2H5-4

# 動向情報を表すテキスト生成

A Study on Text Generation Explaining Trends Information

渡邉千明*1　　小林一郎*2
Chiaki WATANABE　Ichiro KOBAYASHI

*1お茶の水女子大学大学院　人間文化研究科　数理・情報科学専攻
Graduate School of Humanities and Sciences, Ochanomizu University

*2お茶の水女子大学　理学部　情報科学科
Dept. of Information Sciences, Ochanomizu University

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.

## 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．この要因の一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わかりやすくなく，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ていないことが挙げられる．本研究では，このような現状を踏まえ，情報の内容や表示が誰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同士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．

## 2. 提案手法

一般的に，ユーザがグラフなどの数値情報を把握する際，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．長期的な動向を捉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一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，短期的な動向を捉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，長期的な動向を捉えるために，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捉えるために，視覚的に捉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．これらの二つの手法を用い，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．具体的に，始めにグラフとテキストを表示する．ユーザは，マウスにより閲覧したい視点を選択する．そのときに，1 日のみ，もしくはテキストが存在しない期間を選択された場合，グラフの挙動を説明するテキスト生成システムを使い，それ以外の場合，要約文生成システムを使いテキストを表示する（図 1 参照）．

連絡先: 渡邉千明，お茶の水女子大学 人間文化研究科 数理・情報科学専攻 小林研究室，東京都文京区大塚 2-1-1，03-5978-5709，chiaki@koba.is.ocha.ac.jp

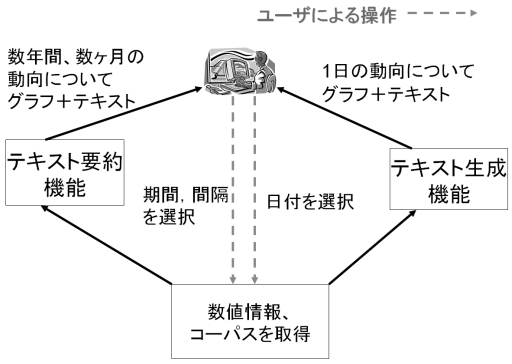



図 1: システム全体図

## 3. 要約文生成機能

### 3.1 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ひとつである「動向情報の要約と可視化に関するワークショップ」（NTCIR-5）[1, 2] で提供されている MuST コーパスを利用する*1．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．

### 3.2 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 つが挙げられる [3].

*1　MuST コーパスの詳細については，http://must.c.u-tokyo.ac.jp/を参照.

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１，３と４を利用する．要約対象となる文の重要度の決定するために，まず tf・idf 法を利用し，ある名詞の文章における相対的な重要度を算出する．以下の計算式を使い，各単語の重要度を決める．

$$\text{各単語の重要度} = tf \times idf \quad (1)$$

**tf:** 文書中 (MuST コーパス) での単語の出現回数
**df:** その単語が出現した文書数 (MuST コーパスの記事の数)
**N:** 文書集合中の全文書数
**idf:** $log(N/df)$

各行の重要度を計算するため，すべての文の重要度を初期値 0 として始める．そして，その文に含まれている名詞を判断し，tf・idf 法の計算で算出された各名詞の重要度を加算していくことで求める．さらに，MuST コーパス中で使用されているタグに基づき重要度を加算する．重要度の計算に使用される二つのタグについて説明する．

- **HEADLINE タグ**
  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．処理の流れとして，まず HEADLINE タグで取り出した見出しを，茶筌で形態素解析する. その結果から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．さらに重要度のランク付けとして，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．また，重要度を相対的に定めるため，各名詞の tf・idf 値を計算した後に，見出しに存在する名詞に tf・idf 値の最大値から平均値を引いた差を加算するという工夫をしている．

- **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．この計算をするために，HEADLINE タグを利用したときと同様に，相対的な値を足して重要度を高くすることを行う．unit タグの付与された文には，コーパス内の各文に含まれている名詞の tf・idf 値の和を求め，その最大値から平均値を引いた値を加算する．

### 3.3 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．ユーザはグラフの目盛間隔を変更すること，及び，特定の範囲を選択することができる．ユーザの操作を受け，グラフの表示詳細度に対応した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 2 に示す.

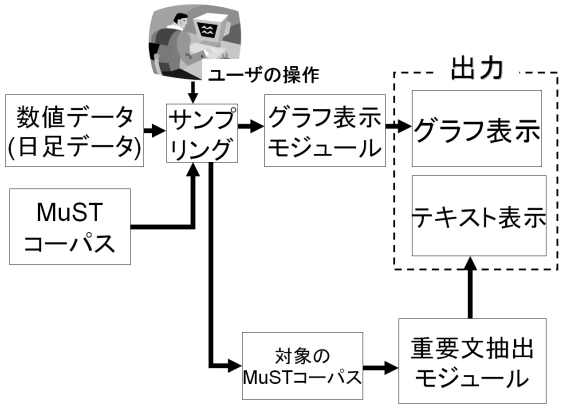



図 2: システム構成図

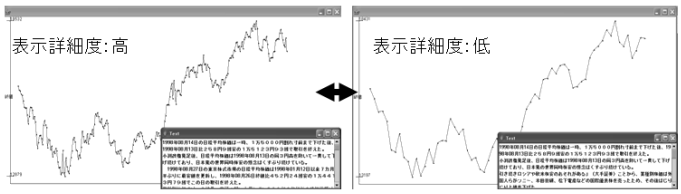



図 3: 実行例 (グラフの目盛り間隔の変更)

**グラフの目盛り間隔の変更**
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．この時，ユーザが設定した文数には関係なく，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．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渡って重要な情報を抽出することができ，全体の傾向を捉えた要約文生成が可能となる．

**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．また，目盛り間隔が変更された場合と異なり，選択した範囲全体の中で重要なニュースを詳細に示すことができる．

### 3.4 実行例

グラフの目盛間隔を変更した場合，特定範囲を選択された場合の実行例を図 3 と図 4 に示す.

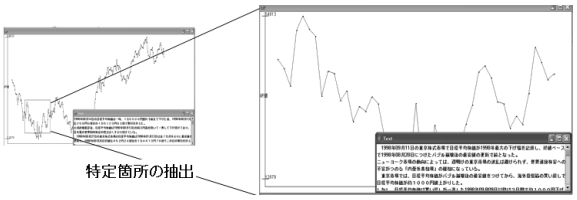



図 4: 実行例 (特定箇所の情報抽出)

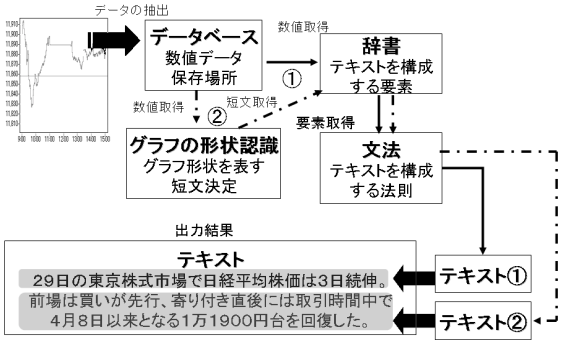



図 5: システム構成図

## 4. テキスト生成機能

### 4.1 提案手法

テキスト生成機能では，数値データをグラフ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捉えることにより，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の処理の流れを変える．

**タイプ①テキスト：** グラフの形状を踏まえることなしに，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：** グラフの形状を踏まえて，かつ，データベースからの情報から生成できるテキスト．

### 4.2 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 5 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 5 中，実線および一点鎖線でそれぞれ示す．

- 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．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，その近似曲線の振る舞いを捉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．

- 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

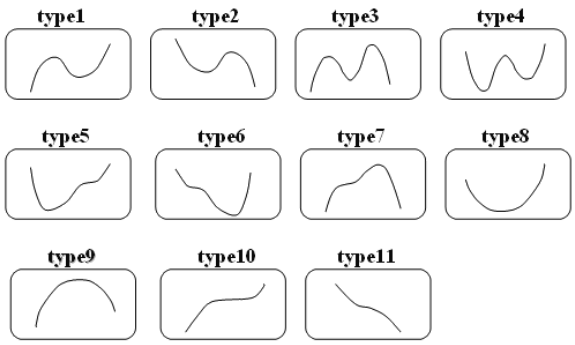



図 6: グラフの全体形状

値の個数などにより 11 のタイプに分割し（図 6 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 1 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 タイプの一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 2 参照）．

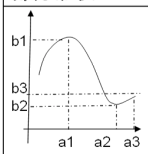
表 1: 全体形状と部分形状

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

- 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．

- 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

表 2: 部分形状の数式的解釈とその諸表現例

| 部分形状 | 特徴 | 短文+時間帯 |
|---|---|---|
|  | \|b2−b1\|/\|MAX−MIN\|>0.4<br>\|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<br>\|b2−b3\|/\|b2−b1\|>0.5<br>\|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<br>\|a3−a2\|/\|max−min\|>0.5 | 中ごろ過ぎにかけて |
| | \|a3−a1\|/\|max−min\|<0.2<br>\|a3−a2\|/\|max−min\|<0.2 | 中ごろに |
| | \|a3−a1\|/\|max−min\|<0.6<br>\|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

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．

### 4.3 テキスト生成過程

**step1. データベースからの数値情報取得**

選択された日の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**

タイプ②のテキスト生成時のみ，step1 で得られた数値情報を元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．

**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．

**step4. テキストを表現する文法選択**

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．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．

### 4.4 実行例

システムの実行例を図 7 に示す．

図 7 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．

## 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．テキスト生成手法に関しては，数値データが


図 7: システムの実行例

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を捉えることより，その挙動を説明するテキスト生成の手法を提案した．コンテンツのさらなる知的化を目指して，新たなタグを追加し重要度を判断する基準とする等，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の開発，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」[10]（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, No4，pp.424-231, 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. In Journal of the Association for Computing Machinery, 16(2), pp. 264.285, 1969.

[7] Barzilay, R. and Elhadad, M. Using lexical chains for text summarization. In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization, pp.10.17, 1997.

[8] Marcu j, D. From Discourse Structures to Text Summaries. In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization, pp.82.88, 1997.

[9] 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, June, pp.406–416, 2000.

[10] 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.