テキスト情報と市場時系列データの統合分析

An Integrative Analysis of Texual Data and Time-Series Market Data

和泉 潔[1*]　後藤 卓[2]　松井 藤五郎[3]
Kiyoshi IZUMI[1]　Takashi GOTO[2]　Tohgoroh MATSUI[3]

[1] 産業技術総合研究所 デジタルヒューマン研究センター
[1] DHRC, National Institute of Advanced Industrial Science and Technology
[2] 三菱東京 UFJ 銀行
[2] The Bank of Tokyo-Mitsubishi UFJ, Ltd.
[3] 東京理科大学 理工学部
[3] Faculty of Science and Technology, Tokyo University of Science

Abstract: In this study, we proposed a new text-mining methods for long-term market analysis. Using our method, we analysed monthly price data of Japanese government bond market. First we extracted feature vectors from monthly reports of Bank of Japan. Then, trends of the JGB market were estimated by regression analysis using the feature vectors. As a result, determination coefficients were over 75%, and market trends were explained well by the information that was extracted from textual data. Finally, we compared the predictive power of textual data with that of numerical data. As a result, Our text mining method had prediction power superior to the numerical data analysis.

1　はじめに

金融市場では常に様々な情報が溢れている．トレーダー達は，市場に影響を及ぼす多様な情報を取捨選択し，現在の市場の状況を分析・予測している．市場の分析に用いる情報には大きく分けて 2 種類がある．一つは，経済指標，マーケットのテクニカル指標等の数値情報である．もう一つは，市場に関わる要人の発言，中央銀行や他の市場参加者の解析記事などのテキスト情報である．これらの多様な情報が瞬時にトレーダー達のもとに，オンラインで送られてきているのである．送られてきた情報の全てを，現場のトレーダーが自分で目を通して市場分析に用いることは不可能に近い．そのため，いくつかの情報技術を市場分析に適用する研究が行われてきた．例えば，数値情報を用いて現在の市場情報を推論するようなエキスパートシステムの構築を行う研究 [7] やニューラルネットや遺伝的アルゴリズムを数値情報による市場分析に用いた研究もある [6]．これらの研究は一定の成果をあげてきた．しかし，数値情報には指標化されていない情報がもともと含まれていないので，分析対象の範囲がテキスト情報よりも狭くなる可能性がある．しかも，指標を集計して発表するには，どうしてもタイムラグが生じてしまうので，分析への反映も遅れがちである．近年，テキスト情報による市場分析に関して，ロイターなどのオンラインの経済ニュースに対する市場の反応を推測する研究もでてきた [1,3,4]．これらの研究は，1 日以内や数日の短期的な市場の反応を分析対象としており，より長期的な市場動向の分析には用いられてこなかった．そこで，我々はオンラインのテキスト情報から，数年にわたる比較的長期の市場動向の変化を分析するための補助を目的とした解析技術を新たに開発した．こういった観点から，市場参加者が特に注目する日本銀行の金融経済月報を題材に，テキストマイニング技術を用いて経済市場分析を試み，また金融経済月報が実際の金利動向をどの程度説明しているのかについて検証を行った．

*連絡先：産業技術総合研究所 デジタルヒューマン研究センター
〒 135–0064 東京都江東区青海 2–41–6
E-mail: kiyoshi@ni.mints.ne.jp

2　テキストデータによる長期市場分析手法

テキストマイニングを長期的な市場分析に用いるには，2 つの重要な点がある．適切な内容と形式をもつテキストデータの選択と，テキストデータと時系列デー

タを関連づける手法である．

最初に，本研究では日本銀行の金融経済月報をテキストデータとして選んだ．金融経済月報は，日本銀行が金融・経済情勢を分析した資料であり，毎月半ばに，A4 で 15-20 ページの分量で公開されている[1]．金融政策の方針を決める金融政策決定会合で内容を審議し，政策決定の基礎資料とする．この情報によって，日本銀行が，当面の経済動向をどう分析しているか対外的に明らかにしている．今回，金融経済月報を分析対象にした理由は 3 つある．第一に金融経済月報は，実際の金融市場のトレーダーが多かれ少なかれ着目している共有の重要テキスト情報であるからである．第二の理由は，会員制の有料マーケットリポート等のテキスト情報と違って，毎月の中旬にサイト上で定期的に発表されていて，誰でもアクセス可能な情報であることである．三番目の理由は，ブログ等のほとんど決まった形式のないテキスト情報と異なり，解説内容の順番や段落構成等がほぼ定式化されていて，月ごとのテキスト内容の変化が比較しやすいからである．

二番目のポイントとして，本研究ではテキストデータと時系列データを関連づけるために，図 1 にある下記の 3 つのステップからなる新たなテキスト解析技術を提案する．

1. 共起関係に基づく主要単語の抽出と可視化
2. 主成分分析による単語のグループ化
3. 重回帰分析による金利データの動向分析

2.1 共起関係に基づく主要単語の抽出と可視化

最初に，各月のテキストデータに KeyGraph [5] を適用し，共起関係を解析した．具体的にはまず，日本語形態素解析システムである Chasen [2] による形態素解析を行い，出現頻度順に名詞・動詞・形容詞等を抽出した．次に，Jaccard 係数（= p(A and B)/p(A or B); ただし A,B は抽出した単語）を段落毎に適用し，段落毎に同時に出現する単語と単語を繋ぎ，共起グラフを作成する．その後，単結合（A,B 間のみの結合部分）を切断し，結合による「島」を作成する．またその後，各単語間の共起度に基づき，上位順に「橋」を作成する．これらによって，各月のテキストデータから主要単語をノードとするネットワークを構築した．

[1] テキストデータは http://www.boj.or.jp/theme/seisaku/handan/gp/ で毎月公開されている．

2.2 主成分分析による単語のグループ化

KeyGraph で作成したネットワークに出現した単語のパターン (単語を月毎の出現状況に従いパターン分類したもの) に対し主成分分析を実施し，30 個の合成変数（主成分）にまとめる．ここで，主成分の数が 30 個であったのは，1998 年から 2007 年までのデータを用いた主成分分析で，累積寄与率が 60%を超えた主成分数が 30 であったからである．各月の 30 個の主成分スコアを，分析対象期間について時系列順に並べることによって，30 次元の時系列データが作成される．これが分析対象期間のテキストデータの特徴の時間的変化を表していると考える．主成分分析の際には，単語に関して品詞を区別せずに分析を実施する．ここで注意してほしいのは，ここまで金利データは全く用いず，純粋に単語の出現パターンのみの分析を行っていることである．つまり，ここまでの分析は，債券市場や株式市場，外国為替市場などの分析対象となる市場の種類に依存せずに，共通であるということである．

2.3 重回帰分析による金利データの動向分析

最後に，各主成分スコアの毎月の動きから月次での金利の動きを解析する．具体的には，さきほどの 30 個の主成分スコアの時系列データを説明変数として，月次の市場データを被説明変数とする重回帰分析を行う．分析対象期間内の金利の動きを推定するだけでなく，分析対象外のテキストデータを与えれば外挿予測を行うこともできる．この外挿予測は，月中に発表される金融経済月報から，約 2 週間後の月末の金利を推定することになる．

3 金融経済月報のテキストマイニング

上述の手法を用いて，日本国債市場の価格 (金利) の月次変動を分析した．1998 年 1 月から 2007 年 12 月までの 10 年間 (120ヶ月) の金融経済月報のテキストと金利データ (月末終値) をサンプルデータとした．

3.1 金融経済月報による月次市場分析

最初に，KeyGraph アルゴリズムと主成分分析を用いて，30 次元の特徴量を金融経済月報のテキストデータから抽出した (表 1)．抽出された主成分には大きく分けて 2 つのタイプがあった．一つは市場の動きに関する特徴量である．例えば，1 番目の主成分は,「横ばい」「圏内」「緩やか」といった動きを表す単語から構

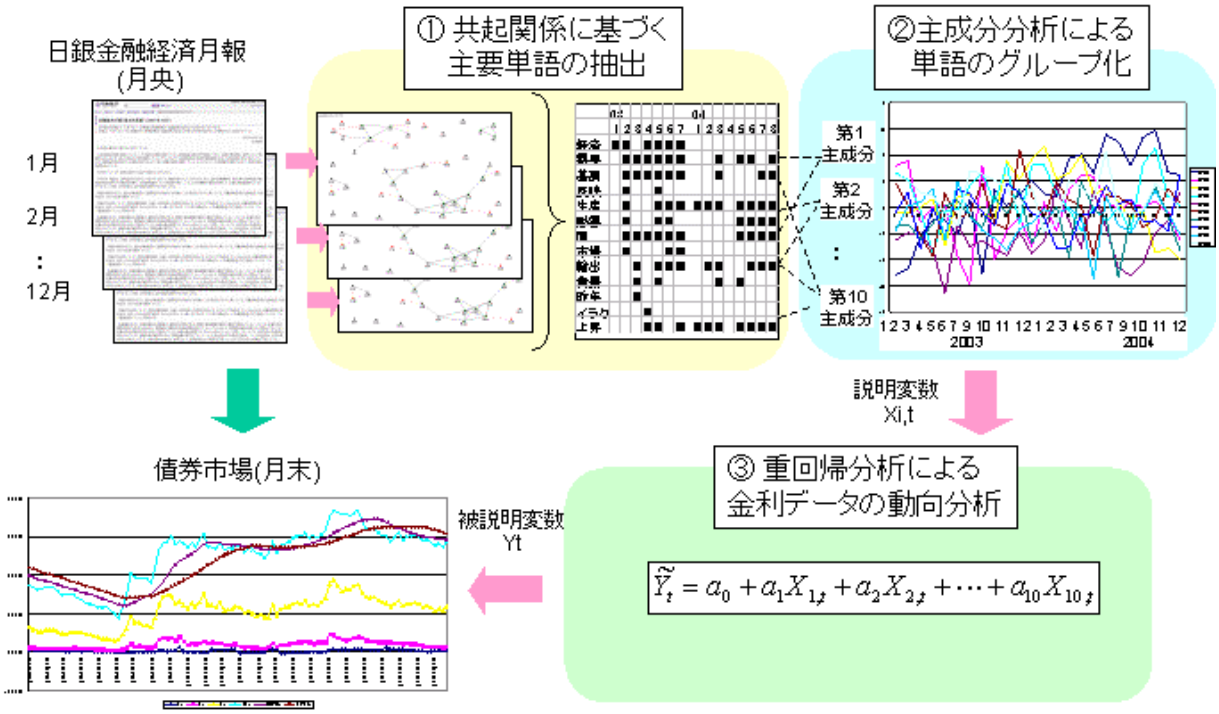

図 1: テキストデータによる時系列データの推定手法

成されていた．他にも，5 番目の主成分は,「上昇」「頭打ち」「軟化」といった単語の寄与が高かった．もう一つのタイプは，経済のファンダメンタルズに関する特徴量である．例えば，2 番目の主成分は「リスク」「国債」「利回り」といった金利に関する単語から構成されていた．他にも，3 番目の主成分は,「需要」「改善」「生産」といった企業活動に関する単語の寄与が高かった．

次に，これらの 30 次元の特徴量の時系列データを用いて，日本国債市場データの回帰分析を行った．回帰結果を表 2 に示す．回帰分析の際に，AIC 基準を用いたステップワイズ選択により，説明変数の絞り込みを行った．日本国債の 1 年物，2 年物，5 年物，10 年物の金利について，23-25 個の説明変数による回帰式を得ることができた．決定係数 R^2 をみると，サンプルデータについて十分な説明力を持つことがわかった．R^2=75.24% (日本国債 1 年物)，78.47% (日本国債 2 年物)，76.76% (日本国債 5 年物)，and 74.65% (日本国債 10 年物)．それから，2008 年 1 月から 6 月までのテキストデータを用いて，外挿予測テストを行った．図 2a-d に，日本国債 1,2,5,10 年物のそれぞれについて，推定されたパスと実際のパスを示す．外挿期間において，推定パスは実際のパスと同様に，一度下がってから上がり，再度下がるという動きを示した．

3.2 数値データによる市場分析との比較

テキストデータを長期市場分析に用いることの有効性をテストするために，数値データによる市場分析と予測力の比較を行った．そのため，通常よく分析に用いられる 19 種類の月次の数値データ[2]を説明変数として，2000 年 1 月から 2007 年 12 月までの日本国債市場データの回帰分析を行った．回帰結果を表 3 に示す．テキストデータの分析と同様に，AIC 基準によるステップワ

[2]マネタリーベース (前年比); 景気先行指数; 貸出動向 (前年比); マネーサプライ M2+CD(前年比); 機械受注 (前月比); 経常収支 (季調済); 工作機械受注 (前年比); 企業倒産件数 (前年比); 第三次産業活動指数 (前月比); 企業向けサービス価格指数 (前年比); 通関ベース貿易収支 (季調済); 鉱工業生産 (前月比); 失業率; 有効求人倍率; 東京消費者物価指数 (除生鮮/前年比); 全国消費者物価指数 (除生鮮/前年比); 住宅着工戸数 (前年比); 消費者態度指数; 消費支出 (前年比)

主成分 1		主成分 2		主成分 3		主成分 4		主成分 5		主成分 6	
横ばい	-0.845	リスク	-0.631	背景	0.655	設備	0.468	足許	-0.458	量的	-0.639
圏内	-0.75	軟調	-0.537	伴う	0.494	国内	-0.432	上昇	-0.436	停滞	-0.639
環境	0.718	国債	-0.537	需要	0.452	低迷	0.421	実体	-0.401	持続	-0.639
資金	0.706	利回り	-0.537	改善	-0.424	輸出	-0.411	年末	-0.394	強い	-0.639
伸び	-0.705	格差	-0.537	生産	-0.421	歯止め	0.36	頭打ち	-0.394	実施	-0.5
基調	-0.702	根強い	-0.537	鈍化	-0.404	掛かる	0.36	先行き	-0.381	歯止め	0.462
緩やか	-0.697	投資	0.532	軟調	-0.394	総合	0.36	厳しい	0.374	掛かる	0.462
民間	0.656	窺う	-0.531	国債	-0.394	対策	0.36	間	-0.363	総合	0.462
金融	0.651	横這い	0.497	利回り	-0.394	ベース	0.358	軟化	-0.355	対策	0.462
低下	0.639	拡大	-0.492	格差	-0.394	踏まえる	0.354	ベース	0.352	窺う	0.415

主成分 7		主成分 8		主成分 9		主成分 10		主成分 11		主成分 12	
調整	-0.632	歯止め	-0.643	マクロ	-0.553	システム	-0.463	年末	0.329	同時	0.625
雇用	0.477	掛かる	-0.643	ギャップ	-0.553	銀行	-0.388	頭打ち	0.329	テロ	0.625
関連	-0.452	総合	-0.643	超過	-0.553	不安	-0.381	受ける	0.322	事件	0.625
厳しい	0.378	対策	-0.643	市況	-0.473	済	-0.381	間	0.312	社債	0.47
銀行	0.369	中小	-0.478	国際	-0.473	傾向	-0.366	軟調	-0.303	機械	0.456
量的	0.365	見込む	-0.454	プラス	-0.44	個人	-0.356	国債	-0.303	米国	0.442
停滞	0.365	収益	-0.369	商品	-0.389	幅	-0.336	利回り	-0.303	システム	0.357
持続	0.365	ベース	0.36	均す	-0.349	大幅	-0.325	格差	-0.303	財	0.35
強い	0.365	指標	-0.354	考える	0.349	伴う	0.32	根強い	-0.303	発行	0.322
維持	0.359	窺う	-0.302	海外	-0.316	経済	0.32	足許	0.291	幅	-0.315

主成分 13		主成分 14		主成分 15		主成分 16		主成分 17		主成分 18	
作用	0.488	雇用	0.46	もと	0.379	不透明	-0.434	減少	-0.354	アジア	-0.322
進行	0.488	縮小	0.367	効果	0.346	生産	0.389	反動	-0.331	米	0.29
昨秋	0.46	受ける	-0.346	同時	0.344	金利	0.375	金利	0.329	前年	-0.278
公共	0.432	イラク	0.343	テロ	0.344	調達	0.358	わが国	0.282	効果	0.274
ベース	-0.377	情勢	0.343	事件	0.344	イラク	-0.355	弱まる	0.282	伴う	0.273
不安	-0.374	必要	-0.312	結果	0.334	情勢	-0.355	相場	0.276	年末	0.267
済	-0.374	不透明	0.299	支出	0.333	低調	-0.351	部品	-0.269	頭打ち	0.267
結果	-0.343	賃金	0.284	アジア	0.281	銀行	-0.308	たどる	-0.265	その後	0.261
季節	-0.319	アジア	-0.275	財	0.275	長期	0.289	強まる	0.254	不安	-0.255
及ぼす	-0.314	為替	-0.271	不安	-0.271	一部	0.283	資本	-0.251	済	-0.255

主成分 19		主成分 20		主成分 21		主成分 22		主成分 23		主成分 24	
着実	-0.361	乏しい	0.501	賃金	0.336	製品	-0.335	調査	0.381	一部	0.372
高め	-0.355	流通	0.501	消費	0.319	年末	0.321	本年	0.381	受ける	0.365
反動	-0.307	需給	-0.333	一部	0.301	頭打ち	0.321	不透明	-0.352	圧力	-0.335
昨年	-0.302	減少	-0.328	発行	-0.284	状況	-0.308	乏しい	-0.341	既往	0.32
マクロ	0.302	自動車	0.326	不透明	-0.282	必要	-0.306	流通	-0.341	弱い	0.256
ギャップ	0.302	明確	-0.326	需要	-0.27	減少	-0.281	減少	0.34	緩和	0.254
超過	0.302	維持	0.307	既往	0.263	その後	0.279	イラク	-0.286	需給	0.239
雇用	0.265	弱い	0.3	サービス	0.261	マクロ	-0.273	情勢	-0.286	不透明	0.239
調査	0.251	好影響	-0.294	持ち直し	0.257	ギャップ	-0.273	高水準	0.265	最終	-0.232
本年	0.251	東アジア	0.292	イラク	-0.247	超過	-0.273	圧力	-0.252	イラク	0.228

主成分 25		主成分 26		主成分 27		主成分 28		主成分 29		主成分 30	
後退	0.326	米価	-0.383	押し上げ	0.443	米価	-0.446	ドル	0.48	住宅	-0.318
調査	0.306	一時	-0.383	働く	0.443	一時	-0.446	相場	0.446	米価	0.285
本年	0.306	調査	-0.376	個人	-0.39	発行	0.324	方向	0.427	一時	0.285
意識	0.295	本年	-0.376	需要	0.265	強まる	0.302	イラク	0.257	既往	-0.27
発行	0.291	圧力	-0.343	着実	-0.259	意識	-0.29	情勢	0.257	一部	-0.259
米	0.288	高水準	-0.303	輸出	-0.254	後退	0.255	基調	0.255	為替	0.25
サービス	0.285	最終	-0.295	収益	-0.239	当面	0.237	米	0.25	伸び	-0.245
緩和	-0.261	作用	0.269	要因	0.229	電気	-0.222	発行	-0.218	後退	-0.245
既往	0.256	進行	0.269	相場	0.222	アジア	0.217	テンポ	0.207	変化	0.238
もと	-0.255	家電	0.268	大幅	0.221	機械	-0.216	本格	-0.206	要因	0.232

表 1: 1997 年 1 月から 2007 年 12 月までのテキストから抽出された主成分と，各主成分で負荷量の絶対値が上位 10 個のキーワード.

	従属変数			
	JGB1Y	JGB2Y	JGB5Y	JGB10Y
主成分 1	−0.0135***	−0.0188***	−0.0109**	−0.0083*
主成分 2	−0.0361***	−0.0396***	−0.0360***	−0.0147**
主成分 3	0.0411***	0.0559***	0.0675***	0.0435***
主成分 4	–	–	0.0217***	0.0176**
主成分 5	−0.0117**	−0.0238***	−0.0561***	−0.0548***
主成分 6	0.0215***	0.0256***	0.0227***	–
主成分 7	0.0089·	0.0088·	–	−0.0289***
主成分 8	–	–	–	−0.0214**
主成分 9	0.0000	−0.0079	−0.0209**	−0.0173*
主成分 10	0.0247***	0.0259***	0.0247**	–
主成分 11	−0.0146**	−0.0141*	−0.0196*	−0.0198**
主成分 12	−0.0092·	−0.0081	−0.0107	–
主成分 13	0.0071	0.0088	0.0164*	0.0291***
主成分 14	−0.0122*	−0.0097	–	–
主成分 15	−0.0121*	−0.0099	–	−0.0163*
主成分 16	–	–	0.0174*	0.0374***
主成分 17	–	–	–	–
主成分 18	–	–	−0.0262**	−0.0359***
主成分 19	–	–	–	–
主成分 20	–	–	–	–
主成分 21	–	–	–	–
主成分 22	−0.0132*	−0.0102	–	–
主成分 23	–	–	–	–
主成分 24	−0.0091	–	–	−0.0131
主成分 25	–	–	–	–
主成分 26	–	–	–	–
主成分 27	−0.0095	−0.0169*	−0.0329***	−0.0333***
主成分 28	–	–	–	−0.0137
主成分 29	–	–	–	–
主成分 30	–	–	–	–
定数	0.2012***	0.3358***	0.8278***	1.4909***
N	120	120	120	120
F 値	21.07	25.27	24.77	18.96
R^2	0.7524	0.7847	0.7676	0.7465

***: 0.1%有意, **: 1%有意, *: 5%有意, .:10%有意.
–: ステップワイズ選択で選択されなかったことを示す.

表 2: テキストデータによる国債市場の回帰式の係数

	従属変数			
	JGB1Y	JGB2Y	JGB5Y	JGB10Y
マネタリーベース (前年比)	−0.0084***	−0.0073**	−0.0083***	–
景気先行指数	0.0013*	0.0020**	0.0027**	0.0019*
貸出動向 (前年比)	0.0236*	0.0542***	–	−0.0262
マネーサプライ M2+CD(前年比)	–	–	–	–
機械受注 (前月比)	–	–	–	–
経常収支 (季調済)	–	–	–	0.0001
工作機械受注 (前年比)	–	–	−0.0021*	−0.0023*
企業倒産件数 (前年比)	0.0040***	0.0060***	0.0079***	0.0041**
第三次産業活動指数 (前月比)	–	–	–	–
企業向けサービス価格指数 (前年比)	0.2046***	0.1875***	0.1964***	–
通関ベース貿易収支 (季調済)	–	0.0001	0.0002*	0.0001
鉱工業生産 (前月比)	0.0097	0.0114	0.0179·	–
失業率	–	–	–	−0.1509
有効求人倍率	−0.0020·	−0.0020	−0.0031·	−0.0069***
東京 CPI(除生鮮/前年比)	–	0.0648	–	1.5320·
全国 CPI(除生鮮/前年比)	–	–	0.2570***	0.1253
住宅着工戸数 (前年比)	–	0.0025	0.0033	–
消費者態度指数	−0.0222***	−0.0215***	–	0.0175
消費支出 (年比)	–	–	–	–
定数	1.2663***	1.3420***	0.7056***	1.1010*
N	96	96	96	96
F 値	59.09	55.11	38.84	18.83
R^2	0.8461	0.8796	0.8222	0.7139

***: 0.1%有意, **: 1%有意, *: 5%有意, .:10%有意.
–: ステップワイズ選択で選択されなかったことを示す.

表 3: 数値データによる国債市場の回帰式の係数

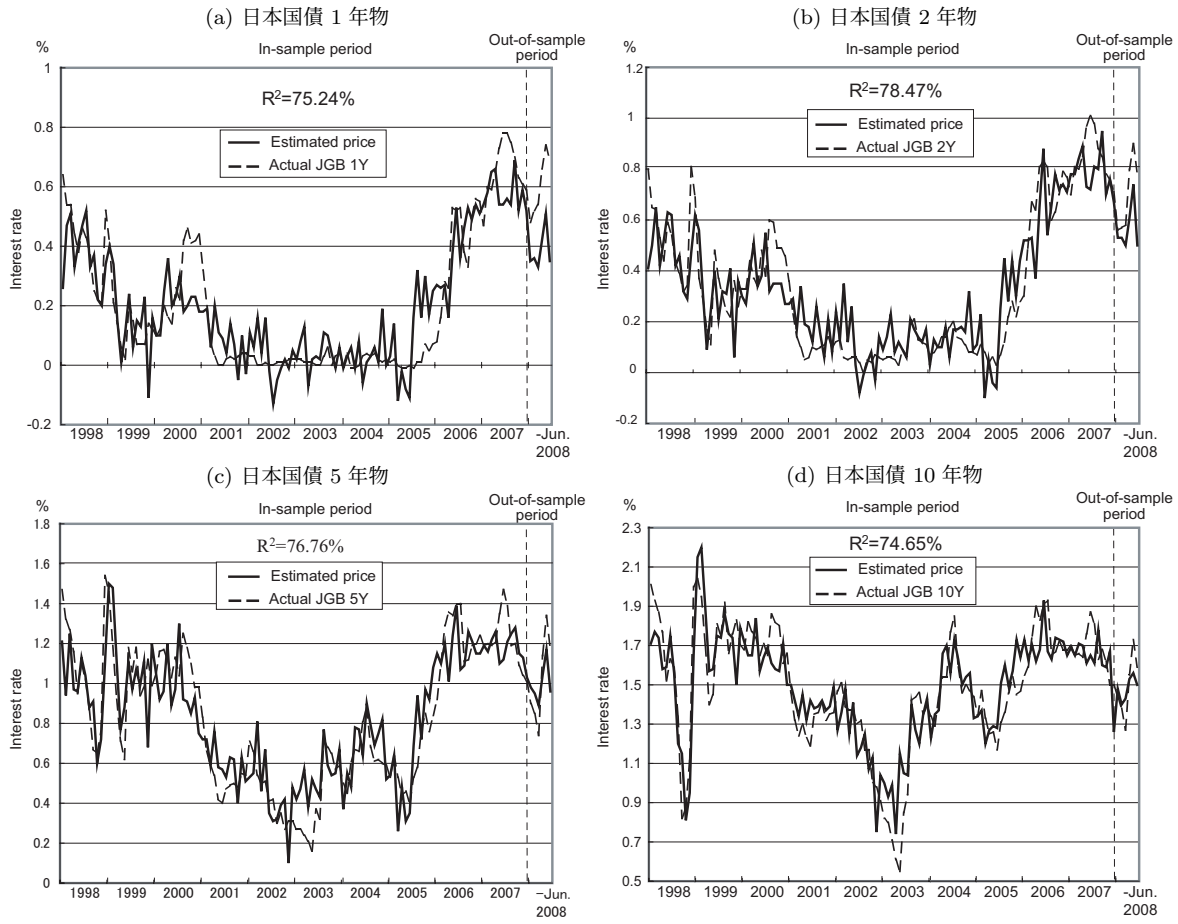

図 2: 日本国債 1,2,5,10 年物の市場トレンドの推定

イズ選択を行った結果，8-11 個の説明変数が選ばれた．そして，得られた回帰式はサンプルデータを非常に良く説明できるものであった．決定係数 R^2 は 84.61% (日本国債 1 年物)，87.96% (日本国債 2 年物)，82.22% (日本国債 5 年物)，71.39% (日本国債 10 年物) であった．

しかしながら，2008 年 1 月から 4 月までの数値データを用いて外挿テストを行った結果，数値データによる市場分析はテキストデータによる分析より明らかに予測力が劣っていた．図 3 が示すように，同じ期間の外挿予測の誤差を比較すると，日本国債 2,5,10 年物に関して，数値データの予測誤差はテキストデータの予測誤差の約 3 倍程度にもなった．この結果より，数値データはサンプルデータの説明力は高いが，サンプルに過剰学習を行って，市場の背景にある構造を適切に推定することができなかったと思われる．これに対して，適切なテキストデータを分析対象とすれば，テキストに含まれる豊かな情報を用いて，市場構造をより適切に分析できる可能性があることがわかった．

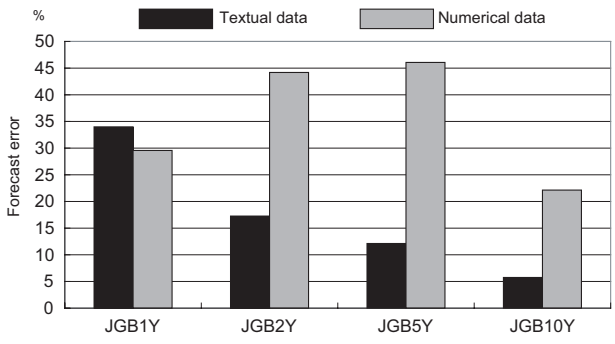

図 3: 外挿予測誤差の比較．Y 軸は外挿期間の実際の価格の平均を 100 とした場合の，予測誤差の割合を示す

4　まとめ

本研究では，テキストデータを用いた長期的な市場分析の新たな手法を提案した．本手法により，月次の日本国債市場データの分析を行った結果，サンプルデータの説明と外挿テストの両方において，優れた結果を示した．さらに，数値データによる市場分析と予測力の比較を行ったところ，本手法の方が既存の数値データによる長期市場分析よりも，高い予測力を得ることができた．これにより，今まで日次以下の短期の市場分析にしかテキスト分析が用いられてこなかったが，週次や月次といった長期市場分析にもテキスト分析が有効である可能性を示した．

本研究では，分析に好条件であると思われるテキスト情報を用いたが，今後は本手法をマーケットリポートやブログ等のより条件の厳しい情報に適用を試みる予定である．またテキストマイニングに市場分析と，市場シミュレーションを統合することによって，市場参加者の行動によるフィードバックを考慮したより動的な市場分析を行うことを目指す．

参考文献

[1] K. Ahmad, L. Gillam, and D. Cheng. Textual and quantitative analysis: Towards a new, e-mediated social science. In *Proc. of the 1st International Conference on e-Social Science*, 2005.

[2] ChaSen ホームページ. `http://chasen.naist.jp/hiki/chasen/`.

[3] Young-Woo Seo, Joseph Andrew Giampapa, and Katia Sycara. Financial news analysis for intelligent portfolio management. Technical Report CMU-RI-TR-04-04, Carnegie Mellon University, 2004.

[4] 高橋悟, 高橋大志, 津田和彦. 株式市場におけるヘッドラインニュースの効果についての研究. ファイナンス学会第 15 回大会, pp. 373–383, 2007.

[5] 大澤幸生. チャンス発見のデータ分析―モデル化+可視化+コミュニケーション→シナリオ創発. 東京電機大学出版局, 2006.

[6] 電気学会（編）. 学習とそのアルゴリズム, 第 6 章. 森北出版, 2002.

[7] 日本ファジィ学会（編）. ファジィ・エキスパート・システム. 日刊工業新聞社, 1993.