時系列データの解説を行うテキストの自動生成

渡邉 千明　(指導教員：椎尾 一郎)

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

株価の動向を言葉で解説するためには，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には，分足ベースの 1 日の株価データが利用できる．これらを用いて，大まかな情報をテキストで提供する際は，文章の要約技術を，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．

3　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．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 ．

4　解説テキスト生成手法

長期動向のテキスト生成では，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

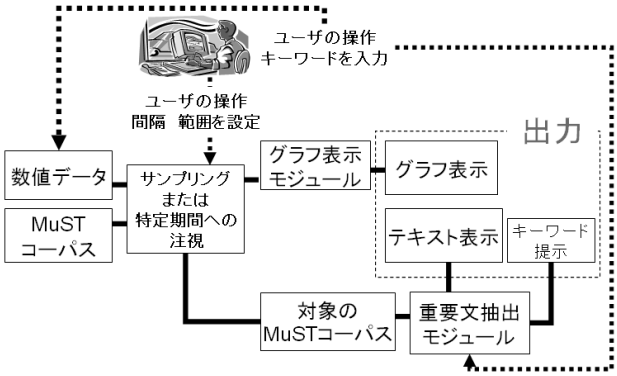

図 1: システム構成図

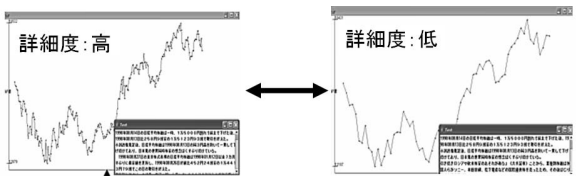

図 2: 実行例 (グラフの目盛り間隔の変更)

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．要約対象となる文の重要度の決定方法は，tf・idf 法を利用し，ある文章における名詞の相対的な重要度を算出する．各行の文の重要度を初期値 0 として始め，その文に含まれている名詞に対し，tf・idf 法の計算で算出された，各単語の重要度を足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．また，キーワードを入力することができ，ニュース記事に含まれている，キーワードが含まれる文の重要度が高くなるように計算する．グラフも同様に，キーワードと関係している数値データを利用して表示させる．この二つを同時に表示させ，グラフとテキストを協調させる．また，そこで新たに表示されたグラフから範囲を選択することも出来る．このように，グラフの表示詳細度，キーワードを繰り返しユーザが設定することができ，ユーザが望む情報をユーザが望む詳細度で得ることができる．要約処理部のシステム構成を図 1 に示す.

4.1　長期動向解説テキストの生成

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．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 2 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．

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.

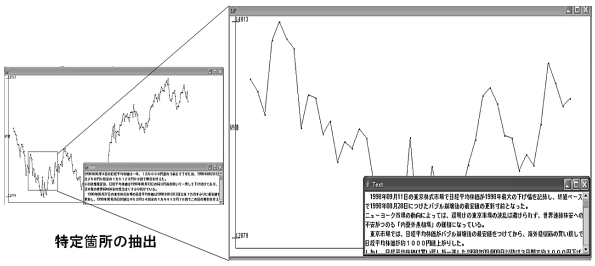

図 3: 実行例 (特定箇所の情報抽出)

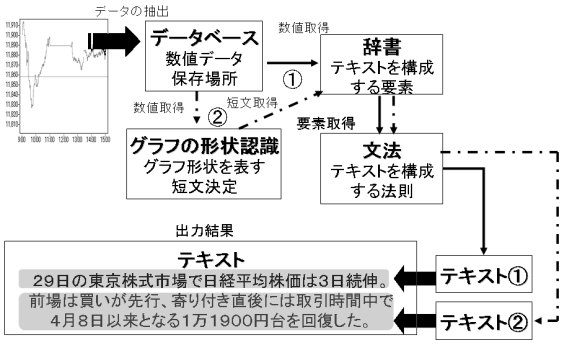

図 4: システム構成図

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（図 3 参照）．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．ここで選択された範囲が 1 日という短期的な場合は，以下で解説する．

4.2 短期動向解説テキストの生成

選択された範囲が 1 日という短期的な場合，テキスト生成システムを実行する．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，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タイプ（1）のテキスト，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タイプ（2）のテキストに分類され，タイプごとにテキスト生成の処理の流れを変える．システムの構成を図 4 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 4 中，実線および一点鎖線でそれぞれ示す．図 5 に短期動向解説テキストの生成例を示す．

以上のようにして作成されたテキストは，必要に応じて音声合成ソフトを使い読み上げられる．Galatea Toolkit という擬人化エージェントツールキットのうちの，音声合成フリーソフト Galatea Talk を使用している [7].

図 5: 実行例（特定箇所の情報抽出）

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手法を提案した．その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，ユーザのグラフに対する表示操作からその意図を判断し，テキスト要約・生成手法を用いてユーザの求める情報を提示するシステムの実装を行った．今後は，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を開発する予定である．

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」[5]（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] 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.

[5] 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.

[6] 奥村奈緒子，小林一郎：グラフの挙動を表すテキスト生成，言語処理学会第 12 回年次大会ワークショップ「言語処理と情報可視化の接点」，pp.17-18，2006.

[7] http://hil.t.u-tokyo.ac.jp/ galatea/index-jp.html

[8] 小林一郎，渡邉千明，奥村奈緒子：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 Vol.48 No.3 Mar.2007