日経平均株価を対象とした時系列データの言語化への取り組み

関亜沙美† 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．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言語化が存在する．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[?]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

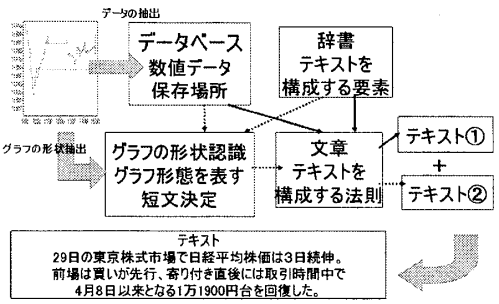


図1: システムの概要

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　: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　: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　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　の生成処理の流れを示す．

A study on Verbalizing Time-Series Data with an Example of Nikkei Stock Average
†Asami SEKI(g0620524@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

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5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5次と導いた．5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）．

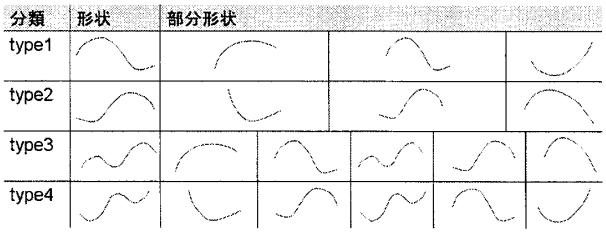


図2: タイプごとに分類された部分形状(一部)

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，5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する(図3参照)．

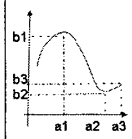
部分形状	短文+時間帯	特徴
	売りが優勢だった	|b2-b1|/|MAX-MIN|>0.4 |a1-a2|/|max-min|<0.7
	売りが広がった	|a1-a2|/|max-min|>0.7
	売りが優勢になる場面があった	|b2-b1|/|MAX-MIN|>0.4 |b2-b3|/|b2-b1|>0.5 |a1-a2|/|max-min|<0.7
	中ごろ過ぎにかけて	|fsh-a2|/|max-min|<0.7 |fsh-a2|/|max-min|>0.5
	中ごろに	|fst-a1|/|max-min|<0.2 |fsh-a2|/|max-min|<0.2
	中ごろ過ぎから	|fsh-a1|/|max-min|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

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

2.3 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 4 参照）.

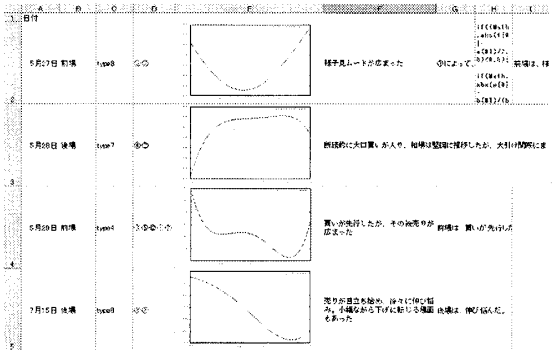


図 4: グラフの形状と語彙の対応

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彙，文を収集し辞書を構築した．構築された辞書は，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 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．

2.4 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「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．

2.5 実行例

図 5 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.」

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．

辞書中の語彙表現を「グラフの状態」「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である．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．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．

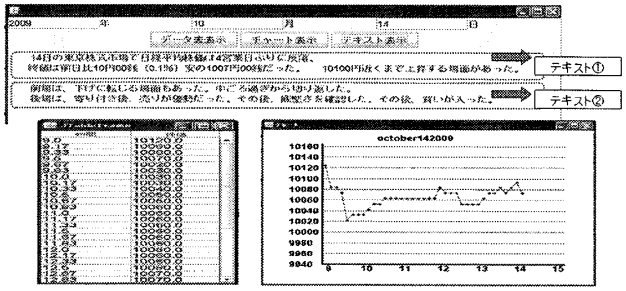


図 5: システムの実行例

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			グラフの挙動表現に対する一致
		完全	同意	不適切	
グラフの状態	4	1	2	3	11
変化率	25	5	23	2	12
変動量	6	1	3	0	11
その他	16	3	16	0	13

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．今後の課題としては，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] 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).

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.