1S-5

時系列データの形状認識に基づく言語化への取り組み

土橋亮子† 小林一郎‡
†お茶の水女子大学理学部情報科学科
‡お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース

1 はじめに

我々の周囲で観測されるデータの多くは時系列データである．時系列データの解釈はその分野の専門家でない限り，迅速かつ的確に解釈することは難しい．そのため専門家に代わって，時系列データの内容を解釈し，ユーザに情報を提供する手法が望まれる．時系列データの振る舞いを解釈する際，人は大局的な動向を視覚で把握し，言葉で理解する．その点に着目し，本研究では株価や為替の動向について時系列データのグラフの形状を捉え，言葉で表現する手法を提案する．

2 日経平均株価テキスト生成システム

2.1 システム概要

先行研究 [1] において開発された「日経平均株価テキスト生成システム」の概要を図 1 に示す．このシステムを利用し，円・ドルの為替レートの動向を説明する言語化を行う．

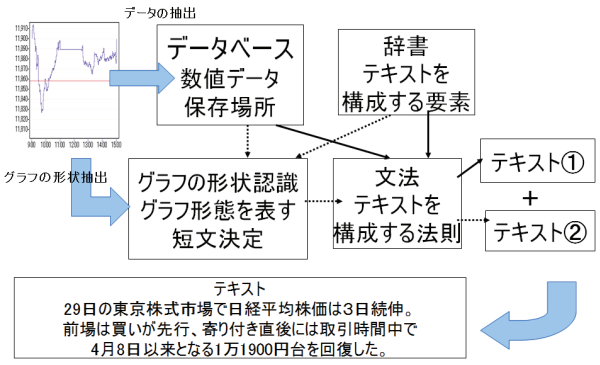


図 1: システムの概要

このシステムによって生成されるテキストは以下の２つのテキストタイプに分類され，タイプごとにテキスト生成の処理の流れが異なる．

テキスト①: 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．

テキスト②： 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．

A study on Verbalizing Time-Series Data based on Its Graphical Shapes
†Ryoko TSUCHIHASHI(g0720548@is.ocha.ac.jp),
‡Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
†Dept. of Information Sciences, Faculty of Science, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610
‡Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

本研究においては，テキスト②の自動生成に着目し，テキスト生成の性能向上およびその評価を行う．以下にシステム各部の説明，および，システムの実行例を示す．

2.2 グラフの形状認識

グラフの動向を把握するとき，グラフが「下がって，上がっている」などの形状によって認識される．グラフを視覚的に把握するために，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表す．近似曲線は 4 次多項式で表現されており，この多項式の次数は，グラフの形状を表現している語彙の実際のコーパス (約 1ヶ月分の日経平均株価動向の解説記事) を分析することにより，その最適な次数を 4 次と導いた．4 次多項式が表現する典型的な曲線の全体的な形状を 11 タイプとし，その形状のパラメータ値のとり方により，さらに 13 種類の部分形状が導けるとした（図 2 参照）．

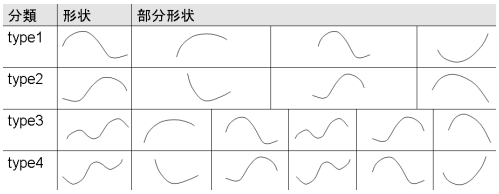


図 2: タイプごとに分類された部分形状 (一部)

この分類は，実際のコーパスから抽出されたグラフの挙動を説明するために使われる語彙表現の観点から導いた．任意の全体形状のタイプはどの部分形状を含むかが決まっているため，4 次多項式で認識されたグラフの形状は，始めに分類された全体形状の特定のタイプを選別する．次に，その部分形状を数式的に解釈することにより最終的なグラフの形状を認識し，これを説明する適切な言語表現を選択する (図 3 参照)．

2.3 辞書

辞書は，実際のコーパスとそれに対応する株価動向を示すグラフの部分形状の対応関係を観測することにより構築される（図 4 参照）．

為替動向の時系列データを表現する語彙として，日経平均株価の動向を表現する語彙と類似している語彙も多々あるが，為替動向を説明するための特有な言葉も多数存在する．そのような語彙を含む辞書を構築し

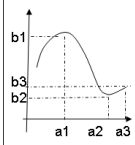
部分形状	短文+時間帯	特徴
	売りが優勢だった	|b2-b1|/|MAX-MIN|>0.4 |a1-a2|/|max-min|<0.7
	売りが広がった	|a1-a2|/|max-min|>0.7
	売りが優勢になる場面があった	|b2-b1|/|MAX-MIN|>0.4 |b2-b3|/|b2-b1|>0.5 |a1-a2|/|max-min|<0.7
	中ごろ過ぎにかけて	|fsh-a2|/|max-min|<0.7 |fsh-a2|/|max-min|>0.5
	中ごろに	|fst-a1|/|max-min|<0.2 |fsh-a2|/|max-min|<0.2
	中ごろ過ぎから	|fsh-a1|/|max-min|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

図 3: 部分形状の数式的解釈とその言語表現

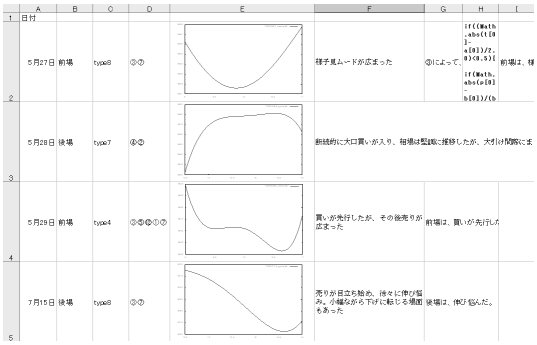


図 4: グラフの形状と語彙の対応

た．為替の動向を示す語彙の辞書構築にあたり，2010年8月2日から2010年9月9日までの114個の円・ドルの為替レートに関する実際のコーパスを分析することにより，グラフの部分形状を適切に表現する語彙・文を収集し，辞書を構築した．

株価動向と為替動向を示す記事の差異は，それぞれの市場の特徴が反映される．株価市場の開場時間は，午前9:00〜11:00，午後12:30〜15:00であるのに対し，東京為替レートは日曜を除き，全世界の市場を併せると24時間体制である．コーパス分析より，為替の動向を説明するには，世界の市場が持つ時間帯に着目する点が株価動向を説明する記事にはない特徴として挙げられることがわかった．時間帯は，大別すると「東京午前」,「東京午後」,「欧州時間」,「ＮＹ時間」の4つに分類されることを確認した．

構築された辞書は，上述のような為替レート説明に特有の表現，および，図3に示すようにグラフの形状を数式的に解釈したものが語彙や文と対応するようにシステム内に実装されている．

辞書内には，部分形状で表現できる短文が約60種類(例：「小動きが続いた」,「もみ合いとなっている」,「反発」)，時間帯が6種類(例：「東京午前」,「序盤」)，接続詞が4種類(例：「その後」,「しかし」)登録されている．

2.4 文法

テキスト は，短文，時間帯，接続詞の適切な組み合わせ規則により生成される．その例を以下に示す．

- 時間帯によって先頭に「東京午前は」,「東京午後は」をつける．
- 部分形状によっては，時間帯によって「序盤」,「現在は」，などが短文の前につけられる．

2.5 実行例

図5では,「2010年9月8日」と入力すると，以下のようなテキストが生成された．

タイプ①テキスト

「8日の東京株式市場で日経平均株価は3日続落．終値は前日比82円62銭（100.0%）東京午前は，小幅な動きが続いた．今日の高値で引けた．1日を通じて高い水準で推移した.」

タイプ②テキスト

「前場は，買いが先行した．その後，底堅さを確認した．後場は，買いが先行した．その後，売りが広がった.」

図 5: システムの実行例

3 おわりに

本研究において，先行研究[1]で構築された株価動向を言語化するシステムを基に，為替レート（円・ドル）の動向を示す言語化を可能にする機能の拡張の一貫として，辞書の拡張を検討した．為替レートの動向を示す言語表現は，株価動向を示す言語表現とほぼ類似しているものの，為替ドメイン特有の表現もあり，コーパスを分析することにより，そのような表現を新たに辞書に追加した．

今後の課題として，実際に為替のデータを用いて為替レートの動向を示す言語化を実現すること，および，株価と為替の変動に関する相関関係の言語化などを行うつもりである．

参考文献

[1] 小林一郎，渡邊千明，奥村奈穂子：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，情報処理学会論文誌，48（3），pp.1058-1070,2007

[2] 加藤，松下: 動向情報の要約・可視化から情報編纂へ，第21回人工知能学会全国大会，2H5-11，(2007).