後は，統計量名と統計量の相対値の再現率と適合率の向上，そして抽出した相対値による数値の対応付けを可能とするテキストマイニングシステムの実装を行う．

参考文献

[1] 松下 光範，加藤 恒昭，： 動向情報に基づく情報可視化の基礎検討，JSAI2005-1E3-03 (2005).

[2] 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研究報告 2005-NL-168,pp67-74 (2005).

[3] 藤畑 勝之，志賀 正裕，森 辰則：係り受けの制約と優先規則に基づく数量表現抽出，情報処理学会　自然言語処理研究会研究報告 2001-NL-145,pp119-125, (2001).

[4] 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 2004-NL-164(15), pp.89-94, (2004)

[5] 動向情報の要約と可視化に関するワークショップ，http://must.c.u-tokyo.ac.jp/

[6] 毎日新聞社，CD-毎日新聞 98 年版

[7] 毎日新聞社，CD-毎日新聞 99 年版

[8] 係り受け解析器「Cabocha」，http://www.chasen.org/taku/software/cabocha/