情 報 学 基 礎 61−19
自 然 言 語 処 理 142−19
（2001. 3. 6）

株価データと新聞記事からのマイニング

小川 知也　渡部 勇
{ogawa.tomoya, watanabe.isamu}@jp.fujitsu.com
富士通研究所
〒211-8588 川崎市中原区上小田中 4-1-1

本稿では，株価データと新聞記事の関連性に関するマイニング手法について論ずる．本手法では，分類により新聞記事に付与されたテーマ情報を用い，新聞記事の株価変動への影響分析と株価変動の外部要因分析という相補的な視点から，株価データと新聞記事の関連性の概要抽出を行う．新聞記事の株価変動への影響分析では，あるテーマの新聞記事が一般に株価変動に及ぼす影響を分析する．株価変動の外部要因分析では，大きな株価変動の要因を新聞記事のテーマを用いて分析する．実際にいくつかの銘柄に関する株価データと新聞記事を用いた実験を行い，実データにおける本手法の有効性を確認した．

Mining of Stock Prices and News Articles

Tomoya OGAWA　Isamu WATANABE

Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 Japan

In this paper, we propose a mining method for finding a correlation between stock prices and news articles. We use a theme information of news articles which assigned by classification system, and extract a correlation between stock prices and news articles by analysing how news articles influence on stock prices and what kind of news articles causes a stock price changes. Through some experiments on stock prices and news articles, we proved a effectiveness of our method.

1 はじめに

電子化された大量の株価データや新聞記事が利用できるようになってきた．新聞記事などのニュースは投資判断のひとつの重要な素材であり，株価データと新聞記事の間には関連性があると考えられる．もしそのような関連性を抽出することができれば，新聞記事を用いて株価変動を予測する，あるいは投資家のための有効な情報提供を行うことができる．

そこで我々は，株価データと新聞記事からそれらの間の関連性の自動抽出を試みた．

我々のアプローチの特徴は，株価データと新聞記事の関連性の分析を行う際，あらかじめ分類システムにより新聞記事に付与されたテーマ情報を用いること，そして新聞記事の株価変動への影響分析および株価変動の外部要因分析という相補的な視点からの分析を行うことである．

テーマ情報を利用することで，抽象化された理解容易な情報を得ることが可能となる．

新聞記事の株価変動への影響分析では，あるテーマの新聞記事が一般に株価変動に及ぼす影響を分析する．株価変動の外部要因分析では，大きな株価変動の要因を新聞記事のテーマから分析する．このように相補的な視点からの分析を行うことで，いろいろな面での株価データと新聞記事の間の関連性抽出が可能となる．

2 関連研究

株価データとニュース記事からのマイニングを行う研究としては，Web上のニュース記事から株価変動予測を行うAEnalyst[1][2]がある．本研究よりも細かい時間情報に基づいた短期的な株価変動予測を行っている．実際に投資シミュレーションを行った結果，利益を上げることができたと報告している．

システムの活動をモニタリングし，その時系列的変化を検出する研究としては，Activity Monitoring[3]がある．評価の一つとして，ニュース記事により大きな株価変動がもたらされるかどうかの判定を行っている．

3 基本方針

株価データと新聞記事の間の関連性を探る際，我々は次のことが重要であると考えた．

- 変動の要因が理解容易な形で把握できること

株価データと新聞記事の関連性を自動抽出することができるかは必ずしも明確ではない．新聞記事を用いることで，長期間にわたる記事データの利用が可能となるが，その反面1日よりも細かい時系列情報を利用することは難しい．

新聞記事には，提携や新製品開発などその記事で伝えようとするテーマが存在することが多い．そこで我々は，AEnalystのアプローチとは異なり，株価変動と新聞記事の関連性を新聞記事のテーマに注目して抽出することにした．これにより，抽象化された理解容易な形で情報を得ることができる．また，株価変動に大きな影響力のある情報を何らかの理解し易い形で得ることができれば，その情報の有用性が増すと思われる．

また，いろいろな角度から株価データと新聞記事の間の関連性を分析するため，次の2つの相補的な視点から分析することとした(図1)．

- 新聞記事の株価変動への影響分析
- 株価変動の外部要因分析

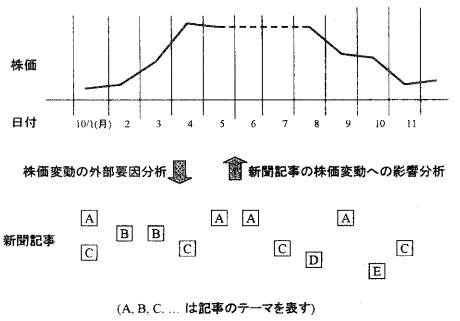

図1: アプローチ

新聞記事の株価変動への影響分析は，あるテーマの新聞記事が株価変動に一般にどのような影響を与えるかを分析する．

株価変動の外部要因分析は，大きな株価変動の要因の分析を新聞記事のテーマを用いて行う．これにより，どのようなことが要因で株価変動がもたらされるかが分かり，これを進めることで新聞記事を用いた株価変動の予測にもつながるものと期待される．

新聞記事の株価変動への影響分析を行う場合は，同じテーマに属する各新聞記事について，その記事の発行時点における株価変動を統計的に求める．

株価変動の外部要因分析を行う場合は，大きな株価変動があった時点における新聞記事に共通してみられるテーマを抽出する．

なお今回の実験においては，新聞記事データ管理などに，文書セットを様々な視点から分析を行える情報分析ツール ACCENT[4, 5] を利用している．

4 分析手法

4.1 株価変動

株価変動としては，単位期間辺りの株価変動の比 r を用いる．すなわち，

$$r = (y - y0)/y0$$

ここで，y は当日の株価，y0 は単位期間前の株価である．

株価としては，四本値の終値を用いる．

実際には，ある銘柄の株価変動を分析する際その一銘柄に注目するだけでは必ずしも十分ではない．株価は，同業他社の銘柄や市場全体と同じような動きを示す場合があり，このような条件を考慮に入れてモデル化する必要がある．そこで，複数銘柄の比較による株価変動の抽出を行う．

銘柄 B に対する銘柄 A の株価変動の抽出を行うには，銘柄 A の株価変動と銘柄 B の株価変動との相対的な値を用いる．相対的な値としては，2 つの銘柄の株価変動の差 $r_{A/B}$ を用いる．

$$r_{A/B} = r_A - r_B$$

ここで，r_A は銘柄 A の株価変動，r_B は銘柄 B の株価変動である．なお，以降「A / B」で銘柄 B に対する銘柄 A の相対的な株価変動を表す．

実験では，市場全体の動きを表すものとしての TOPIX(東証指数) に対する各銘柄の相対的な株価変動を用いた．

株価変動を抽出する単位期間を変更することで，異なる視点での変動抽出を行うことができる．

実験対象とした 1997 年 1 月から 2000 年 5 月までの富士通と TOPIX の株価を図 2 に，単位期間が 1 日の株価変動を図 3 に，単位期間が 1ヶ月の株価変動を図 4 に，それぞれ示す．

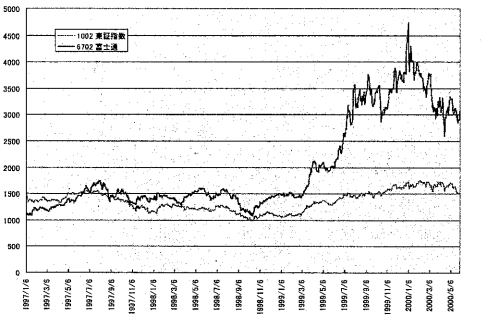

図 2: 富士通と TOPIX の株価

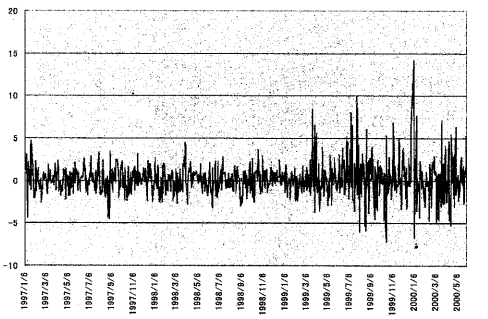

図 3: 富士通 /TOPIX の 1 日単位の株価変動

これをみると，単位期間が 1 日の場合には 1999 年暮れから 2000 年明けにかけてが最も大きな株価上昇を示していたが，単位期間 1ヶ月ではその時期は必ずしも最も大きな変動ではなく，代わりに 1999 年 7 月頃の株価上昇が最も大きな株価上昇であることがみえてくる．

次に，株価変動の大きい時期の新聞記事を実際に調べてみる．単位期間 1ヶ月で最も大きな株価上昇である 1999 年 7 月の中で，単位期間 1 日で大きな上昇を示した 1999 年 7 月 28 日近辺の新聞記事見出し一覧を図 5 に示す．

見出しから，この時の株価上昇の要因としては，7 月 27 日の「さくら銀行と富士通、共同でネット専業銀行を設立」などが考えられる．こうして，確かに株価変動の大きい時期近辺に株価変動に大きな影響を与えそうな新聞記事があることが確かめられる．しか

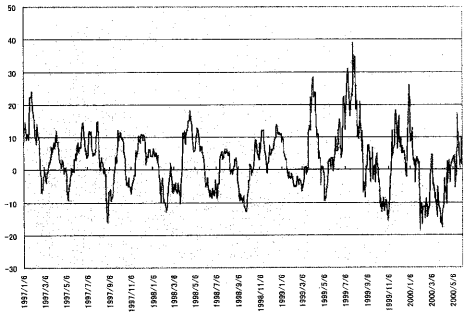

図 4: 富士通 /TOPIX の 1ヶ月単位の株価変動

1999/07/26 郵政省、「宅内情報通信・放送高度化フォーラム」を立ち上げ。1系統化目指す
1999/07/26 検証/次世代ネットビッグ3、国内通信機業界揺らす
1999/07/26 富士通、セキュリティー機能付きMO装置技術を開発
1999/07/27 さくら銀行と富士通、共同でネット専業銀行を設立
1999/07/27 富士通北陸システムズ、オラクルのリナックス版RDBMSを販売・サポート
1999/07/27 富士通、音声通信のIP化対応PBXを発売。10月末から出荷
1999/07/28 富士通、超大型UNIXサーバを正式発表－サン製品を上回る最上位機
1999/07/28 あさひ銀行、4次オンラインシステム開発でベンダー3社と協力体制構築

(日刊工業新聞社提供)

図 5: 1999.7.28 近辺の富士通関連新聞記事見出し一覧

し, 実際にどの記事が株価変動に大きな影響を与えているかは必ずしも明らかではない. それを過去の株価データと新聞記事の分析に基づき明らかにしていくのが本研究の狙いのひとつである.

4.2 新聞記事とテーマ

新聞記事は, 日刊工業新聞の記事データを用いた. 総記事数は 145,558 件, うち富士通関連記事は 2,189 件である.

各新聞記事には, 分類システム JSort[6] によりテーマ付与を行う. JSort はルールベースの分類システムおよびカテゴリー体系である. 分類カテゴリーとしてのテーマ数は 158 であり, テーマは階層構造を成す. 今回はすべての階層のテーマを分析対象とした. ひとつの記事に複数のテーマが割り付けられることがある. また, ひとつもテーマが割り付けられない記事も存在する. 富士通関連記事の場合, 割り付けられているテーマ数は 115, ひとつ以上のテーマが割り付けられている記事数は 1,660 である.

なお, 低頻度 (頻度が 9 以下) のテーマはノイズになり易いため実験対象から除外した. 富士通関連記事の場合, 除外されなかったテーマ数は 57 である.

4.3 新聞記事の株価変動への影響分析

新聞記事の株価変動への影響分析手順は次の通りである.

1. 分析対象銘柄の記事を検索する
2. その記事集合について, 各記事に付与されているテーマの集合を求める
3. 各テーマについて, そのテーマに属する記事の発行日における株価変動を求める
4. 株価変動値を統計処理する

各テーマ t_i (件数 n_i) に対応する株価変動値に関しては平均値 μ_i と標準偏差 σ_i を求める. この値が, 各テーマの株価変動への影響と考えられる.

その中から株価変動への影響の大きなテーマを抜き出すために, 分析対象銘柄の全体的な株価変動値の平均 μ と標準偏差 σ に対する検定統計量 Z

$$Z = \frac{\mu_i - \mu}{\frac{\sigma}{\sqrt{n_i}}}$$

を求める [7]. この検定統計量の順にテーマをソートし, その両端から設定した有意水準を満たすテーマを選択することで, 株価変動に関連が深いテーマを抽出する.

4.4 株価変動の外部要因分析

株価変動の外部要因分析手順は次の通りである.

1. 株価変動の抽出を行う
2. 株価変動と新聞記事の対応付けを行う
3. 株価上昇および株価下降に対応する新聞記事群の特徴テーマを抽出する

株価変動の抽出

株価変動の抽出では, 単位期間辺りの変動が, ある閾値以上の変動を抽出する.

抽出する期間が多過ぎると, 変動があまり大きくない時期の記事まで含まれることによるノイズの増加がある. 逆に, 抽出する期間が少な過ぎると, 変動が大きな時期にたまたま含まれた記事がノイズとして含まれることになる.

そこで, 閾値を株価変動の標準偏差 σ を目安に設定し, 複数のクラスの変動を抽出するようにした.

株価変動と新聞記事の対応付け

株価変動と新聞記事の対応付けに関しては, 基本的に株価変動の単位期間の開始時点周辺に発行された新聞記事を対応付ければよいと考えられる.

最近ではインターネットからの情報入手などにより, 新聞記事よりも早く情報を入手し, それを元に投資行動を起こすということも考えられる. その場合には, 株価変動の開始時点周辺よりも遅い時点の新聞記事とも対応付ける方がより適切ということも考えられる. 今回はその影響は小さいとし, 株価変動の単位期間の最初の 1 日を新聞記事との対応付けに用いた. 株価としては終値を用いているため, 新聞記事の発行時点が株価変動の期間よりもほぼ先立っており, 株価予測などに本手法を適用する場合にも都合が良いと考えられる.

特徴テーマの抽出

株価変動に対応する新聞記事群の特徴テーマの抽出手順は, 次の通りである.

1. 特徴素選択の手法に基づき, 各クラスの特徴テーマの候補を抽出する
2. 最大エントロピー法に基づき, 抽出された特徴テーマ候補が適切かどうか確認する

特徴素選択にはいくつかの手法がある [8], [9] が, Kullback-Leibler 情報量に基づく方法などいくつかの手法を比較した結果, 比較的良好な結果が得られることの多い χ^2 値に基づく方法 [10] を用いる. すなわち, 各クラス (株価上昇, 株価下降など) における特徴テーマを求めるために, クラス c_j におけるテーマ t_i の出現頻度の x_{ij} の理論度数 m_{ij} からのずれをテーマの評価点 $score(i,j)$ とし, あるクラスにおける評価点の大きい順に特徴テーマとして選択する.

$$score(i,j) = \chi^2_{ij} = \frac{(x_{ij} - m_{ij})^2}{m_{ij}}$$

$$m_{ij} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij}$$

ここで m はテーマ異なり数を, n はクラス数を, x_{ij} はクラス c_j におけるテーマ t_i の出現頻度を, m_{ij} はクラス c_j におけるテーマ t_i の理論度数を, それぞれ表す.

特徴素選択に基づく手法では, 各クラスを特徴付けるテーマを抽出できるが, クラス間でのテーマの比較を行うことは単純にはできない. 何故なら, 例えば上記の χ^2 値に基づく方法や Kullback-Leibler 情報量に基づく方法では, クラスに属する文書件数が小さいほど特徴量は大きなものが割り当てられる傾向があるからである. 株価変動の大きさを反映するようにクラスが設定されている場合, 各テーマが株価上昇, 株価下降などどのクラスにどの程度寄与しているのかは株価変動分析のための重要な情報となると考えられる.

そこで, 特徴テーマ候補が適切かどうかを確認するために, 最大エントロピー法 [11], [12] を用いて次の手順でテーマ候補 t の出現時におけるクラス a の条件付き確率 $P(a|\{t\})$ の推定を行う.

1. 株価上昇, 株価下降などをクラス a とする
2. 各記事の属するクラス a と, その記事に付与されたテーマの集合 b とから, feature function $f(a,b)$ を設定する
3. 最大エントロピー法により, 各テーマ候補 t について, そのテーマ候補の出現時における各クラス a の確率 $P(a|\{t\})$ を推定する.

各クラス a のテーマ候補 t について, 確率 $P(a|\{t\})$ が他のクラス a' における確率 $P(a'|\{t\})$ よりも大きければ, t はクラス a の特徴テーマとして適切であるとする.

5 実験

5.1 新聞記事の株価変動への影響分析

新聞記事の 富士通 / TOPIX (TOPIX に対する富士通の株価変動) への影響分析として, 単位期間を 1

週間, 1ヶ月とした場合の株価上昇に関する結果を表1, 表2にそれぞれ示す. 今回の実験では, 有意水準を10%とした.

表1: 新聞記事の富士通/TOPIXへの影響分析 (単位期間: 1週間)

テーマ	平均	σ	Z
◇高齢者 (14)	4.00	6.69	2.44
◇特許・商標 (14)	3.79	4.50	2.28
◇新株発行・社債発行 (16)	2.66	4.76	1.51

表2: 新聞記事の富士通/TOPIXへの影響分析 (単位期間: 1ヶ月)

テーマ	平均	σ	Z
◇企業業績 (123)	5.53	9.31	3.25
◇企業財務 (75)	5.92	10.14	2.83
◇合弁・提携 (144)	4.71	10.80	2.43
◇高齢者 (14)	7.29	10.89	1.89
◇倒産・事業縮小 (18)	6.40	7.85	1.67
◇会社概況 (271)	3.74	9.20	1.60
◇大企業 (122)	3.90	9.60	1.30

単位期間1週間の場合の株価上昇と関連の高いテーマとして, <特許・商標>や<新株発行・社債発行>などがある. <特許・商標>には, 海外で重要技術の特許化を行った, あるいは特許関連の裁判に勝訴した, などの記事が含まれる. <新株発行・社債発行>には, ベンチャー企業への出資の記事などが含まれる.

単位期間1ヶ月の場合の株価上昇と関連の高いテーマとしては, <合弁・提携>などがある. これらのテーマ関連のニュースは, より長期的な影響力を持つものと考えられる.

ここでは単位期間を固定して影響力の大きなテーマを求めたが, テーマを固定して単位期間による影響力の大きさを分析することも考えられる.

次に, 電機メーカー4社 (日立, 東芝, NEC, 富士通) について同様に株価上昇と関連の高いテーマを単位期間1週間で求めた所, 共通して出現するテーマがいくつか存在した.

4社中2社以上に共通するテーマとしては, <価格 (物価) >, <特許・商標>, <新株発行・社債発行>, <技術開発>などがあった. これらはいずれも電機メーカーの特徴をよく反映している. <価格 (物価) >は主にパソコンの販売状況に関する記事である. このテーマに属する記事は特定のメーカーに関する記事というよりも複数のメーカーに関連する記事であるため, 共通して影響力を持つテーマとして抽出されたと考えられる.

5.2 株価変動の外部要因分析

富士通/TOPIXの外部要因分析に関し, 単位期間を1週間とした場合の株価上昇に関する実験を行った. 株価変動の標準偏差 σ は 4.7(%) なので, 株価変動の抽出として

【上昇】 10% ～ の株価変動のクラス

【やや上昇】 5% ～ 10% の株価変動のクラス

【平坦】 -5% ～ 5% の株価変動のクラス

【やや下降】 -10% ～ -5% の株価変動のクラス

【下降】 ～ -10% の株価変動のクラス

を抽出した.

特徴素選択に基づく株価上昇時の特徴テーマ候補の抽出結果を表3に示す.

表3: 株価上昇時の特徴テーマ候補

順位	上昇	やや上昇
1	合併・買収 (M&A)(2.75)	高齢者 (5.45)
2	社会 (2.69)	社会問題 (3.20)
3	企業財務 (2.51)	新株発行・社債発行 (2.93)
4	企業動向 (1.47)	企業業績 (1.83)
5	合弁・提携 (1.38)	製品用途 (1.37)
6	ビジネス (1.26)	特許・商標 (0.78)
7	契約・注文 (0.76)	ビジネスソフトウェア (0.46)
8	価格 (物価)(0.76)	株式・債券市場 (0.22)
9	会社概況 (0.74)	合併・買収 (M&A)(0.21)
10	大企業 (0.74)	ライセンス・特許 (0.20)

さきほどの新聞記事の株価変動への影響分析と違い, クラス【上昇】において<合併・買収 (M&A) >, <合弁・提携>といったテーマが上位に出現している. これは, <合併・買収 (M&A) >, <合弁・提携>の記事すべてが株価上昇に関連するわけではないが, それらの記事の中には大きな株価上昇の要因となるものがある, ということを示している.

代わりに, 新聞記事の株価変動への影響分析で出現していた<特許・商標>や<新株発行・社債発行>といったテーマはクラス【やや上昇】の方に出現している. これらのテーマの記事は, 平均的に株価上昇に寄与するだけではなく, 株価上昇に寄与する記事の中でもある程度の割合を占めているテーマであるとみられる.

これらの結果はテーマ毎の記事数の影響もある. <特許・商標>や<新株発行・社債発行>は記事件数が比較的小さく (十数件程度), かつ株価変動と言う面からみて比較的似かよった傾向の記事が集まっていたのに対し, <合併・買収 (M&A) >や<合弁・提携>は記事件数が百件前後とかなり大きく, 株価上昇と関連の高い記事もそうではない記事も含まれていたため, このような結果が得られたと推測される.

次に, クラス【上昇】の特徴テーマ候補<合併・買収 (M&A) >, <合弁・提携>, および【やや上昇】の特徴テーマ候補<特許・商標>, <新株発行・社債発行>について, 最大エントロピー法に基づく確率 $P(a|\{t\})$ の推定結果を表 4 に示す.

表 4: 確率 $P(a|\{t\})$ の推定結果

テーマ	上昇	やや上昇	平坦	やや下降	下降
合併・買収 (M&A)	0.27	0.24	0.19	0.14	0.15
合弁・提携	0.29	0.21	0.21	0.14	0.16
特許・商標	0.40	0.53	0.02	0.02	0.03
新株発行・社債発行	0.24	0.40	0.05	0.15	0.16

これをみると, <合併・買収 (M&A) >および<合弁・提携>はクラス【上昇】の確率が最も大きく, 【上昇】の特徴テーマといってよいと考えられる. 同様にして, <特許・商標>および<新株発行・社債発行>は【やや上昇】の特徴テーマであると確認できる.

6 まとめ

新聞記事に付与されたテーマ情報を用いた新聞記事の株価変動への影響分析および株価変動の外部要因分析により, 株価データと新聞記事の関連性を明らかにした. また実際に, いくつかの銘柄に関し株価変動と関連のあるテーマを抽出した.

今後は, テーマ情報に限らず記事の本文情報や書誌的情報 (何面の記事か, 記事のサイズなど) も利用して, より質の高い株価データと新聞記事の関連性抽出を行うことが課題のひとつである.

現在扱えていない問題としては, テーマの意味的な粒度が十分細かくはないということが挙げられる. 例えば<合弁・提携>というテーマには提携した記事と, 提携を解消した記事とが混在している可能性がある. 提携と提携解消では, 株価変動という面からは大きく異なるはずである.

ネットワークから利用可能な株価データやニュース記事などの時系列データを用いることで, よりリアルタイムで時間精度の高い実験を行うことが可能になると思われる. そのような実験にも取り組みたいと考えている.

謝辞

日刊工業新聞データの使用に関して, 記事データの研究利用許諾をいただいた日刊工業新聞社およびジー・サーチ社に感謝いたします.

最大エントロピー法の計算にあたり, 富士通研究所の颯々野学研究員の開発されたプログラムを利用させていただきました. ここに感謝いたします.

参考文献

[1] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J.: Mining of Concurrent Text and Time Series, *KDD-2000 Workshop on Text Mining* (2000).

[2] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J.: Language Models for Financial News Recommendation, *Proceedings of the Ninth International Conference on Information and Knowledge Management* (2000).

[3] Fawcett, T. and Provost, F.: Activity Monitoring: Noticing interesting changes in behavior, *Proceedings of the 5th International Conference on KDD* (1999).

[4] 渡部勇, 三末和男：単語の連想関係によるテキストマイニング, 情処研報 情報学基礎 55-8, pp. 57–64 (1999).

[5] 三末和男, 渡部勇：テキストマイニングのための連想関係の可視化技術, 情処研報情報学基礎 55-8, pp. 65–72 (1999).

[6] 内野寛治, 宗意幸子, 橋本三奈子, 武智 峰樹, 松井くにお, 菊田泰代：ルールベースを用いたテキスト分類サービス －自動分類技術のビジネスへの応用－, INFOSTA シンポジュウム 2000 (2000).

[7] 東京大学教養学部統計学教室 (編)：統計学入門, 東京大学出版会 (1991).

[8] Yang, Y. and Pedersen, J. O.: A Comparative Study on Feature Selection in Text Categorization, *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)* (1997).

[9] Mladenic, D. and Grobelnik, M.: Feature Selection for classification based on text hierarchy, *Conference on Automated Learning and Discovery (CONALD-98)* (1998).

[10] 渡辺靖彦, 竹内雅人, 村田真樹, 長尾真：χ^2 法を用いた重要漢字の自動抽出と文献の自動分類, 信学技法 NLC94-25, pp. 23–30 (1994).

[11] 北研二：確率的言語モデル, 東京大学出版会 (1999).

[12] Manning, C. D. and Schutze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press (1999).