グラフの表示に基づいた要約文生成システムの提案

渡邉 千明 †
chiaki@koba.is.ocha.ac.jp

小林 一郎 ‡
koba@koba.is.ocha.ac.jp

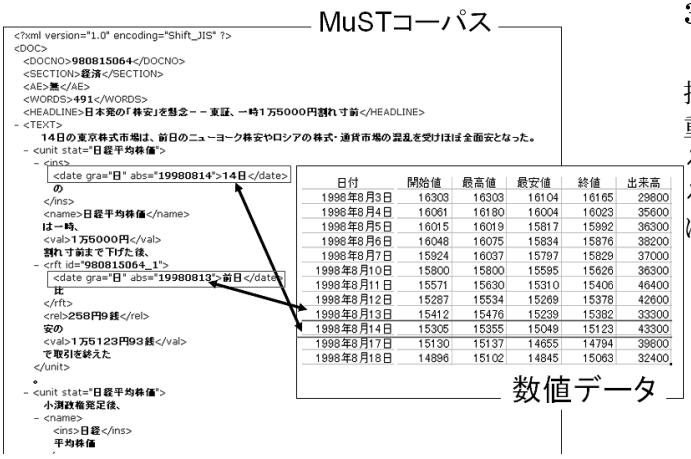

図 1: MuST コーパスと日経平均株価の数値情報

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．この現状を踏まえて，情報の内容や表示を誰にでも理解されやすいよう，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もつ知的情報提示システムの開発を試みる．具体的には，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ことを目的とし，グラフの表示状態に協調してテキストの詳細情報も変化するテキスト要約手法の機能を備えたシステムの開発を行う．

2. 提案手法

本研究では，対象となるニュース記事から重要文を抜き出すことにより，ユーザによって変更されたグラフの状態に対応している要約文を生成する．本システムでは，日経平均株価の数値情報と，MuST コーパスによって得られるその日の株価の動向情報を対応させ，グラフとテキストを関連付けておく．MuST コーパスと日経平均株価の数値情報の対応関係は，MuST コーパスにおける株価の値に言及しているタグの値より確認できる（図 1 参照）．

†お茶の水女子大学 人間文化研究科 数理・情報科学専攻
‡お茶の水女子大学理学部情報科学科

3. 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 [4].

1. テキスト中の単語の重要度 [5, 6].
2. テキスト中あるいは段落中での文の位置情報 [7].
3. テキストのタイトルなどの情報 [7].
4. テキスト中の手がかり表現 [7].
5. テキスト中の文あるいは単語間のつながりの情報 [8].
6. テキスト中の文間の関係を解析したテキスト構造 [9].

本システムでは，tf・idf 法，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させるため，上記手法（1）,（3）と（4）を利用する．

3.1 要約対象となる文の重要度の決定方法

- **tf・idf 法**
  重要文抽出法 (1) の利用として，tf・idf 法を利用した以下の計算式を使い，各単語の文章における相対的な重要度を算出する．

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

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．

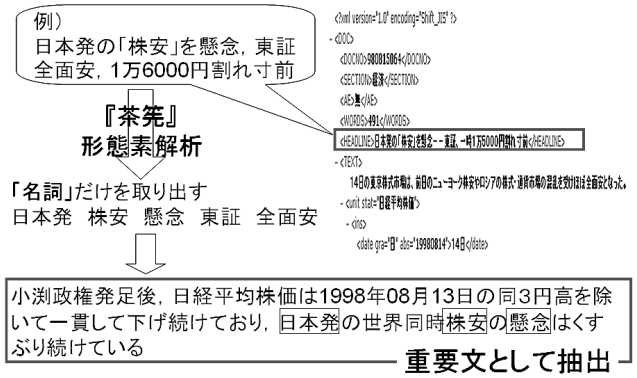

図 2: HEADLINE タグの利用

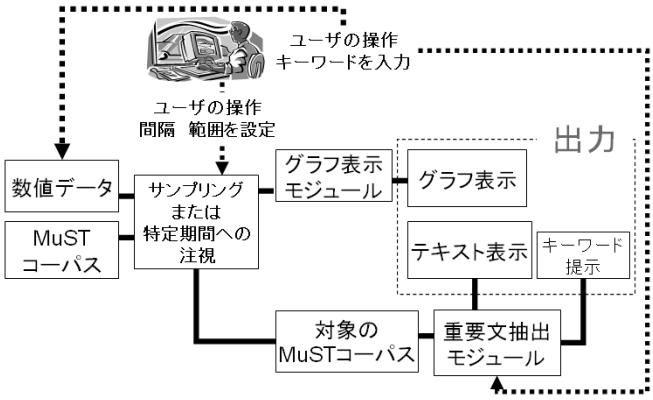

図 3: システム構成図

- **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．このことを重要文抽出に反映させるため，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．そのため，各名詞の tf・idf 値を計算した後に，HEADLINE タグが付与されている文中に存在する名詞にタグ別ポイントを加算する．その加算する値は，事前に既定値を与えるのではなく，そのとき計算された tf・idf 値の最大値から平均値を引いた差を，タグ別ポイントとして加算することにより，タグ別ポイントを全体の名詞の重要度に対して，相対的に与えることができるようにする．処理の流れを図 2 に示す.

- **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．また，unit タグが付与されている文の中でも,「前日比」というように，短い期間の挙動について言及しているものから,「12 年 8ヶ月ぶりに」というように，長い期間を言及している文もある．期間が長いということは，滅多に起こらないことが起こったということだと考え，より重要な文とする．このとき，期間の表現に付与されている dur タグの属性である gra 属性を参考にし，長い期間か短い期間なのかを判断している．扱っている記事の中では,「日」と「月」が存在していたため,「月」のときのみ加算した．さらに，グラフ上で表されているものが日経平均株価なので，それに対応するように,「業種別株価」について言及している文より,「日経平均株価」について言及している文を重要とする．このため，unit タグの属性である stat 属性を参考にし，日経平均株価となっている文に加算した.

4. システム構成

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．さらに，キーワードを入力することができる．キーワードが入力されると，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．また，株式の重要銘柄についても同様に情報を得ることが出来る．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．システム構成を図 3 に示す.

4.1 数値データ・MuST コーパスのサンプリング

グラフの目盛り間隔の変更

グラフの目盛り間隔が変更され粗くなった場合，ユーザは細かい流れよりも全体の傾向が知りたいと思うようになると考えられる．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（図 4 参照）．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

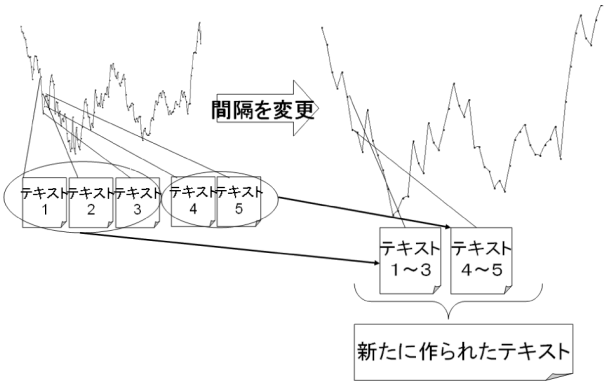

図 4: グラフの目盛り間隔の変更

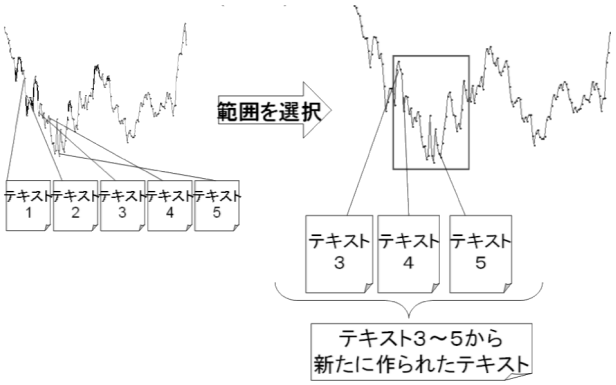

図 5: 特定箇所の情報抽出

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．また，目盛り間隔が広くなれば抽出されるテキストが減り，情報の詳細度は低くなる．

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（図 5 参照）．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．

範囲が狭くなればなるほど，変更する前には抽出されなかった重要度の低い文も抽出されるようになり，その範囲のみをより詳しく説明した要約文となる．ユーザは，新しいテキストから重要なニュースが起きている部分を判断し選択することで，そのニュースに関係する情報を詳細に表示させることもできる．

キーワードの入力

要約文が生成されると同時に，対象テキスト中のキーワードが出力される．ユーザは，その中に興味のあるキーワードがある場合はそれを用いて入力し，新たに興味あるキーワードがある場合はそれを入力することができる．ユーザが入力したキーワードが含まれている文に，unit タグのときと同様に，文の重要度の最大値と平均の差を足す．それにより，文の重要度が変化し，再度重要文を選びなおす．もし，ユーザが「安田信託銀行」などのキーワードで，数値データが取り出せるものを入力したときに関しては，再度数値データをグラフ表示し，安田信託銀行と関連するテキストと同時に表示する．これにより，

図 6: キーワードの入力

ユーザが望む情報を，より柔軟に取り出せるようにする．図 6 に，キーワードを入力することによって得られた結果を示す．上のグラフとテキストは，最初に出てくる，日経平均株価のグラフとテキストである．下のグラフとテキストは，キーワード「安田信託銀行」と入力したことで得られるグラフとテキストである．

5. 性能評価

MuST コーパスで提供されている 1998 年 9 月 9 日から 1998 年 10 月 24 日までの記事全 76 文のテキストを被験者 10 人に与え，ニュースとして重要な事柄について言及していると思われる文を 10 文選んでもらう．その結果を表 1 に示す．被験者は，大学生 10 人，22 歳～25 歳であり，株価に対する知識は，新聞を読み理解できる程度であり、知識の偏りがない被験者を選別して実験を行った．システムおよび被験者 10 人が選んだ文番号には○印を記す．この結果からシステムの性能を評価するため，次のような計算を行う．文番号 0 の文は 9 人の被験者が選んでいるので 9 点，文番号 1 の文は 5 人選んでいるので 5 点というように，1 つの文を選んでいる被

表 1: 被験者実験によるシステムの評価

Statement	System	Subject A	Subject B	Subject C	Subject D	Subject E	Subject F	Subject G	Subject H	Subject I	Subject J
0	○	○	○	○	○	○	○	○	○	○	
1		○			○		○		○		○
2						○					
4			○	○						○	
7	○		○							○	
8	○		○		○					○	
10								○			
15	○		○		○	○	○		○	○	
16			○		○		○				
17						○					
18						○					
19	○	○		○							
22									○		
23	○			○	○						
24				○							
25								○			○
29								○			○
30		○							○		○
33							○	○	○		
34				○			○				
37				○							
39					○					○	
44		○									○
45								○			
50	○		○	○	○	○	○		○	○	○
51	○	○	○				○				
52						○			○		
53		○						○			○
54			○							○	
58	○		○	○						○	
61		○		○	○	○	○	○			○
62						○			○	○	
63						○	○		○		
66					○						
70	○										
71		○									○
73		○						○			
74								○			○

験者の人数をそれぞれの文の点数とする．選んだ文につけられた点数を累積計算をする方法で，被験者 10 人の総得点をそれぞれ計算する．その平均値を求めると 40.6 点となる．システムの点数は 38 点となり，平均値には至らなかったが，多くの被験者が選んだ文を重要と判断できていたと思われる．

表 1 より，7 人の被験者が選び，システムが選ばなかった文 (文番号 61) があることが分かる．この文は，次に挙げる (a) の文である．そして，システムが選んだ文で，被験者が選ばなかった文 (文番号 70) もある．この文は次に挙げる (b) の文である．

(a) 業種別株価では，金融システム不安を背景に大きく売り込まれてきた銀行株が急反発している

(b) 東京株式市場の日経平均株価がバブル後の安値を更新したのも，銀行株が主導した

この 2 つの文は，同じ記事の中にある．日経平均株価についての文をより重要としていたため，業種別株価を言及していた文よりも重要としていたと考えられる．今後は，数値データから，幅がある部分，最小値をとっている日，最大値をとっている日を判断し，重要文を抽出する基準として付け加えていきたい．

6. まとめ

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，グラフとテキストの異なる 2 つのモダリティ情報を用いた，グラフの表示状態に対応したテキストの表示を行った．これにより，ユーザがグラフを変更し，キーワードを入力することで，ユーザの情報閲覧の焦点を判断し，その要望に対応した提示方法を提案した．また，表示された情報から，さらに関連する情報を閲覧することができる．ユーザのその時々の興味にあわせて，ユーザの望む情報提示をより柔軟に行うことができる．今後の課題として，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コンテンツの開発を進める．

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」[3]（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] 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.

[4] 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.

[5] Luhn, H. P. The automatic creation of literature abstracts. IBM journal of Research and Development, Vol. 2, No. 2, pp. 159.165, 1958.

[6] Salton, G. Automatic Text Processing. Addison-Wesley, 1989.

[7] Edmundson, H. P. New methods in automatic extracting. In Journal of the Association for Computing Machinery, 16(2), pp. 264.285, 1969.

[8] 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.

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