グラフとテキストの協調による知的な情報提示手法 —日経平均株価テキストとグラフの提示を例にして

小林一郎[†] 渡邉千明[††] 奥村奈穂子[†††]

本研究では，知的情報提示技術開発の一環として，日経平均株価のグラフ（チャート）の表示状態および指定された期間に合わせて，そのグラフと対応した日経平均株価の動向を説明するテキストを動的に生成するシステムの開発を行う．日経平均株価の動向を説明するテキストを生成する際に，その期間のニュースを要約する機能，および，1 日の株価の動向をそのグラフの形状から認識し，言語で説明する機能を提案する．テキスト要約機能に関しては，要約する対象となるニュースとして，国立情報学研究所主導のワークショップ「動向情報の要約と可視化」（NTCIR-5）により提供されている，MuST コーパス中の日経平均株価について集められた本文を利用する．そのテキスト情報と数値情報を連携させることにより，グラフの表示状態と協調して，MuST コーパスから重要文を抽出する手法を用いてテキスト要約を行う．また，テキスト生成機能は，線形最小二乗法を用いて，5 次多項式によるグラフの近似曲線を作り，その近似曲線の振舞いをパターンによってとらえる．コーパスを分析することから得られた日経平均株価の挙動を説明するのに適切な言語表現をあてはめることにより，グラフの挙動を説明するテキスト生成を行う．

Intelligent Information Presentation Based on Collaboration between 2D Chart and Text — With an Example of Nikkei Stock Average Text and Its 2D Charts Presentation

Ichiro Kobayashi,[†] Chiaki Watanabe[††] and Naoko Okumura[†††]

As a study of developing an intelligent information presentation technology, we develop a system with two functions: one is the function of summarizing news articles about Nikkei stock average corresponding to its 2-D chart representation state and generating a text which explains the behavior of 2-D chart of the stock average trends. As the target news article to be summarized, we use an annotated corpus called MuST corpus that is provided by a pilot workshop "A Workshop on Multimodal Summarization for Trend Information (MuST)" (NTCIR-5) hosted by National Institute of Informatics. By linking the news articles of the MuST corpus with numerical data of the stock average, a summarized text is produced by extracting important sentences from the news articles corresponding to its 2-D chart representation state. As for the text generation function, we have applied linear least squares to produce an approximate chart to the original 2-D chart so that we can observe the behavior of the 2-D chart through recognizing the shape of the approximate chart. The shape of the approximate chart is broken down into several patterns of partial shape that are expressed with linguistic expressions extracted from the news articles about the stock average. By finding proper words and short sentences to express the shape of the approximate chart, a text explaining the behavior of 2-D chart is generated.

1. 研究背景と目的

インターネットが普及するにつれ，インターネット上の膨大な情報を利用できる人，そうでない人の格差であるデジタルデバイドという社会現象が起きている．この要因の 1 つとして考えられるのが，インターネットから得られる情報の内容や表示が必ずしも分かりやすくなく，また情報を提供する側において，ユーザが欲しい情報を欲しい形で提供するなどの工夫がなされ

† お茶の水女子大学理学部情報科学科
Department of Information Sciences, Faculty of Science, Ochanomizu University
†† お茶の水女子大学人間文化研究科数理・情報科学専攻
Graduate School of Humanities and Sciences, Ochanomizu University
††† 株式会社 NTT データ
NTT Data Corporation

ていないことがあげられる．本研究では，このような現状をふまえ，情報の内容や表示がだれにでも理解しやすいよう，情報提示の形態を動的に変化させることができる機能を持つ知的情報提示手法を提案する．その一例として，テキストとグラフという異なるモダリティどうしを協調させることにより，大まかな情報を必要とするユーザ，または，詳細な情報を必要とするユーザなど，それぞれのユーザに適した情報を提示する手法を提案する．一般に大まかな情報をテキスト形式で提供する際は，文章の要約技術が，また詳細な情報を提供する際は新たなテキストを生成する技術が必要になる．

具体的に，日経平均株価の動向に対する情報を取り上げる．ユーザは一般的に年月単位の長期的な動向に関する大局的な情報と，速報性を重視する日単位の短期的な動向に関する 2 種類の情報を必要とする．長期的な動向をとらえるための情報源として，株価の日足ベースの始値，最高値，最安値，終値の数値データおよび新聞記事などによる 1 日の株価の動向を伝えるテキスト情報が利用できる．一方，短期的な動向をとらえるための情報源としては，分足データなど観測される数値データは存在するが，グラフの挙動を説明する適切なテキスト情報が存在しない場合もある．これらのことを考慮して，長期的な動向をとらえるために，グラフの表示状態に協調してテキストの情報も変化するテキスト要約手法を提案し，また短期的な動向をとらえるために，視覚的にとらえられるグラフの挙動を説明するテキスト生成手法を提案する．これらの 2 つの手法を用い，**図 1** に示されるような，様々な動向情報を出力することができる知的な情報提示システムを開発する．

以下，2 章でグラフに連動したテキスト要約機能について説明し，3 章ではグラフからのテキスト生成機能について説明する．4 章では，本研究に関連する研究を紹介し，5 章では結論と今後の課題について述べる．

2. テキスト要約機能

2.1 提 案 手 法

テキスト要約機能では，対象となるニュース記事から重要文を抜き出すことにより，ユーザによって変更されたグラフの状態に対応している要約文を生成する．本システムでは，日経平均株価の数値情報と，MuST コーパス（2.1.1 項：対象コンテンツで説明）によって得られるその日の株価の動向情報を対応させ，グラフとテキストを関連付けておく．MuST コーパスと日経平均株価の数値情報の対応関係は，MuST コーパスにおける株価の値に言及しているタグの値より確認できる（**図 2** 参照）．

まず，ユーザに要約文の文数を指定してもらう．そして，全体のグラフと，20 日分の記事から作られた指定した文数からなる要約文を表示する．次に，ユーザがグラフから一部を選択，または目盛り間隔の変更を行うと，選択された範囲や変更されたグラフの詳細度に対応して，MuST コーパス内に記されたニュース本文が要約され，グラフの表示とテキストが協調した情報提示が行われる．要約文は，ニュース本文の中から重要文を抽出することにより生成される．この際，各文の重要度はグラフに対する変更処理に応じて動的に変更され，グラフの表示状態に応じた重要文からなる要約文が生成される．システムの処理概要を **図 3** に示す．

2.1.1 対象コンテンツ

本研究では，日経平均株価の動向を示すテキストとグラフを対象とする．テキストデータとして，国立情報学研究所の主催で実施されている評価型ワークショップの 1 つである「動向情報の要約と可視化に関するワー

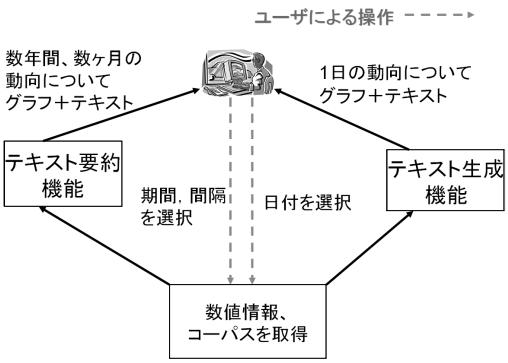


図 1 システム全体図
Fig. 1 Overview of system.

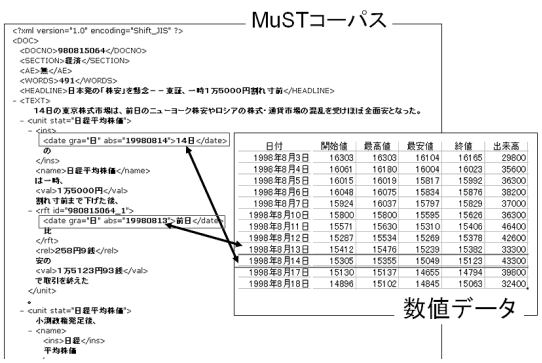


図 2 MuST コーパスと日経平均株価の数値情報
Fig. 2 MuST corpus associated with numerical data.

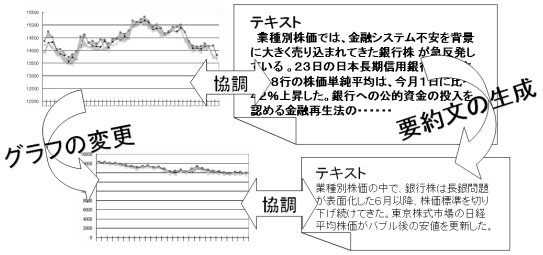


図 3 グラフ情報と協調した要約処理の概要
Fig. 3 Overview of summarization process collaborating with 2D chart.

クショップ」(NTCIR-5)[1),2)] で提供されている MuST コーパスを利用する．MuST コーパスとは，1998 年と 1999 年の 2 年分の毎日新聞から，ガソリン価格やパソコン出荷状況など 20 トピックについて時系列になっている記事を収集し，各トピックにつき 3 つ前後の統計量を選び，これらの統計量の可視化に必要な要素に対して，XML 文書として，人手でタグを付与したものである☆．日経平均株価の動向を示すテキストは，2 年分の記事からピックアップされた 20 日分の記事がある．その中で用いられているタグのうち主なものを次にあげる．

- **unit タグ**
  具体的にグラフの挙動（数値情報が得られる箇所）が記載されている文に付与される．属性として言及している統計量（stat）や出来事（event）が存在する．
  例：`<unit stat="日経平均株価">`
- **ins タグ**
  unit タグが付与されている文中の表現においてその文だけで意味が通じるように省略されている主語など，補完された部分に付与される．主に，統計量名やパラメータ名，に適当な助詞を加えたものが対象となる．
- **val タグ**
  統計量の値を示す部分に付与される．「前後」「約」「程度」などの範囲表現や概数表現も含めて付与される．
- **date タグ**
  「10 日」「今月」「昨年」などの時刻の表現している部分に対して付与される．gra 属性として，時，日，週，旬，月，4 半期，半期，年，不明などの，時刻表現が記述される．abs 属性として，示している日付を，西暦 4 桁，日の場合は 8 桁で記述される．
  例：`<date gra="日" abs="19980814">14 日</date>`
- **dur タグ**
  「4 カ月（ぶり）」などの期間の表現．date 要素と同様に gra 属性を持つ．
  例：`<dur gra="月">7 カ月半</dur>`ぶりに

☆ MuST コーパスの詳細については，
http://www.kecl.ntt.co.jp/scl/workshop/must を参照．

2.1.2 要約手法

要約文を生成する方法として，本システムでは重要文抽出法を用いる．この手法は，各文の重要度を計算し，重要度の高い文から順に，設定された要約の長さに達するまで文を選択するというものである．重要度を計算する際に判断基準として利用できる情報に以下の 6 つがあげられる[3)]．

(1) テキスト中の単語の重要度[4),5)]．
(2) テキスト中あるいは段落中での文の位置情報[6)]．
(3) テキストのタイトルなどの情報[6)]．
(4) テキスト中の手がかり表現[6)]．
(5) テキスト中の文あるいは単語間のつながりの情報[7)]．
(6) テキスト中の文間の関係を解析したテキスト構造[8)]．

本システムでは，MuST コーパスで与えられているタグ，および MuST コーパスの基となる毎日新聞コーパスに付与されているタグを利用しグラフと対応した要約文を生成するため，上記の (3) と (4) を利用する．

2.1.3 要約対象となる文の重要度の決定方法

上記 (3)，(4) の情報を利用してその重要度を決定するために以下のタグを用いる．

- **HEADLINE タグ**
  MuST コーパスで定義されているタグとは異なり，毎日新聞コーパスにおける記事の見出しに付与されている HEADLINE タグを参考にし，見出しで取り上げられている話題に言及している文を重要とする．具体的には，見出しを日本語形態素解析器「茶筌」にかけて名詞のみを取り出し，その名詞が含まれている文を重要と判断する．この際，見出しに含まれている名詞がより多く含まれている文をより重要とする．見出しは，記事の内容で最も主張したいことを端的に示す．この方法より，見出しに関連した内容を判断し，重要な文として抽出することができる．処理の流れを**図 4** に示す．

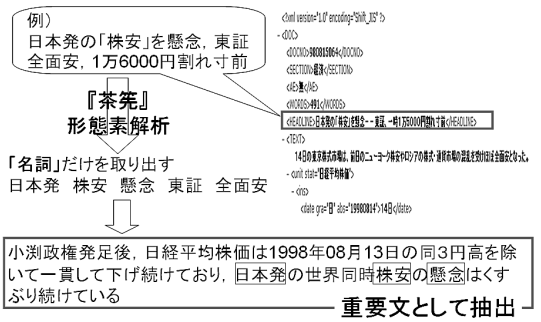


図 4 HEADLINE タグの利用
Fig. 4 The use of HEADLINE tag.

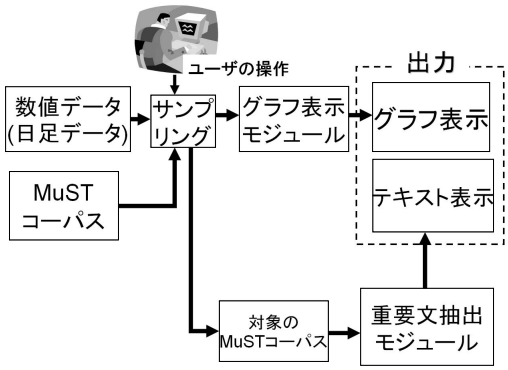


図 5 システム構成図
Fig. 5 System constitution.

- **unit タグ**
  unit タグは，具体的にグラフの挙動（数値情報が得られる箇所）が記載されている文に付与されている．このような文には，日経平均株価について重要と思われる値動きの部分が記載されているため，他の文と比べて重要度が高いと判断される．また，unit タグが付与されている文の中でも，「前日比」というように短い期間の挙動について言及しているものから，「12 年 8 カ月ぶりに」というように長い期間に言及している文もある．期間が長いということは，頻繁には起こらないことが起こったということであると見なし，より重要な文と判断する．そのために，期間の表現に付与されている dur タグを参考にし，長い期間か短い期間なのかを判断する．さらに，グラフ上で表されているものが日経平均株価であることから，それに対応するように，「業種別株価」について言及している文より，「日経平均株価」について言及している文を重要とする．

重要度の判定には，すべての文の重要度を初期値 0 として始め，上述した条件を判断し，重要度を累積していく．たとえば，HEADLINE に含まれる名詞が 1 つあるときには 1 ポイント加算し，3 つあるときには，3 ポイント加算する．より長い期間の挙動について言及している文を重要度がより高いとするため，dur タグの属性である gra 属性を参考にする．扱っている記事の中では，「日」「月」のみが存在することから，

```
<dur gra="月">
```

となっている，「月」のときのみ 1 ポイント加算する．unit タグが付与されていて，「日経平均株価」について言及している文は，unit タグの属性である stat 属性を参考にし，

```
<unit stat="日経平均株価">
```

とある文に，重要度を 1 ポイント加算する．

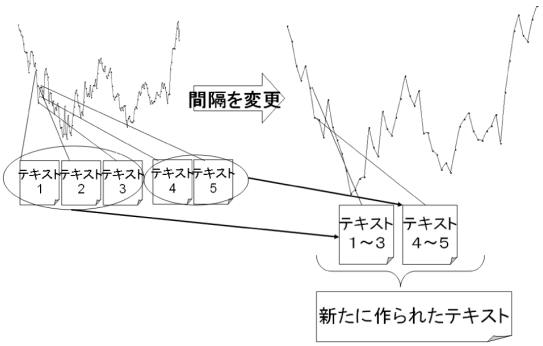


図 6 グラフの目盛り間隔の変更
Fig. 6 Change of sampling interval.

2.2 要約処理部のシステム構成

要約処理部では，ユーザの情報を閲覧したい視点に従い，変更されたグラフの状態に対応して限定されたニュース記事から重要文を抜き出すことにより要約文を生成する．ユーザは，数値データから興味がある範囲を選択し，グラフとして表示させる．MuST コーパスも同様に，グラフの表示詳細度に対応してニュース記事がサンプリングされ，重要度の高い文が抽出されて要約文として表示される．これにより表示されるグラフとテキストの協調が実現される．要約処理部のシステム構成を**図 5** に示す．

2.2.1 数値データ・MuST コーパスのサンプリング

グラフの目盛り間隔の変更

グラフの目盛り間隔が変更され粗くなった場合，ユーザは細かい流れよりも全体の傾向が知りたいと思うようになると考えられる．グラフが変更され，2 日おき，4 日おきのように目盛りの間隔が広くなった場合，2 日ごと，4 日ごとのように，重要文を抽出してテキストをまとめる（**図 6** 参照）．このとき，ユーザが設定した文数には関係なく，

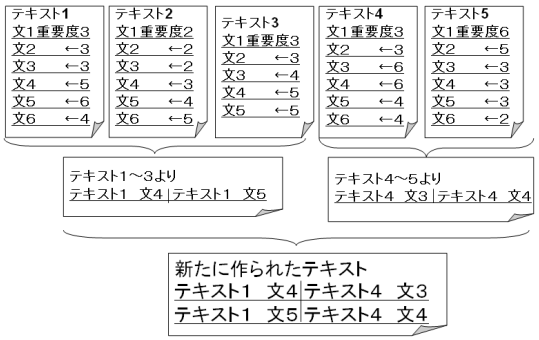


図 7 重要度による抽出例（グラフの目盛り間隔の変更）
Fig. 7 Summarization process based on change of sampling interval.

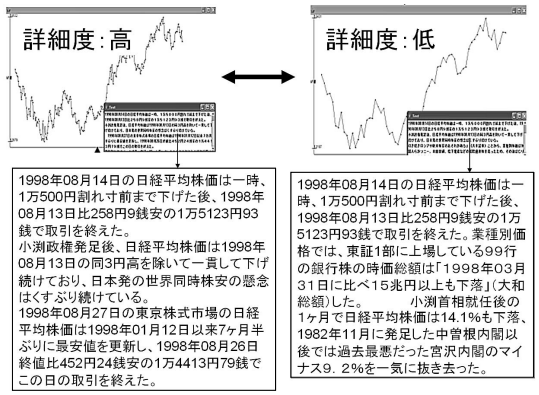


図 8 実行例（グラフの目盛り間隔の変更）
Fig. 8 An example of change of sampling interval.

それぞれ 2 文だけ抽出するように設定している．さらに，それぞれから抽出されたテキストから，新しい要約文を生成する（**図 7** 参照）．

この処理により，ある特定期間に集中した重要度が高いニュースを偏って抽出するのではなく，変更した目盛り間隔の各区間から全範囲にわたって重要な情報を抽出することができ，全体の傾向をとらえた要約文生成が可能となる．また，目盛り間隔が広くなれば抽出されるテキストが減り，情報の詳細度は低くなる．システムの実行例を**図 8** に示す．

範囲の選択

グラフの一部分が選択された場合，選択された日付の範囲にあるテキストの中から重要度の高い文を抽出する．このとき，抽出する文の数はユーザによって指定可能である．

この処理により，テキストも選択した範囲を焦点とした内容となる（**図 9** 参照）．また，目盛り間隔が変更された場合と異なり，選択した範囲全体

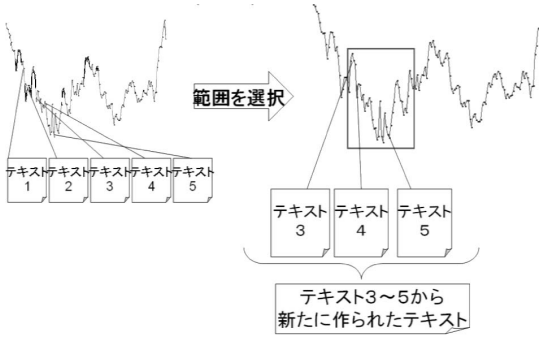


図 9 特定箇所の情報抽出
Fig. 9 Focus on a particular range.

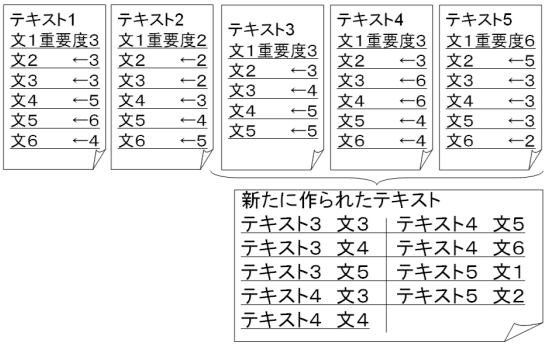


図 10 重要度による抽出例（特定箇所の情報抽出）
Fig. 10 Summarization based on a particular range.

の中で重要なニュースを詳細に示すことができる．範囲が狭くなればなるほど，変更する前には抽出されなかった重要度の低い文も抽出されるようになり，その範囲のみをより詳しく説明した要約文となる（**図 10** 参照）．ユーザは，新しいテキストから重要なニュースが起きている部分を判断し選択することで，そのニュースに関係する情報を詳細に表示させることもできる．

システムの実行例を**図 11** に示す．右側の図では，特定期間のグラフが拡大表示され，その期間内の要約文が抽出されている．

2.3 性能評価

MuST コーパスで提供されている 1998 年 9 月 9 日から 1998 年 10 月 24 日までの記事全 76 文のテキストを被験者 10 人（A〜J とする）に与え，ニュースとして重要な事柄について言及していると思われる文を 10 文選んでもらう．その結果を**表 1** に示す．被験者は，大学生 10 人，22 歳〜25 歳であり，株価に対する知識は，新聞を読み理解できる程度であり，知識の偏りがない被験者を選別して実験を行った．システムおよび被験者 10 人が選んだ文番号には○印を記す．この結果からシステムの性能を評価するため，次のよう

図 11 実行例（特定箇所の情報抽出）
Fig. 11 An example of focusing a particular range.

表 1 被験者実験によるシステムの評価
Table 1 Result of a subject experiment.

文番号	システム	A	B	C	D	E	F	G	H	I	J
0		○	○	○	○	○	○	○	○	○	
1		○			○		○		○		○
2	○					○					
4			○	○						○	
7			○							○	
8			○		○					○	
10								○			
15	○		○		○	○	○		○	○	
16			○		○		○				
17						○					
18						○					
19		○		○							
22									○		
23				○	○						
24				○							
25								○			○
29								○			○
30		○							○		○
33							○	○	○		
34				○			○				
37				○							
39					○					○	
44		○									○
45								○			
50	○		○	○	○	○	○		○	○	○
51	○	○	○				○				
52	○					○			○		
53	○	○						○			○
54			○							○	
58	○		○	○						○	
61	○	○		○	○	○	○	○			○
62						○			○	○	
63	○					○	○		○		
66					○						
71		○									○
73	○	○						○			
74								○			○

な計算を行う．文番号 0 の文は 9 人の被験者が選んでいるので 9 点，文番号 1 の文は 5 人選んでいるので 5 点というように，1 つの文を選んでいる被験者の人数をそれぞれの文の点数とする．選んだ文につけられた点数を累積計算をする方法で，被験者 10 人の総得点をそれぞれ計算する．その平均値を求めると 40.6 点となる．システムの点数は 38 点となり，平均値には至らなかったが，最高点数 9 点となっていた文番号 0 の文をシステムが選択できなかったことを考慮すると，多くの被験者が選んだ文を重要と判別できていたと判断できる．

提案システムと被験者による重要文選択の傾向の違いを見るために，下記 3 文について考慮する．このうち，文番号 0 は，次にあげる（a）〜（c）文中，（a）の文である．

(a) 1998 年 9 月 11 日の東京株式市場で日経平均株価が 1998 年最大の下げ幅を記録し，終値ベースで 1998 年 8 月 28 日につけたバブル崩壊の最安値の更新寸前となった

(b) 1998 年 10 月 09 日の東京株式市場は国際優良株の下落に引っ張られて日経平均株価はバブル後最安値を更新した

(c) ただ，業種別株価の中で銀行株の急反発は「空売り勢による買い戻しが中心で自律反発の域を出ない」（外資系証券）との見方が支配的だ

（a）の文は，日経平均株価について言及している文のため，他の文よりも重要度が高いと評価している．しかし，その文を含んでいる記事の見出しは『「内憂外患」波乱含み　NY の動向がカギ「売りが売り呼ぶ展開も」――東証全面安』というものであったため，本システムで採用している重要文抽出法では，重要として認識されない．一方，（b）の文を含む記事の見出しは，『［動乱マーケット］株価最安値更新　国際優良株が標的に』であり，本システムも抽出することができる．このように，本文中には「1998 年最大の下げ幅を記録」，「最安値の更新寸前」という重要なキーワードを含んでいるにもかかわらず，見出しにその重要なキーワードが含まれていないため，この要約手法では抽出ができなかった．逆に，システムが選択しているが被験者があまり選択していない文（文番号 2，73）もある．文番号 73 は，上にあげた（c）の文である．これは，書き手の個人的な意見を含む文である．このため，被験者は重要度が低いと判断したと推測されるが，採用した重要文抽出法ではそのような文を判断することができずに重要文として抽出している．また，実験で使ったテキストには，「株価最安値更新」という見出しの記事と，「銀行株が急反発，再生法施行で不安やわらぐ」という見出しの記事がある．本システムの場合，記事どうしでどちらの出来事が重要なのかという判定を行わない．数値データから，幅がある部分，最小値をとっている日，最大値をとっている日を判断し，重要文を抽出する基準として付け加えることを今後の課題としたい．

3. テキスト生成機能

3.1 提 案 手 法

テキスト生成機能では，数値データをグラフ（チャート）表示した際のグラフの形状を線形最小二乗法により近似し，近似曲線の部分形状のパターンを言語的にとらえることにより，グラフの挙動を説明するテキスト生成を行う．

本システムによって生成されるテキストは以下の 2 つのタイプに分類され，タイプごとにテキスト生成

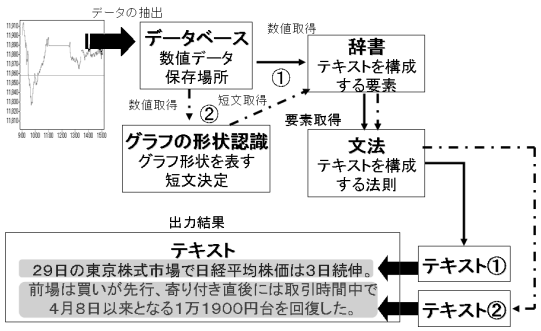


図 12 システム構成図
Fig. 12 System architecture.

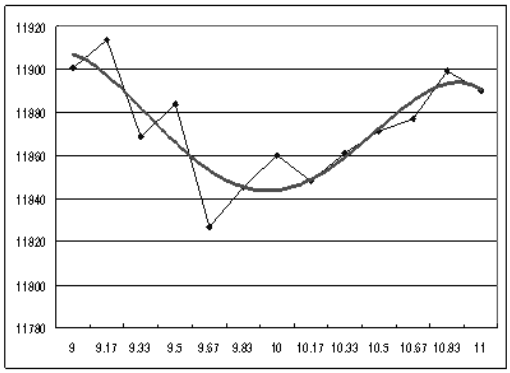


図 13 線形最小二乗法の適用例
Fig. 13 Example of least squares.

の処理の流れが異なる．

タイプ ① テキスト：グラフの形状をふまえることなしに，データベースからの情報のみから生成できるテキスト．

タイプ ② テキスト：グラフの形状をふまえて，かつ，データベースからの情報から生成できるテキスト．

3.2 テキスト生成処理部のシステム構成

システムの構成を**図 12** に示す．タイプ ①，および，タイプ ② テキストの生成の流れは，図 12 中，実線および一点鎖線でそれぞれ示す．

以下，処理の流れに沿って，システム各部の説明を行う．

3.2.1 データベース

テキスト生成における最初の処理として，日経平均株価 2005 年 7 月 25 日から 8 月 30 日までの分足データ，始値，終値，高値，安値から必要な数値情報を取得する．これらの数値データの管理を容易に行うため，Microsoft Office Access を用い，Java 言語のデータベース管理機能（DBMS）を通じてシステム本体からアクセス可能とする．

3.2.2 グラフの形状認識

午前の相場である前場と午後の相場である後場のグラフの形状を認識する．**図 13** に示す折れ線グラフの動向を視覚的に把握すると，「下がって，上がっている」と認識される．このようなグラフの視覚的特徴を把握するために，本研究では線形最小二乗法を用いてグラフの近似曲線を作り，その近似曲線の振舞いをとらえることによりグラフの動向を言語で認識する．

線形最小二乗法の適用

まず，多項式の次数を設定するために，上記期間内の前場，後場のグラフに対して異なる次数の近似曲線を作成し，実際のコーパスとの対応を調べた．その結果より，グラフの挙動を表現するのに使用される言語

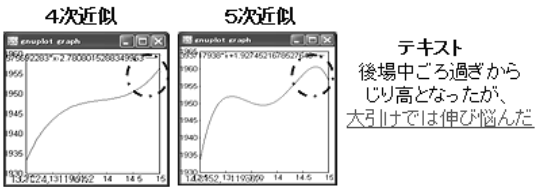


図 14 4 次近似と 5 次近似の比較
Fig. 14 Comparison between fourth and fifth approximation.

表現を最も的確に表す近似曲線として，5 次多項式を採用した．**図 14** にグラフの挙動を示す表現と，近似曲線の対応を調べた一例をあげる．テキスト「後場中ごろ過ぎからじり高となったが，大引けでは伸び悩んだ」に対応したグラフを，4 次近似と 5 次近似で表す．テキストから，グラフは右端は上がった後，下がる必要がある．これより，テキストをより正確に表現できているのは 5 次多項式近似であることが分かる．同様に 6 次多項式の可能性を検証したが，近似曲線の挙動が複雑となり，5 次多項式の方がコーパスに現れる言語表現を適切に表しているという結果が得られたため，最適な次数は 5 次であると判断した．

グラフの全体形状と部分形状

5 次多項式が表現する典型的な曲線の全体的な形状を極値の個数などにより 11 のタイプに分割し（**図 15** 参照），その形状のパラメータの値のとり方，および，グラフの挙動を説明するために使われる言語表現の観点からさらに 13 種類の部分形状を定義する（**表 2** に type4 までのその一例を示す）．極値に関しては，5 次多項式において微少な時間幅での傾きを求め，傾きの極性が変化する座標の値を求めるアルゴリズムにより求める．5 次多項式で認識されたグラフの形状は，上述した全体形状の 11 タイプの 1 つのタイプとして認識される．次に，そのグラフの部分形状の特徴量を数

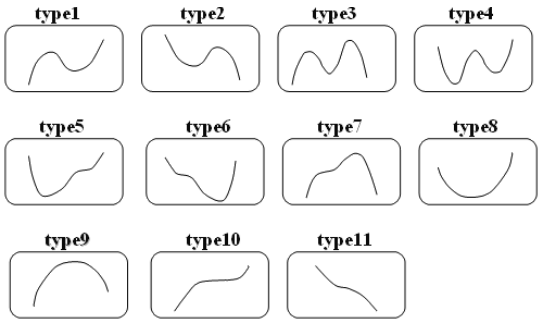


図 15 グラフの全体形状
Fig. 15 Eleven types of whole shape.

表 2 全体形状と部分形状
Table 2 Example of partial shapes.

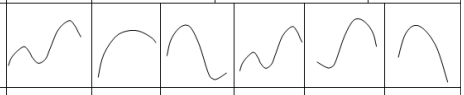
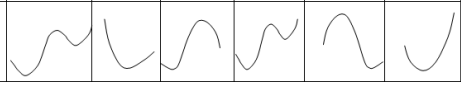
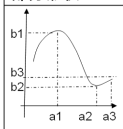
分類	形状	部分形状
type 1		
type 2		
type 3		
type 4		

表 3 部分形状の数式的解釈とその諸表現例
Table 3 Partial shape and verbal expressions.

部分形状	特徴	短文+時間帯
	|b2−b1|/|MAX−MIN|>0.4 |a1−a2|/|max−min|<0.7	売りが優勢だった
	|a1−a2|/|max−min|>0.7	売りが広がった
	|b2−b1|/|MAX−MIN|>0.4 |b2−b3|/|b2−b1|>0.5 |a1−a2|/|max−min|<0.7	売りが優勢になる場面があった
	|a3−a2|/|max−min|<0.7 |a3−a2|/|max−min|>0.5	中ごろ過ぎにかけて
	|a3−a1|/|max−min|<0.2 |a3−a2|/|max−min|<0.2	中ごろに
	|a3−a1|/|max−min|<0.6 |a3−a1|/|max−min|>0.45	中ごろ過ぎから

MAX，MIN：前場（後場）での株価の最大値，最小値
max，min：前場（後場）での時間の最大値，最小値

式的に解釈することにより，これを説明する適切な言語表現を選択する（**表 3** 参照）．

3.2.3 辞　　書

2005 年 7 月 25 日から 8 月 30 日までの 27 のテキストを分析することにより，グラフの挙動を説明するのに頻繁に使用される語彙の抽出を行った．それらは，38 種類の部分形状を表現する短文（例：「売りが広がった。」，「じり高歩調となった。」），19 種類の特定水準（データ）から判断できる短文（例：「反発」，「続伸」，「1 日を通じて高い水準で推移した。」），10 種類の時間帯および比較表現（例：「前場は、」，「大引けで」，「中ごろ過ぎから」，「前週末比」），3 種類の接続詞（「そして、」，「なので、」，「しかし、」）である．これらの中から，認識されたグラフの形状（グラフの挙動）を表現する適切な語彙を選択する．

3.2.4 文　　法

タイプ ① テキストでは，データベースから得られた数値情報を基に，あらかじめ用意されたテキスト生成用テンプレートと文法規則に基づきテキストを生成する．タイプ ② テキストでは，グラフの形状が認識され，言語表現された短文を時間軸に沿って，状況語や理由などを示す接続詞を適切に追加することによりテキストを生成する．タイプ ① およびタイプ ② の文法規則は，上記 27 のテキストを分析することにより得ている．

タイプ ① テキスト生成用テンプレートと文法規則

前日と当日の終値の比較の観点に基づいて作成された 3 種類の短文テンプレート（以下に示す（a），（b），（c））と，それらを補足する，数値情報の特定点の比較によって判断されるグラフの状態を表現する 10 種類の短文（例：「1 日を通じて高い水準で推移した。」，「今日の高値で引けた。」）を用意する．この 2 つからそれぞれ適切なものを選択し，タイプ ① のテキストを生成する．

（a） __日の東京株式市場で日経平均株価は__。終値は__円__銭__（__%）__の__万__円__銭だった。

（b） __日の東京株式市場で日経平均株価は__。終値は__円__銭__（__%）__の__万__円__銭で、__円台を割り込んだ。

（c） __日の東京株式市場で日経平均株価は__。終値は__円__銭__（__%）__の__万__円__銭で、__円台を回復した。

タイプ ② テキスト生成用文法規則

グラフの形状を表現した短文をヒューリスティックに得られた 5 種類の時間帯を表す表現，および，原因を表す 3 種類の接続詞を用いてつなぎテキストを生成する．時間帯および接続詞挿入のヒューリスティック規則を以下に示す．

- 時間帯の挿入
  - 時間帯によって先頭に「前場は、」，「後場は、」をつける．
  - 2 種類の部分形状において「売りが優勢だった」や「上昇に転じた」など，特定の

短文が該当する 5 次多項式の形状の時間幅に応じて「中ごろ過ぎにかけて」,「中ごろに」,「中ごろ過ぎから」が選択され,短文の前に補われる.

– 短文の前に時間帯,接続詞がなく,またその短文の前に 2 種類の部分形状を言及する短文(例:「売りが先行した」)が存在する場合,その間に「その後、」を補足する.
– 特定の部分形状において「売りが増えた」など一定状態の継続を表現する短文が該当する 5 次多項式の形状の時間幅に応じて「中ごろ以降は」をその短文の前に補足する.
– 特定の部分形状において,短文「下げ幅をじりじりと広げた」が該当するとき,5 次多項式の形状の時間幅に応じて「中ごろにかけて」をその短文の前に補足する.

- 接続詞の挿入
  – 短文「上げ幅は小幅にとどまった」の前に,その原因を示す「底堅さを確認した」という短文の有無に応じて,存在する場合は「なので、」,存在しない場合は「しかし、」をつける.
  – 「下げ渋った」の前にその原因を示す「買いが入った」という短文がくると,接続詞「そして、」を「下げ渋った」の前につける.
  – 上げ幅,下げ幅について言及している短文には,その短文の前に事象の推移を明確に示すため接続詞「そして、」をつける.

3.3 実 行 例

システムの実行例を**図 16** に示す.

図 16 の例は,日付 2005 年 8 月 22 日を入力し,「データ表示」ボタン,「チャート表示」ボタンをクリックすると入力日の分足時系列数値データ,グラフが表示され,「テキスト表示」ボタンをクリックすることにより,入力日の日経平均株価の挙動を説明するテキストが生成されたものである.以下に,テキストの生成過程をテキストのタイプごとに示す.

タイプ ① のテキスト生成過程

step1. データベースからの数値情報取得

2005 年 8 月 22 日のグラフ数値情報と過去の始値,終値,高値,安値の数値情報を取得する.

step2. グラフの形状に対する語彙選択

step1 で得られた数値情報をもとに,適切な語彙が選択される.

図 16 システムの実行例
Fig. 16 Example of text generation.

- 入力日が「2005 年 8 月 22 日」から日付「**22**」が取得される.
  出力:「22 日の東京株式市場で」
- 終値を前日と比較して入力日は「高い」,前々日と比較して前日は「安い」,3 日前と比較して前々日は「高かった」であることより,その状態を表現する「**反発。**」が選択される.
  出力:「22 日の東京株式市場で日経平均株価は反発。」
- 入力日が月曜で 1 営業日前の終値と比較して 160.78 円高く,入力日の終値での百分率が 1.38%であることより「**前週末比 160**」,「**78**」,「**1.38**」,「**高**」.
  出力:「22 日の東京株式市場で日経平均株価は反発。終値は前週末比 160 円 78 銭(1.38%)高の」
- 入力日の終値が 12452.51 円であることより「**1**」,「**2452**」,「**51**」.
  出力:「22 日の東京株式市場で日経平均株価は反発。終値は前週末比 160 円 78 銭(1.38%)高の 1 万 2452 円 51 銭」

step3. テキストを表現する文法の適用

タイプ ① のテキストでは,step1 で得られた数値情報をもとに,あらかじめ用意された短文テンプレートを適切に選択する.

- 入力日の終値が 12400 円台で,その前日終値が 12400 円以下であることより「**12400**」が取得される.
- 前日と当日の終値の比較により,以下の短文テンプレートが選択される.

「__日の東京株式市場で日経平均株価は__終値は__円__銭__（__%）__の__万__円__銭で、__円台を回復した。」

出力：「22 日の東京株式市場で日経平均株価は反発。終値は前週末比 160 円 78 銭（1.38%）高の 1 万 2452 円 51 銭で、12400 円台を回復した。」

- 前日の終値と比較して 25 円より高く相場が終了したことから短文「高値水準をつけた。」を短文テンプレート以降に補足する．
- 前場，後場ともに始値より終値の方が高くまた，終値が始値より 100 円より高いことから短文「相場上昇。」を短文テンプレート以降に補足する．
- 前日の終値と比較して，前場・後場の平均値がともに 25 円より高いことから短文「1 日を通じて高い水準で推移した。」を短文テンプレート以降に補足する．

上記処理例において生成されるタイプ ① テキストは以下となる．

「**22** 日の東京株式市場で日経平均株価は**反発。**終値は**前週末比 160** 円 **78** 銭（**1.38** %）**高**の **1** 万 **2452** 円 **51** 銭で、**12400** 円台を回復した。**高値水準**をつけた。**相場上昇。1 日を通じて高い水準で推移した。**」

タイプ ② のテキスト生成過程

step1. データベースからの数値情報取得

2005 年 8 月 22 日のチャート数値情報と過去の始値，終値，高値，安値の数値情報を取得する．

step2. グラフの形状認識

step1 で得られた数値情報をもとに線形最小二乗法を用いて，午前の相場である前場と午後の相場である後場のグラフの形状を認識する．

- 前場は type1 と判別．
- 後場は type7 と判別．

step3. グラフの形状に対する語彙選択

step1 で得られた数値情報と step2 で得られたグラフの形状（部分形状）から，それを表現する適切な短文，および語彙（短文に付随する時間帯）を選択する（**図 17** 参照）．前場は，「**買いが先行した。**」，「**底堅さを確認した。**」，「**上げ幅は小幅にとどまった。**」が選択され，後場は，「**じり高歩調となった。**」，「**上げ幅を広げた。**」，「**伸び悩んだ。**」が選択される．

step4. テキストを表現する文法選択

タイプ ② のテキストでは，step3 で選択した語彙に付随する時間帯，接続詞を選択する．

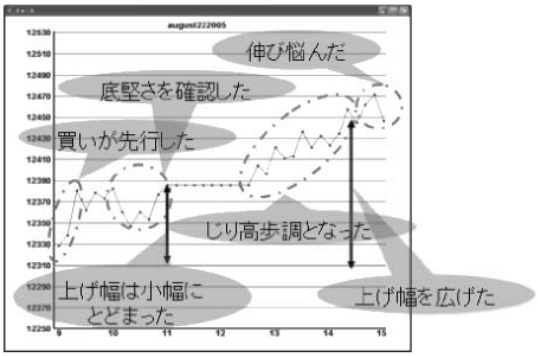


図 17 グラフの形状と選択された語彙表現
Fig. 17 Shapes and their verbal expressions.

- 時間帯に合わせて，「前場は、」，「後場は、」が補足される．
- 「買いが先行した。」と「底堅さを確認した。」の間に「その後、」が補足される．
- 「上げ幅は小幅にとどまった。」の前に「なので、」が補足される．
- 「伸び悩んだ。」の前に「**大引けで**」を補足される．

以上より生成されたタイプ ② テキストを以下に示す．

「前場は、**買いが先行した。**その後、**底堅さを確認した。**なので、**上げ幅は小幅にとどまった。**後場は、**じり高歩調となった。上げ幅を広げた。大引けで伸び悩んだ。**」

3.4 性能評価

本研究で取り上げた 27 のコーパスとシステムが生成したテキストにおいて，「状態を表現するテキスト」，「状態に付随する時間帯（比較）」，「数値」について言及している箇所がどの程度一致しているかという観点からシステムの評価を行った．その結果を**表 4** に示す．

本論文で提案しているテキスト生成機能の特徴は，グラフの挙動を視覚的にとらえることによってグラフを言語表現で表すことができることである．したがって「状態を表現するテキスト」の再現率の高さが本機能の有用性を決めると考えられる．その点について表 4 より判断してみると，グラフと過去 3 日間のデータより判断可能な状態を表現するテキストの再現率は，約 90%（119 文中 107 文）であった．また，実際のコーパスでは，「裁定買い」などグラフだけでは判断不可能な表現を含む文が状態を表現するテキスト全体の中で約 34%含まれていた．一方，システムが生成

表 4 テキスト生成機能の性能評価
Table 4 Evaluation of the ability of our system.

実際のコーパス		生成したテキスト		
状態を表現するテキスト		状態を表現するテキスト		
グラフ + 最低過去 3 日間のデータより判断可能	グラフ + 最低過去 3 日間のデータより判断不可能	本文と一致	本文よりも詳細	不適切
119	62	107	76	11
状態に付随する時間帯（比較）		状態に付随する時間帯		
グラフ + 最低過去 3 日間のデータより判断可能	グラフ + 最低過去 3 日間のデータより判断不可能	本文と一致	本文よりも詳細	不適切
61	10	59	42	0
数値		数値		
グラフ + 最低過去 3 日間のデータより判断可能	グラフ + 最低過去 3 日間のデータより判断不可能	本文と一致	本文よりも詳細	不適切
181	66	176	0	1

したテキストには不適切な表現もいくつか含まれていた．その原因としては，グラフを構成する数値データに 10 分足データを利用し，その数値データをさらに 5 次多項式で近似したものに対して言語表現を行ったため，あいまいさの拡大が起こってしまったことが考えられる．より時間間隔の短い数値データを用いることができるならば，生成される不適切な表現は減少すると推測される．

本来，システムの性能評価を行うには，システムを開発する際に使用したデータとは異なるものを利用する必要性がある．表 4 に示す結果は，利用可能なデータ数の不足により，それを行えていない．しかし，グラフの挙動を 11 個のタイプと 13 個の部分形状でとらえると設定し，それらの形状のパターンを比率で定義したものに言語表現を対応づけたことを考えると，約 90%の高い適合率は，2 つと存在しないグラフの詳細な動きを的確に近似し，パターンと適合し，同じ言語表現に対応させることができていることを示唆している．この点において，現在のシステムは今後他のデータを利用して性能を評価するのに十分に値する完成度を持っているといってよいと考える．

4. 関 連 研 究

テキストとグラフの協調に関する先行研究として，文献 9) や 10) などがあげられる．また，文献 11) では，複数のモダリティの情報をユーザにとって適切な情報量のマルチメディア文書として生成する研究を行っている．これらは，マルチメディア文書として，テキストやグラフなどがどのような双方向の対応関係にあるかについて分析をしている．一方，本研究におけるテキストとグラフの協調に基づく要約手法は，グラフの表示状態に合わせてテキストが要約されるという 1 方向の協調関係に基づくものである．上記，先行研究は，本研究において協調関係を双方向に拡張する際に有益なアイディアを提案するものであり，今後，参考とするつもりである．また，グラフの挙動を自然言語で表現する研究として，文献 12) においては，時系列データに Wavelet 解析を行い，気温の長期・短期変動をとらえ，自然言語で表現する手法を提案している．また，マルチメディア文書の要約として，テキスト要約と同様にグラフ情報を要約する研究に文献 13) があげられる．ここでいうグラフ情報の要約とは，グラフが意味する情報を自然言語で表現するというものである．同様に，グラフの挙動を自然言語で説明する研究に文献 14) があげられる．ここでは，システミック言語理論を用い，言語表現とグラフ特徴の関係を分析し，それに基づいたテキスト生成を提案している．

ほかに，数値情報などを自然言語表現する研究として，気象データから天気予報の生成を行うシステムの開発[15),16)] などがある．また，時系列データからパターンを取り出す研究には，経済データを使ったもの[17)]，プロセス制御データを使ったもの[18)]，また，薬品データを使ったもの[19)] などがある．関連研究の中に，グラフを解析する際に Wavelet 解析を用いているものが多く見受けられる．しかし，解析によって得られた結果と自然言語表現を結び付けることにおいて明確な手法を提案しているものは数少ない．我々は，グラフで表現される時系列データを言語によって表現するために，Wavelet 解析などによってグラフの特異な変化量をとらえた特徴量に言語表現をあてはめることではなく，グラフを説明する言語表現の下で，視覚的に表現されているグラフの挙動をいかにとらえるかというアプローチをとっている．このため，グラフの挙動を説明する表現を収集するのに最も適しているテキストの 1 つである株価の動向が記述されたテキストを対象に新たな手法の提案を行った．

5. 結　　論

本研究では，異なるモダリティが協調することにより情報を効果的に提示する技術開発の一環として，グラフとテキストという異なる 2 つのモダリティ情報を用い，テキスト要約・生成手法を用いた情報提示方法の提案，および，システムの実装を行った．テキスト要約手法を利用した情報提示に関しては，グラフの表示

状態に対応しテキストの表示内容を変更する手法の提案を行った．これにより，ユーザの情報閲覧の興味に応じた情報提示が行われる．要約文生成には，重要文抽出手法を用いたが大別して 6 つある手法のうち，現時点では 2 つしか用いていない．また，HEADLINE タグから取り出した名詞にも一律 1 ポイントの重要度を加算しているなど語彙の重要度を考慮した重要文の判定機能が未実装となっている．今後，要約対象となるテキストから tf-idf などを用いて判定される重要語を利用し，また，テキストの修辞構造をとらえることなどにより重要文抽出の精度を向上させる予定である．また，コンテンツのさらなる知的化を目指して，MuST コーパスで用意されているタグが付与されていない文の中で，重要度が高いとする文には，新たにタグを追加し重要度を判断する基準とするなど，グラフとテキストの情報がより協調する仕組みを工夫し，提示方法を自由に変化させることができる手法として開発を進めるつもりである．

テキスト生成手法に関しては，数値データが視覚的に表現されたグラフの形状を線形最小二乗法による近似曲線の部分形状のパターンととらえることより，その挙動を説明するテキスト生成の手法を提案した．先行研究においては，グラフを Wavelet 解析することにより，その挙動を言語表現する研究もあるが，本研究において提案する手法はグラフの挙動を表現する語彙の立場からグラフ解析の精度を決定するという新しい手法を提案している．

提案したテキスト生成手法は，視覚障害者に対して，グラフ（チャート）を説明する際に有益なインタフェースとして活用することができる．また，グラフ表示される統計データに対して，言語によるインデックス化を行うことができ，特定の統計データの事例などを言語によって検索できる検索エンジンの中核技術として活用することが可能である．

今回，生成されたテキストに対する評価においては，システムを開発する際に用いたデータを評価データとして用いたため，今後は異なるデータを用いて，より正確なシステムの評価を行う予定である．

備　　考

本研究においては，国立情報学研究所主導における NTCIR-6 パイロットワークショップである「動向情報の要約と可視化に関するワークショップ」[20]（URL: http://must.c.u-tokyo.ac.jp/）における毎日新聞 98 年および 99 年の記事に注釈づけされた研究用データセット（MuST コーパス）を利用している．

参 考 文 献

1) 加藤恒昭，松下光範，神門典子：動向情報の要約と可視化—その研究課題とワークショップ，知能と情報（日本知能情報ファジィ学会誌），Vol.17, No.4, pp.424–431 (2005).
2) 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，第 19 回人工知能学会全国大会予稿集，1E3-03 (2005).
3) 奥村　学，難波英嗣：知の科学テキスト自動要約，人工知能学会，株式会社オーム社 (2005).
4) Luhn, H.P.: The automatic creation of literature abstracts, *IBM journal of Research and Development*, Vol.2, No.2, pp.159–165 (1958).
5) Salton, G.: Automatic Text Processing, Addison-Wesley (1989).
6) Edmundson, H.P.: New methods in automatic extracting, *Journal of the Association for Computing Machinery*, Vol.16, No.2, pp.264–285 (1969).
7) Barzilay, R. and Elhadad, M.: Using lexical chains for text summarization, *Proc. ACL Workshop on Intelligent Scalable Text Summarization*, pp.10–17 (1997).
8) Marcu, D.: From Discourse Structures to Text Summaries, *Proc.ACL Workshop on Intelligent Scalable Text Summarization*, pp.82–88 (1997).
9) Corio, M. and Lapalme, G.: Integrated Generation of Graphics and Text: A Corpus Study, *Proc. CVIR98*, pp.63–68 (1998).
10) 加藤恒昭，松下光範：新聞記事におけるテキストとグラフの協調に関する分析，言語処理学会第 9 回年次大会，pp.47–50 (2003).
11) Matthiessen, C., Zeng, L., Cross, M., Kobayashi, I., Teruya, K. and Wu, C.: The Multex generator and its environment: Application and development, *Proc. 9th International Workshop on Natural Language Generation* (*INLG-98*), pp.228–237 (Aug. 1998).
12) Boyd, S.: TREND: A System for Generating Intelligent Descriptions of Time-series Data, *Proc. IEEE-ICIPS* (1998).
13) Carberry, S., Elzer, S., Green, N., McCoy, K. and Chester, D.: Extending Document Summarization to Information Graphics, *Proc. ACL Workshop on Text Summarization* (2004).
14) 小林一郎：グラフ情報の自然言語表現に関する研究，日本ファジィ学会誌，Vol.3, No.12, pp.406–416 (2000).
15) Goldberg, E. and Driedger, N.: Using Natural language processing to produce weather forecasts, *IEEE Expert* (Apr. 1994).
16) Coch, J.: Interactive generation and knowledge administration in multimeteo, *Proc. 9th*

International Workshop on Natural Language Generation (*INLG-98*), pp.300–303 (Aug. 1998).
17) Berndt, D.J. and Clifford, J.: Finding patterns in time series: A dynamic programming approach, *Advances in Knowledge Discovery 1996*, AAAI MIT Press (1996).
18) Bakshi, B.R. and Stephanopoulos, G.: Reasoning in time, *Advances in Chemical Engineering*, Vol.22, pp.485–548, Academic Press, New York (1995).
19) Himowitz, I.J., Le, P.P. and Kohane, I.S.: Clinical monitoring using regression-based trend templates, *Artificial Intelligence in Medicine*, Vol.7, pp.473–496 (1995).
20) 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，2004-NL-164 (15), pp.89–94 (2004).

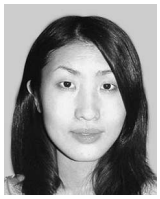
(平成 18 年 6 月 20 日受付)
(平成 18 年 12 月 7 日採録)

小林　一郎（正会員）

1965 年生．1995 年東京工業大学大学院総合理工学研究科システム科学専攻単位取得退学．博士（工学）．1995 年法政大学経済学部研究助手．1996 年同助教授．2000 年理化学研究所脳科学総合研究センター非常勤研究員．2003 年お茶の水女子大学理学部情報科学科助教授．現在に至る．言語情報処理を主体とする知能情報処理技術の開発に従事．第 16 回人工知能学会全国大会ベストプレゼンテーション賞．日本機能言語学会理事，人工知能学会，言語処理学会，日本知能情報ファジィ学会各会員．

渡邉　千明（学生会員）

1983 年生．2006 年 3 月お茶の水女子大学理学部情報科学科卒業．同年 4 月お茶の水女子大学大学院人間文化研究科博士前期課程入学．The 2nd International Symposium on Computational Intelligence and Industrial Applications, Excellent Student Paper Award 受賞．日本知能情報ファジィ学会会員．



奥村奈穂子

1982 年生．2006 年 3 月お茶の水女子大学理学部情報科学科卒業．同年株式会社 NTT データ入社．法人ビジネス事業本部コンサルティングビジネスユニットビジネスデザイン担当．