社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－NL－168 (11)
2005／7／22

文書横断文間関係を考慮した動向情報の抽出と可視化

難波 英嗣[1] 国政美伸[2] 福島志穂[3] 相沢輝昭[1] 奥村学[4]

1 広島市立大学情報科学部 〒731-3194 広島市安佐南区大塚東 3-4-1

2 中国サンネット 〒730-0036 広島市中区袋町 4-21

3 マイティネット 〒733-0834 広島市西区草津新町 1-21-35

4 東京工業大学精密工学研究所 〒226-8503 横浜市緑区長津田町 4259

E-mail: 1 {nanba, aizawa}@its.hiroshima-cu.ac.jp, 4 oku@pi.titech.ac.jp

あらまし 「日経平均株価」や「内閣支持率」のように数値が時間とともに常に変動するような情報のことを動向情報と呼ぶ．本稿では，動向情報の抽出を一種の複数文書要約であると考え，複数文書要約技術を用いて，あるトピックに関する複数の文書から動向情報を自動的に抽出し，グラフ化する手法について述べる．複数文書からの要約の作成は，様々な要素技術を組み合わせることで実現できる．こうした技術のひとつとして，我々は文書横断文間関係理論(CST)に着目する．CST とは，Radev らが提唱している理論で，文書中の各文の機能を特定し，文間の依存関係を特定する修辞構造理論(RST)を，文書間関係に拡張したものである．本研究では，CST の一部を計算機上で実現し，それを用いてグラフ化に必要な数値情報と時間情報の抽出を行う．

キーワード 動向情報，可視化，文書横断文間関係理論，自動要約

Extraction and Visualization of Trend Information Based on the Cross-document Structure

Hidetsugu NANBA[1] Yoshinobu KUNIMASA[2] Shiho FUKUSHIMA[3]

Teruaki AIZAWA[1] Manabu OKUMURA[4]

1 Hiroshima City University 3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

2 Chugoku Sunnnet Corporation 4-21, Fukuromachi, Naka-ku, Hiroshima, 730-0036 Japan

3 Mighty Net 1-21-35, Kusatsu-shinmachi, Nishiku, Hiroshima, 733-0834 Japan

4 Tokyo Institute of Technology 4259, Nagatsuta, Yokohama 226-8503 Japan

E-mail: 1 {nanba, aizawa}@its.hiroshima-cu.ac.jp, 4 oku@pi.titech.ac.jp

Abstract Trend information is defined as information obtained by synthesis and organization of temporal information such as cabinet approval ratings and stock movements. In this paper, we describe a method for visualizing trend information extracted from multiple documents. We focus on cross-document structure theory (CST) which Radev et al. proposed. The theory expands the notion of Rhetorical Structure Theory (RST) to the relationships between sentences in the different documents. We implement this theory partially, use it to extract trend information, and visualize it as a graph.

Keyword Trend Information, Visualization, Cross-document Structure Theory, Automatic Summarization

1. はじめに

電子化された文書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．そこで近年，自然言語処理の分野で活発に研究されている研究領域の一つに，複数文書要約がある．複数文書要約とは，あるトピックに関する複数の文書の内容をまとめ，要約を作成する技術のことをいう．一般的には，ここで出力する要約とは文章を指すが，ある種の情報は，文章よりもグラフとして出力した方が分かりやすい場合がある．例えば「ある期間の日経平均株価の推移」や「内閣支持率の推移」といった内容は，文章よりもグラフとして提示される方が，ユーザにとって直感的に理解しやすい．ここで，株価や内閣支持率の推移のように，数値が時間ととも

に常に変動するような情報を動向情報と呼ぶ．本研究では，あるトピックに関する複数の文書から，動向情報を抽出し，グラフを自動的に作成するシステムの構築を目指す[1]．

文書集合から動向情報を抽出・グラフ化するには，まず数値情報を抽出し，次にその数値に対応する時間情報を抽出する必要がある．しかし，文書中にはグラフ化する上で必要な数値情報や時間情報と，必要でないものとが混在しているため，両者を区別する必要がある．この処理を行うため，本研究では文書横断文間関係理論(CST)に着目する．CST とは，Radev ら[7]が提唱している理論で，文書中の各文の機能を特定し，文間の依存関係を特定する修辞構造理論(RST)を，文書間関係に拡張したものである．本研究では，CST の一部を計算機上で実現し，それを用いてグラフ化に必要な数値情報と時間情報の抽出を行う．

本稿の構成は以下のとおりである．次節では，文書横断文間関係理論について述べ，3 節では，文書横断文間関係の解析手法について，4 節では，動向情報抽出の手順について説明する．本研究では提案手法の有効性を調べるため，実験を行ったが，5 節では，その実験方法および結果を報告する．6 節で本稿をまとめ，7 節で今後の課題について述べる．

2. 文書横断文間関係理論

文書の文脈や文同士の関係など，文書の構造を談話構造という．これまでに，談話構造を自動的に解析する様々な手法が提案されてきた[5,8]．談話構造は１つの文書内の構造であるが， Radev ら[7]はこれを文書間の関係に拡張しており，24 個の関係を定義している．また，衛藤らは，Radev らの関係を参考に日本語文書に対して 14 種類の関係を定義し，文間及び段落間に関係タグを付与したコーパスを作成している[1]．衛藤らの定義する 14 種類の関係のうち，ここでは動向情報と関連のある「推移」と「更新」という２つの関係を取り上げる．

以下に，「推移」の例を示す．

(1) さらに円高の進行や三井グループによるさくら銀行支援を好感し，日経平均株価は前週末終値比１９２円２６銭高の１万４１０７円８９銭と４営業日ぶりに反発，１万４０００円の大台を回復した．[毎日新聞 98.09.01]
(2) 日経平均株価は前日終値比２１８円３３銭安と続落し，１万４０００円割れ寸前の１万４０４２円９１銭で取引を終えた．[毎日新聞 98.09.05]

[1]近年，動向情報を可視化するという研究課題に対する研究者の関心が集まりつつあり，この課題に関するワークショップも始まっている[4]．

文(1)と(2)を比べると，下線部の株価の終値が 1 万 4107 円 89 銭から 1 万 4042 円 91 銭に変動していることが分かる．このように，「推移」関係にある 2 文には，一定時期に変動する数値が含まれている．数値が変動するという点では「更新」も同様であるが，2 文に含まれる数値の重要性が異なる．「更新」の例を以下に示す．

(3) トルコからの報道によると，同国南部で２７日午後５時(日本時間同１１時)ごろ，マグニチュード(M)６・３の地震が発生し，崩壊した家屋の下敷きになるなどして少なくとも１０７人が死亡，約８００人が負傷した．[毎日新聞 98.06.29.1]
(4) トルコ南部で２７日夕に発生した地震の犠牲者数は２８日夜までに死者１１２人，負傷者１５１７人に達した．[毎日新聞 98.06.29.2]

文(3)と(4)では，下線部の死亡者が「107 人」から「112 人」に，負傷者が「約 800 人」から「1517 人」に増えている．「更新」の関係にある 2 文では，前の文に比べ後の文(ここでは文(4))の数値が正確な情報であるため，後者のみ重要である．

動向情報の抽出においては，「推移」と「更新」のうち「推移」関係が重要となる．しかし，両者には，同じ単位(例えば文(3)と(4)の場合「人」)を持つ数値情報を含み，2 文間で同じ単語を数多く含むという共通した性質がある．そこで，本研究では，まずこのような共通の特徴を持ち，かつ，ある程度内容の似ている(同じ単語を数多く含む)２文の対を更新あるいは推移の候補として検出したのち，次節で述べる方法を用いてそれらの関係が「推移」であるか「更新」であるかについて判定を行う．

3. 文書横断文間関係の解析

本節では，「推移」と「更新」の検出方法について述べる．まず 3.1 節では「推移」と「更新」の特徴についてそれぞれ説明する．次に 3.2 節では「推移」であるか「更新」であるかの判定方法について述べる．

3.1. 更新と推移

3.1.1. 更新の特徴

更新には，以下に示す例文の下線部のように，数字の後に「～となる」，「～目」といった表現が多く含まれる．

- 負傷者は１５１７人となった．
- 今回の調査で身元が確認されたのは４人目．
- 発生した地震の死者は少なくとも１１６９人に上った．

このような表現の有無が，更新であるかどうかの判定に利用できると考えられる．そこで，上記のような手がかりとなる語を，以下に述べる手順で収集した．

まず上記の3例のうち，2つ目の「4人目」という表現に着目した．更新には「4人目」のような「数値」＋「単位」＋「目」という表現が頻出する．このパターンは更新の有力な手がかりとなりうる．しかし，中には「一丁目」のような住所の一部もこのパターンに当てはまるため，上記のパターンに当てはまるものすべてを更新として判定するわけにはいかない．そこで上記のパターンに当てはまる様々な単位を収集し，その中から判定に使用できる単位を選別した．具体的には，1999年の毎日新聞記事から数値の直後に表れる1～4バイトの文字を単位として検出し，そこから使用できない単位を削除した．削除した表現には，「日本一を目指す」の「を」や「一際目立って」の「際」などがある．最終的に，95種類の単位が得られた．以下にその一部を掲載する．

回　年　度　日　勝　人　番　例　代　種　期
つ　点　枚　球　戦　作　試合　敗　件　場所

次にこれらの手がかり語と共に良く現れる表現で，更新を検出する上で有用なものを52種類選定した．以下にその一例を示す．

となる．　になる．　に突入する．を迎える．
に当たる．　に上る．　を喫した．

3.1.2. 推移の特徴

推移には，以下の例の下線に示すように，数値の相対的な差異を表す表現や，数値の変動を示す表現が出現することが多い．

- 日経平均株価は前日終値比２１８円３３銭安と続落し，…
- ９７年度末に比べ３６万４０００台減の６３６万３０００台となった．
- 気象庁は1日，4月の平均気温が全国的に平年を上回り，…

このような表現を，本研究では相対表現と呼ぶことにする．我々は，相対表現を新聞記事から人手で以下に示す26種類を選定した．

強　弱　アップ　ダウン　高　安　台　代
増(増加，増し)　減(減少，減り)　伸び　延び
多い(多く)　少ない(少なく)　上回る　下回る
(を)割(割る)　当たり　超(え)　長い　短い
以上　以下　(の)大台　拡大　縮小

3.2. 更新と推移の判定手順

2節で述べた更新と推移の特徴を考慮し，以下の手順で「推移」および「更新」関係の文を抽出する．

1.　単位の一致

まず数値に添えられる単位が同じである2文を検出する．茶筌を用いて入力文書を形態素解析し，品詞が接尾-助数詞である接尾辞が，比較する2文に共に出現しているかどうかを調べる．接尾-助数詞とは数値の後に続く語のことで，これが一致していると単位が同じであると言える．このとき，日付と年齢も数値と判断されるが，これらは更新と推移の特徴である数量的な変化ではないので除外する．ただし，日付と年齢以外に，数量的な変化を示す表現が含まれる場合はこれを検出する．

2.　類似文の検出

次に検出された2文の対のうち，内容が似ているものを検出する．対象となる2文から名詞，動詞，形容詞を抽出し，2文間の類似度をコサイン距離で算出し，閾値以上の対を検出する．

3.　更新・推移の判定

最後に3.1節で述べた更新の手がかり語が含まれていれば更新，推移の手がかり語が含まれていれば推移と判定する．なお，両方の手がかり語が含まれている場合は，どちらかを優先する必要がある．この点については6節で述べる．

4. 動向情報抽出の手順

あるトピックに関する動向情報を抽出・可視化するには，まず検索エンジンを用いて，そのトピックに関する文書を収集し，次にそこから日経平均株価等の具体的な数値情報と，それに対応する時間情報を抽出する必要がある．これらの処理のうち，数値情報抽出を4.1節で，時間情報抽出を4.2節でそれぞれ述べる．なお，今回の実験では新聞記事を対象にしているが，より良い精度で動向情報を抽出するため，新聞記事に特化した制約もいくつか組み入れている．その制約について4.3節で述べる．

4.1. 数値情報の抽出

数値情報の抽出は，収集された文書集合に対し文書横断文間関係を解析した結果，「推移」と判定された文から抽出する．しかし，これだけでは目的の数値情報とは異なる情報が抽出される場合がある．例えば，「センター試験出願者数」に関する新聞記事集合を過去数年の新聞記事データベースから収集した場合，これらの記事集合中には出願者数だけでなく受験者数の推移情報も含まれる．このように，文書集合中に推移する

情報が2種類以上存在する場合には，その中から目的の情報のみを選択する必要がある．そこで，本研究では，「推移」と判定され，かつ文書集合を収集するのに用いたキーワードと関連度の高い文から数値情報を抽出する．

相対表現の取り扱い

3.1.2節で述べたとおり，「推移」関係にある文中にはしばしば相対表現が出現する．相対表現を利用することで，より多くの数値情報を得ることができる．例えば，2節の例文(2)から，98年9月4日の日経平均株価が「1万4042円91銭」であることが分かるが，「前日終値比218年33銭安」という表現に着目すると，98年9月3日の終値も算出できる．しかし，すべての相対表現から同様に算出できるわけではなく，その場合分けが現状ではまだ十分に検討できていないため，今回は，相対表現からの数値情報の算出は行わない．

4.2. 時間情報の抽出

本研究では，三重大の桝井らが開発した情報抽出器NExT[6]を使用し，時間情報を抽出する．ここで，以下の例のように，文中に複数の時間表現が出現することがある．

> 厚生省は六日，一九九一年の人口動態統計を発表した．女性が生涯に産む子供の平均数(合計特殊出生率)は一・五三人で，前年の確定データ(一・五四人)を下回って史上最低を更新，「少産ショック」に歯止めがかからなかった．[毎日新聞 92.9.5]

例えば上記の文では，「六日」と「一九九一年」という2つの時間表現があり，どちらを抽出するのか決める必要がある．今回は，どの単位(年，月，日)で時間情報を抽出するのかは，事前にユーザに入力してもらうことを想定している．例えば上の例において，ユーザが年単位での抽出をシステムの入力として与えた場合には，「一九九一年」が抽出される．

時間の省略補完

時間情報には日時の省略してあるものが多く，抽出した時間情報そのものではグラフ上にプロットできないものがある．このような場合，日時の省略を補完する必要がある．以下に，補完が必要な日時表現の例を示す．

> 総務庁と労働省は三十一日，それぞれ昨年一年間の労働力調査（平均）と十二月の有効求人倍率（季節調整値）を発表した．景気の減速傾向を反映して就業者の伸びは後半に鈍化，有効求人倍率も昨年八月以来五カ月連続の低下という結果となった．[毎日新聞 92.1.31]

このような場合，記事の書かれた日付(92年1月31日)から，例にある3つの時間表現はそれぞれ，「92年1月31日」，「91年12月」，「91年8月以来」であると推測できる．

相対表現の取り扱い

時間情報にも，数値情報と同様な相対表現がある．例えば，以下の例では「二十一年ぶり」の「ぶり」が相対表現に該当する．

> 一九九四年の年間出生数は前年より四万七千人も多い百二十三万五千人を記録し，二十一年ぶりに大幅増に転じたことが三十一日，厚生省の九四年人口動態統計年間推計で分かった．[毎日新聞 91.1.1]

この例から年単位で時間情報を抽出する場合，「一九九四年」と「二十一年」の両方に「年」が含まれるため，どちらか不要な情報を削除する必要がある．ここで，「二十一年」の後には「ぶり」という相対表現があるため，「二十一年」が時間情報の候補からはずれ，結果として「一九九四年」が抽出される．

この他，本研究で考慮する時間の相対表現を以下に示す．以下の表現が時間情報の後ろに出現する場合，その直前の時間は相対表現とみなし，除去する．

～連続　～目　～ぶり　～以来　～越し(ごし)
～続伸　～より　～比　～前　～後

この他，「来(年|月|日)」は未来を表すので，相対表現ではないが除去する．また，「前(年|月|日)」という表現は必ず比較時に用いられるので，除去の対象となる．

4.3. 新聞に特化した制約

本研究では，数値情報や時間情報を抽出する際，新聞に特化した制約をいくつかもうけている．まず，対象記事を報道記事に特化している．報道記事は，記事の冒頭で主題とその時間情報を述べ，その後の文で補足説明をする書き方が一般的であるため．記事の後半部から抽出される時間情報はノイズとなる可能性が高い．従って，時間情報は記事の先頭2文以内から抽出する．特に，以下の表現を文末に含む文中の時間表現を優先的に抽出する．

分かった(わかった)　明らかになった　発表した
調べた　報告した　公表した　まとめた
となった　であった

5. システムの構成

システムの概要を図1に示す．グレーで囲まれた個

所が本研究で作成した部分である．なお，文書の収集は人手で行った．

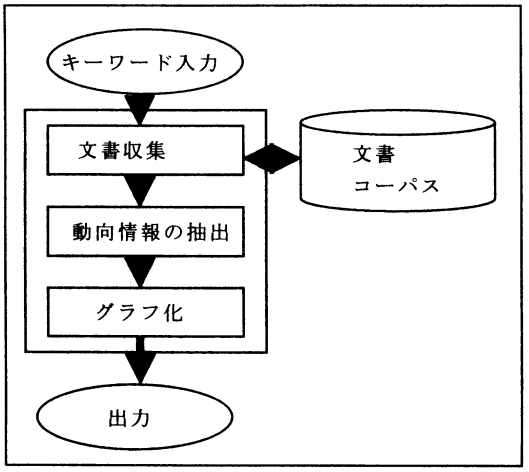

図 1　システム構成図

文書コーパスから入力されたキーワードに関する文書の自動収集を行う．収集した文書集合に対して，以下に示す(1)〜(3)の処理(図 1 グレーの個所)を行う．

(1) 文書集合とキーワードを入力する．入力は文書の収集に用いたキーワード，'単位'，'時間単位'である．
(2) データセットに対して文書横断文間関係を解析し，時間情報と数値情報を抽出する．
(3) (2)の結果を GDGraph[2]という Perl のグラフ作成モジュールに渡し，グラフを出力する．

6. 実験

提案手法の有効性を確認するため，実験を行った．6.1 節では文書横断文間関係解析実験について，6.2 節では動向情報抽出実験について述べる．提案手法で実際にグラフを出力した．その結果を 6.3 で示す．

6.1. 文書横断文間関係の解析

● データセット

NTCIR ワークショップの自動要約タスク TSC2[2]および TSC3[3]の複数文書要約課題で使われたデータに CST タグを付与したデータ 80 トピック分[1]の中から，動向情報に関連のあるものを選び(以後，CST コーパスと呼ぶ)，それらを実験に用いた．文間の関係は同等，簡略など，全 14 種類が定義されているが，この中で推移と更新を対象とする．

● 実験手順

3 節で述べた手法を CST コーパスに適用し，実験を行う．解析方法は 2 節で述べたとおり，まず更新か推移のいずれかにあてはまる候補（更新／推移候補）を検出し，次にそれらが更新であるか推移であるかの判定を行う．評価はそれぞれの段階で行う．

実験 1　更新／推移候補の検出

2 文間の類似度をコサイン距離で測り，閾値を 0 から 1 まで 0.1 ずつ変化させ，更新と推移の候補を検出する実験を行う．評価尺度は以下に示す精度と再現率を用いる．

$$\text{精度} = \frac{\text{システムが検出した文対の中で人手で更新か推移が付与されているものの数}}{\text{システムが更新か推移の候補として検出した文対の数}}$$

$$\text{再現率} = \frac{\text{システムが検出した文対の中で人手で更新か推移が付与されているものの数}}{\text{人手で更新か推移が付与されている文対の数}}$$

実験 2　更新か推移かの判定

3.1.1 節と 3.1.2 節で示した手がかり語を用い，更新であるか推移であるかの判定を行う．評価尺度は以下に示す精度と再現率を用いる．

$$\text{精度} = \frac{\text{人手で更新(推移)が付与されている文対の数}}{\text{システムが更新(推移)と判断した文対の数}}$$

$$\text{再現率} = \frac{\text{システムが更新(推移)と判定した文対の数}}{\text{人手で更新(推移)が付与されている文対の数}}$$

なお，文中に更新の手がかり語と相対表現の両方が現れた場合にどちらかを優先する必要がある．そこで，更新を優先させた場合と推移を優先させた場合の 2 通りで，精度と再現率をそれぞれ求める．

実験 2 では，更新／推移候補が理想的に検出できた場合を考え，CST コーパス中で「推移」または「更新」が(人手で)付与されている文対を入力とし，それらを提案手法で更新か推移か判定する．

● 実験結果

実験 1 更新／推移候補の検出

実験結果を図 2 に示す．

[2] http://search.cpan.org/dist/GDGraph/Graph.pm

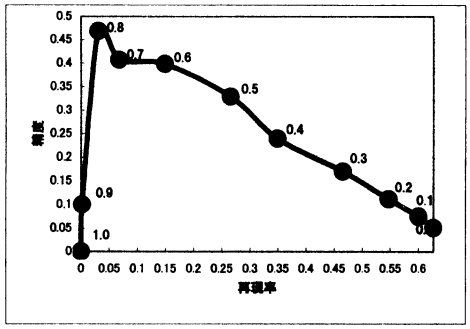

図２　更新／推移候補の検出結果

図２において，グラフ中の数字はコサイン距離の閾値を示す．図から，２文間の類似度が高いほど精度は高くなるが，類似度が 0.8 を超えると急激に精度が低下することが分かる．

実験２　更新か推移かの判定

結果を表１に示す．

表１　更新か推移かの判定結果

		更新優先	推移優先
更新	精度	0.357	0.286
	再現率	0.667	0.267
推移	精度	0.991	0.966
	再現率	0.864	0.896

表１から，推移はかなり正確に判定できているが，更新の判定制度は十分ではないことが分かる．

● 考察

実験１更新／推移候補の検出

２文間の類似度について

閾値を上げすぎると，0.8 を超えたあたりから急激に精度が低下する．これは，「同等」な関係にある２文を検出するためである．従って，極端に類似した２文は候補から外す必要がある．

提案手法で検出できない例

提案手法は，同じ単位を持つ２文を検出対象としているため，単位は異なるが更新あるいは推移の関係である２文を検出することはできない．このような例を以下に示す．

- ＰＨＳは，加入者を１１カ月連続で減らし続けており，８月末も９７年度末に比べ３６万４０００台減の６３６万３０００台となった．[毎日新聞 98.9.8]
- 郵政省が９日発表した１０月末のＰＨＳ(簡易型携帯電話)の累計加入数は，９月末より９万８０００件減って６１６万８０００件となり，１３カ月連続の減少となった．[毎日新聞 98.11.10]

これら２文は推移の例であるが，PHS の加入者を表す単位が前者では「台」，後者では「件」となっており，同じ PHS の加入者数を表すにもかかわらず，異なる単位が使用されていた．そのため，提案手法ではこの２文を推移として検出することができなかった．

実験２　更新か推移かの判定

本システムは更新の手がかり語と相対表現の両方を用いて更新と推移の判定を行うため，両方の手がかり語が含まれている場合，更新と推移のどちらを優先するかで結果が異なる．以下に例を示す．

- トルコ国営アナトリア通信によると死者１１人が確認され，負傷者は数百人に上っており，救援活動が進むにつれ，犠牲者はさらに増えそうだ．[毎日新聞　99.8.17]
- トルコ政府危機管理センターによると，同国北西部で１７日に起きた大地震の被害は，１８日未明までに死者が２１００人を超え，負傷者も１万３０００人に上ったほか，１万人以上が行方不明になっている．[毎日新聞 99.8.18]

この２文は更新の関係であるが，「に上る」という更新の手がかり語と，「増え」，「超」，「以上」などの相対表現の両方が現れるため，更新と推移のどちらの手がかりを優先するかで判定結果が異なった．

また，更新の手がかり語や相対表現が全く現れず，判定ができなかった場合があった．以下に例を示す．

- 九州各地は激しい風雨に見舞われ，午前１０時までの各地の県警の調べでは，熊本県で１４人，福岡県で３人，広島県で４人，大分県で１人が死亡した．[毎日新聞 99.9.24]
- 警察庁の２４日午後１０時現在のまとめによると，熊本県１４人▽広島県５人▽福岡県４人▽山口，岡山，大分県各１人の計２６人が死亡し，宮崎，静岡両県で２人が行方不明となっている．[毎日新聞 99.9.25]

この２文は更新の関係であるが，数値のすぐ後に更新の手がかり語が全く現れず，かつ相対表現も現れな

いので，更新と推移のどちらとも判定されなかった．

6.2. 動向情報の抽出

● データセット

毎日新聞データベース 1991 年～1999 年の中から，26 トピックに関する記事集合を人手で作成した．表 2 にデータセットの一覧を示す．表には，トピック，トピック毎の記事数，推移する数値情報の単位，および時間単位を載せている．時間単位で「月 or 年」のように複数の表記があるものは，同じトピックだが異なる時間単位で文書集合を作成していることを意味する．

表 2　動向情報の抽出に用いたデータセット

トピック	記事数	数値単位	時間単位
出生，死亡数	8	人	年
センター試験出願者数	5	人	年
パソコン売り上げ台数(年，四半期)	12	台	四半期 or 年
就業者，失業者数(年，月)	41	人	月 or 年
ゴールデンウイーク人出数	5	人	年
有効求人倍率	54	%	月
日経平均株価	25	円	日
円相場	21	円	日
平均寿命	5	歳	年
結婚，離婚	9	組	年
PHS 携帯電話の加入数	5	台 or 件	月
新車販売台数(年，月)	60	台	月 or 年
中古車販売台数(年，月)	16	台	月 or 年
子どもの数	4	人	年
小学生のお年玉の金額	9	円	年
レコード大賞視聴率	7	%	年
高齢者数	4	人	年
年男，年女の数	5	人	年
トヨタ自動車のシェア	3	%	年
内閣支持率	34	%	月
一世帯平均年収	4	円	年
新成人の数	8	人	年

● 実験方法

表 2 に示すデータセットを用いて実験を行う．なお，現状では文書横断文間関係解析の精度が十分ではない．そこで，数値情報の抽出は文書横断文間関係解析の結果は使わず，新聞記事の先頭 2 文で文書集合の収集に用いたキーワードとの適合度の高いものから抽出する．この方法では，文書横断文間関係解析で行うような文間で一致する単位の抽出を行わないため，今回の実験では，抽出する数値情報の単位もシステムに与える．

以下に示す再現率と精度を用いて評価を行う．

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

$$精度=\frac{システムが抽出した正解数}{システムが抽出した数}$$

● 実験結果

結果を以下に示す．

表 3　動向情報の抽出精度

	再現率	精度
数値情報	0.848(351/414)	0.862(351/407)
時間情報	0.860(308/358)	0.875(308/352)

● 考察

今回は記事の先頭 2 文から時間情報の抽出を行うという制約を用いた．この方法の有効性は確認できたが，先頭 2 文以内に時間情報が出現しない場合もあった．今後は先頭 2 文だけでなく，キーワードと関連度の高い段落等も抽出対象にすることで，この問題はある程度改善されると思われる．

数値情報の失敗原因の多くは，特殊な表現によるものが多かった．例えば，「結婚，離婚」に関する記事集合中に「2 分半に 1 組離婚」といった表現が使われているものがあったが，このような表現には現在のシステムでは対応できない．また，「出生，死亡数」に関する記事集合中で「出生数」ではなく「一九九三年に生まれた赤ちゃんは百十八万五千人で」という表現が使われているものがあるなど，想定外の表現が使われて失敗するケースがかなりあった．

動向情報抽出実験では文書横断文間関係解析の結果を用いていないが，数値情報，時間情報ともに高い再現率と精度が得られている．しかし，この手法は与えるキーワードが少し変わるだけで再現率と精度が大きく変化することが実験により確認されている．抽出対象の文を単純にキーワードだけで絞り込むのは限界があるため．今後さらに文書横断文間関係解析の解析精度を向上させ，利用する必要があると考えている．

6.3. 動向情報のグラフによる表示

提案手法を用いた抽出されたデータでグラフ化した．図 3 に 1992 年 3 月 1 日～1992 年 3 月 31 日の記事集合から円相場に関する動向情報を抽出しグラフ化した結果を示す．X 軸は年月日，Y 軸は円を表す．図 3 は，動向情報がうまく抽出できた例であり，文章で提示する方法よりも円相場の推移が直感的に分かる．

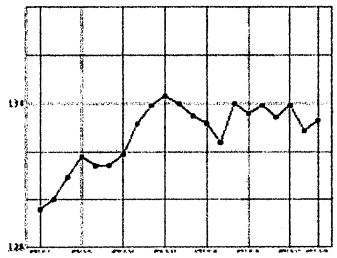

図 3　円相場を表すグラフ

図 4 は，1991 年から 1999 年の月単位での就業者数を表すグラフである．X 軸は「月」，Y 軸は「人数」を表す．このグラフでは正しく抽出されていない個所が数箇所あるため，実際の就業者数の動向が分かりにくい．ただ，図 4 を見れば抽出に失敗していると思われる点が容易に推測できるため，このような点をインタラクティブに削除したり，数値情報を抽出した記事にリンクしたりする機能をシステムに備えれば，ユーザは比較的容易に正しいグラフを得ることができると思われる．

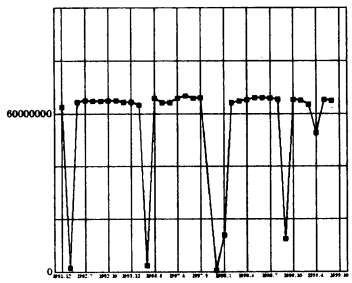

図 4　就業者数を表すグラフ

7. おわりに

本稿では，あるトピックに関する複数の文書から動向情報を自動的に抽出し，グラフ化する手法を提案した．あるトピックに関する文書集合について文書横断文間関係解析を行い，「推移」関係にある文を抽出し，「推移」関係にあり，かつ文書集合の収集に用いたキーワードと関連度の高い文から動向情報の数値を，同時に該当記事の先頭 2 文から時間情報を抽出する．現状では文書横断文間関係解析の解析精度が十分ではないため，今回は文書集合を収集するのに用いたキーワードと適合度の高い文から数値情報を抽出した．実験の結果，約 85%の再現率と精度で数値情報と時間情報の抽出ができることがわかった．

8. 今後の課題

まず文書横断文間関係解析の解析精度向上を目指す．次に文書横断文間関係解析の結果を用いてどの程度動向情報の抽出が可能であるのか調査する．この他，文書から数値情報と時間情報を抽出するだけでなく，「なぜ数値が変動したのか」の理由の記述も文書から抽出し，グラフ上に提示できるよう改良していきたい．

文　献

[1] 衛藤純司，奥村学，“文書横断文間関係タグ付コーパスの構築，” 言語処理学会第 11 回年次大会，2005.

[2] Fukushima, T., Okumura, M., and Nanba, H. “Text Summarization Challenge 2 / Text Summarization Evaluation at NTCIR Workshop3,” Working Notes of the 3rd NTCIR Workshop Meeting, PART V, 1-7, 2002.

[3] T. Hirao, M. Okumura, T. Fukushima and H. Nanba. “Text Summarization Challenge 3 - Text summarization evaluation at NTCIRWorkshop 4,” Working Notes of the 4th NTCIR Workshop Meeting,, 2004.

[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ワークショップの提案，” 情報処理学会研究報告，vol. 2004, no. 108 (2004-NL-164) pp. 89-94, 2004.

[5] D. Marcu, The theory and practice of discourse parsing and summarization, The MIT Press, 2000.

[6] 桝井文人，鈴木伸哉，福本淳一，“テキスト処理のための固有表現抽出ツール NExT の開発，” 第 8 回言語処理学会年次大会発表論文集，pp. 176-179, 2002.

[7] D. R. Radev, H. Jing, and M. Budzikowska, “Summarization of multiple documents: clustering, sentence extraction, and evaluation,” Proceedings of the Workshop on Automatic Summarization, pp.21-30. Association for Computational Linguistics, 2000.

[8] 横山憲司，難波英嗣，奥村学，“Support Vector Machine を用いた談話構造解析，” 情報処理学会研究報告，NL-155, pp. 193-200, 2003.