2J2-NFC2-1

時系列データの言語化への取り組み－日経平均株価を例として－

A Study on Verbalization of Time-series data-With an Example of Nikkei Stock Average Trends-

関亜沙美[*1]　　小林一郎[*2]
Asami SEKI　　Ichiro KOBAYASHI

[*1]お茶の水女子大学　理学部　情報科学科
Dept. of Information Sciences, Faculty of Science, Ochanomizu University

[*2]お茶の水女子大学院　人間文化創成科学研究科理学専攻
Advanced Sciences, Graduate School of Humanities and Science, Ochanomizu University

This paper describes a method to verbalize time-series data, especially we use the trends of Nikkei Stock Average as time-series data. In this study, to verbalize time-series data, we firstly make an approximate function to imitate the behavior of the data and capture the shape of its approximate function, and then verbalize each partial shape of the function with the words obtained from real corpus data, i.e., economic news articles, etc. The generated sentences are evaluated by comparing with the real corpus reporting on the trends of the stock average, and also by humans. With the result of evaluation, we have verified that our proposed method is useful.

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．その時系列データの解釈へのアプローチとしてグラフなどのモダリティに表現を変更する可視化などの手法がある．一方，株価や為替の一日の動向などを示すテキストが新聞やWEBページに掲載されているように，時系列データの振る舞いを言葉で説明するニーズも数多く存在する．そこで，本研究では，時系列データの振る舞いを言葉で説明することに着目し，日経平均株価の動向を例とした時系列データの言語化手法を提案し，システムを開発することを目的とする．

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．

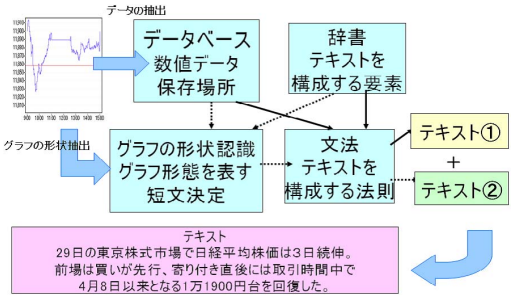


図 1: システムの概要

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．

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，テキスト 2 の生成処理の流れを示す．

2.2　データベース

システムへの入力情報は日経平均株価の分足データと始値・終値・高値・安値である．その数値データを管理するデータベースは MySQL を使用した．日経平均株価の数値データを 2 つのテーブルに分けて管理しており，日付，時間，株価の分足データがまとまったテーブルと，日付，高値，安値，始値，終値の数値データがまとまったテーブルから成る．日経平均株価の分足データを数値データとしてデータベースに入れておくと，データの変更，追加，削除に対応しやすく，また，将来的にリアルタイムでの日経平均株価の分足データの取得を視野に入れているため，保存形態をデータベースとした．

2.3　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，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形状それぞれに対して，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いる．この多項式の次数は，小さすぎるとグラフの挙動を表現しきれない，また，大きすぎると余分な挙動まで表現してしまう．そのため，最適な最小次数を求める必要がある．そこで全データに対しフリーソフトである gnuplot を用いて 4 次近似，5 次近似，6 次近似，7 次近似と行い，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極値の個数などにより 12 タイプとした．その際に極値・変曲点の個数などによりタイプ分けをしている．(図 2 参照) その形状の

連絡先: 小林一郎，お茶の水女子大学院　人間文化創成科学研究科理学専攻　小林研究室，東京都文京区大塚 2-1-1，03-5978-5709，koba@is.ocha.ac.jp

パラメータ値のとり方により，さらに 13 種類の部分形状を導いた（図 3 参照）．

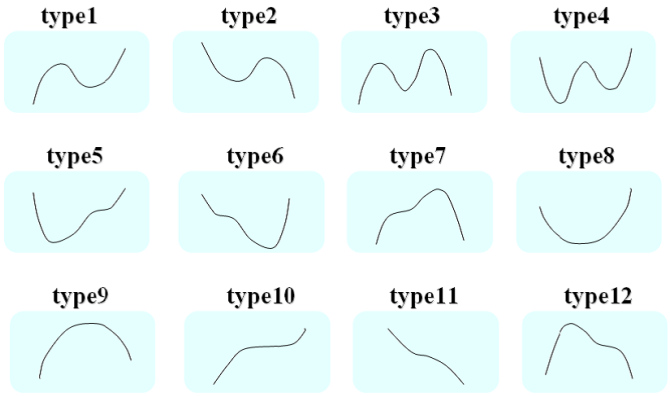


図 2: 12 タイプの部分形状

分類	形状	部分形状
type1		
type2		
type3		

図 3: タイプごとに分類された部分形状 (一部)

図 3 に示す分類は，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．グラフの全体形状を示す 11 のタイプはグラフの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るグラフの形状は，始めに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する (図 4 参照)．

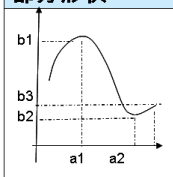
部分形状	短文+時間帯	特徴
	**売りが優勢だった**	|b2-b1|/|MAX-MIN|>0.4 |a1-a2|/|max-min|<0.7
	売りが広がった	|a1-a2|/|max-min|>0.7
	売りが優勢になる場面があった	|b2-b1|/|MAX-MIN|>0.4 |b2-b3|/|b2-b1|>0.5 |a1-a2|/|max-min|<0.7

図 4: 部分形状の数式的解釈とその言語表現

2.4 辞書作成

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 5 参照）．

辞書構築にあたっては，先行研究において使用された 2005 年 7 月 25 日から 8 月 30 日までの 27 個，2009 年 5 月 20 日から 7 月 24 日までの 20 個の株価動向を表す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

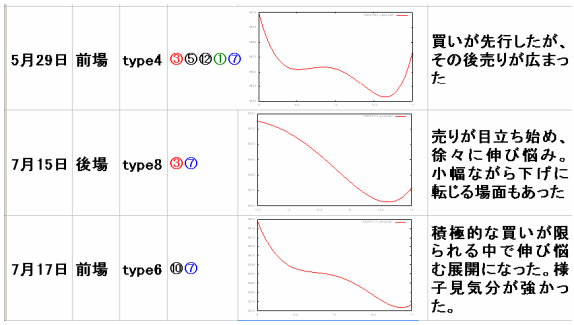


図 5: グラフの形状と言語表現の対応

彙，文を収集し，辞書を構築した．構築された辞書は，図 4 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．

現時点において，辞書内には，部分形状を表現できる短文が 64 種類 (例：「売りが広がった」,「じり高歩調となった」,「反発」)，時間帯が 9 種類 (例：「前場」,「大引けで」)，接続詞が 4 種類 (例：「そして」,「なので」) 登録されている．

2.5 文法

テキスト 2 は，短文，時間帯，接続詞の実際のコーパスを真似た適切な語彙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「前場は」,「後場は」をつける．
- 部分形状によっては，時間帯によって「中ごろ過ぎにかけて」,「中ごろに」，などが短文の前につけられる．
- 上げ幅，下げ幅について言及している短文はその前に接続詞「そして」が前につけられる．

2.6 テキスト生成過程

step1.データベースからの数値情報の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
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．

2.7 実行例

システムは，3 つのウインドウで構成されており，MySQL に蓄積されている日経平均株価データを読み出し表示するウインドウ，グラフを表示するウインドウ，グラフの動向を言語で説明するウインドウからなる．図 6 では,「2009 年 10 月 14 日」と入力すると，以下のようなテキストが生成された．

「後場は，寄り付き後, 売りが優勢だった．その後，底堅さを確認した．その後，買いが入った.」

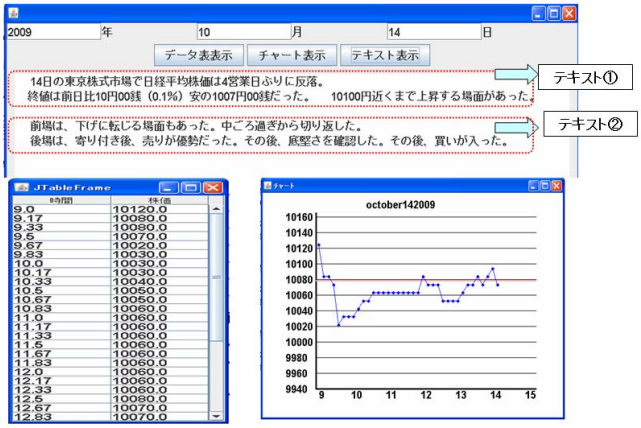


図 6: システムの実行例

2.8　生成されたテキストと実際のコーパスとの比較

生成されたテキストと実際の日経平均株価のコーパスとの比較を行った．2009 年 5 月 28 日を例として，システムより生成されたテキストを以下に示す．

「前場は，買いが入った．相場は堅調に推移した．上昇に転じる場面もあった．後場は，相場は堅調に推移した．上昇に転じる場面もあった．まとまった売りが出た．」

実際の日経平均株価のコーパスを以下に示す．

「28 日前場中ごろの東京株式市場で日経平均株価は上昇に転じる場面があった．相場は堅調に推移したが，大　引け間際にまとまった売りが出て上げ幅を縮小した．」

枠で囲まれた箇所が表現が一致する箇所である．一致する表現が多く見られることがわかる．

3.　システムの評価

表 1 に実際のコーパスとシステムによるテキスト生成結果を比較したものを示す．

表 1: 評価

グラフ特徴	実際のコーパス	コーパスに対する一致		グラフの挙動に対する一致
		完全	同意	
状態	4	1	2	11
変化率	25	5	23	12
変動量	6	1	3	11
その他	16	3	16	13
合計	51	10	44	47

辞書中の語彙表現を「状態」「変化率」「変動量」「その他」の 4 つの種類に分類し，それぞれに対して一致度を評価した (その他の項目には，“もみ合い” など方向性のないテキスト表現が入っている)．また，上図の「同意」を考慮した一致度とは，語彙が完全に一致していなくても，「売りが強まった」と「売りが広がった」など，同じグラフの挙動を意味しているものの一致度のことを指す．「同意」を考慮した一致度の定義としては，辞書のパラメータ設定から図 7 のように，同じ「売り」の挙動を示す語彙の中で包含関係となっているものを「同意」（正確には含意）と考える．例として，「売りが広がった」と「売りが優勢になる場面があった」は，包含関係にあるため同意とみなされるが，「売りが強まった」と「売りが膨らんだ」は，包含関係にはないため同じ「売り」という単語を含んでいても，同意とはみなされない．

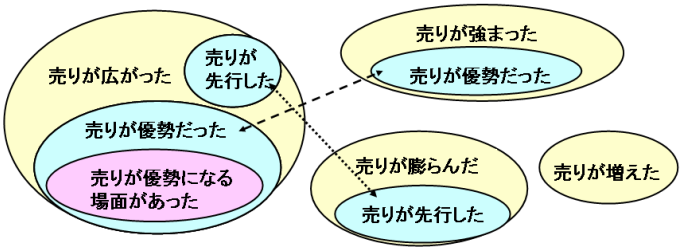


図 7: 言語表現の同意関係

表 1 より，同意を考慮した一致度の場合，システムのコーパスに対する一致度は 0.86(44/51) となり，作成した辞書が適切であり，また，言語化するグラフ挙動の箇所が同じである割合が高いといえる．一方，ひとつのグラフ挙動に対して，それが変化率の特徴で言語表現される場合もあれば，状態の特徴で言語表現される場合もあり，一概にグラフの同じ特徴に対して一致する生成されたテキストの数によってテキスト生成システムの性能評価はできない．このことを考慮して，表 1 中の列項目における「グラフの挙動表現に対する一致」は，同じ期間の同じグラフに対して生成されたテキストの内，グラフの各特徴において正しくグラフの挙動を表現するテキストの数を示している．この表の数値からは，実際のコーパスにおいて，対象とする期間内のグラフの挙動を 51 個（4+25+6+16）のテキストを用いて表しているのに対して，システムはグラフの挙動を表現する 47 個（11+12+11+13）の適切なテキストを生成していることがわかる．システムが生成したテキストの中には実際のコーパスに存在しないものもあるが，それはコーパスの作成者が気がつかなかったグラフの特徴をシステムが言葉で明示化したものとも考えることができる．このことからも我々の開発したシステムがグラフの挙動を説明するのに十分なテキスト生成が行えていることがわかる．

4.　おわりに

本研究において，株価動向を示す時系列データの言語化手法およびその性能評価を示した．言語化手法において，辞書を強化することによりテキストの生成能力の向上を計り，また，性能評価についてはテキスト生成性能の評価を行うひとつの方法を提案した．今後の課題としては，より客観的な性能評価の指標としてグラフの形状認識において得点制を導入するなど，テキスト生成の性能評価指標を確立することや，株価のリアルタイムに基づく言語化などを行うつもりである．

参考文献

[1] 奥村菜穂子，グラフの挙動を表すテキスト生成，2005 年度お茶の水女子大学卒業論文，(2006).
[2] 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).
[3] 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).
[4] 加藤，松下: 時系列情報の抽出と可視化に基づく情報アクセスためのマルチモーダルインタフェース –情報編纂の基盤技術に向けて–，人工知能学会論文誌，Vol. 22，No. 5，pp. 553 - 562，(2007).