英文経済レポートのテキストマイニング分析ツールの開発

Development of the tool to analyze Financial Markets' Fluctuation by English Textual Information

余野 京登[1*]　和泉 潔[1,2]　後藤 卓[3]　松井藤五郎[4]　陳 ユ[1]
Kyoto Yono[1]　Kiyoshi Izumi[1,2]　Takashi Goto[3]　Tohgoroh Matsui[4]　Chen Yu[1]

[1] 東京大学　[2]JST さきがけ　[3] 三菱東京 UFJ 銀行　[4] 中部大学
[1] The University of Tokyo　[2] PRESTO，JST　[3] The Bank of Tokyo-Mitsubishi UFJ，Ltd.
[4] Chubu University

Abstract: 本研究では，経済レポートのテキストマイニング分析ツールを開発した. 英国中央銀行の金融政策会議議事録を利用し，10 年物の英国国債金利の長期予測を行った結果を報告する. 5 年間のテキストと金利のデータセットから回帰式を獲得し，その後の 1 年間のデータを用いて外挿評価した.

1　はじめに

金融市場を取り巻く情報は日々増加している. 物価指数や産業活動指数など数値データもあれば，企業の有価証券報告書，ロイターなどの配信している経済ニュースなどのテキストデータも含まれている. この膨大な量の情報から，投資家やトレーダーは，抽出した必要な情報を基に経済予測し，投資判断を下している.

近年，データマイニング技術を用いて，市場動向を分析する研究が多く行われている. ニューラルネットワークや遺伝的アルゴリズム等を数値データに用いて市場分析を行うものがあり，一定の成果を上げている [1].

本研究では，テキストマイニングを利用した市場分析のツールを開発した. 中央銀行の発行するレポートを対象にテキストマイニングし，金利予測を行い，ツールの有効性を検証した.

2　金利予測

従来の金利予測の多くは，経済指標などの数値データを用いた単一回帰型のモデルや自己回帰モデルなどがある. [2] では，長期金利（国債 10 年物）を 5 変数の回帰式で計測している.

$$\text{長期金利} = 0.60 * x_1 + 0.05 * x_2 + 0.005 * x_3 + 0.36 * x_4 + 0.19 * x_5$$

ここで，x_1 は予測インフレ率，x_2 は生産増加率，x_3 は円相場の変化率，x_4 は政策金利水準，x_5 は米国実質金利である.

一方で，経済指標などの市場における情報だけでなく，中央銀行が市場金利をどのような方向に誘導するかという情報も金利動向を予測する際のポイントである. そのため，各市場関係者は中央銀行の公表資料や幹部の発言内容を注意深く見守っている. 市場金利の政策誘導方針がどのようなものかを知ることができれば，市場金利の先行きを予測することができる. 本研究では，従来の数値データを基にする金利予測ではなく，市場金利への影響力の大きい中央銀行の発行する公表資料を基にテキストマイニングを利用して金利予測を行う.

3　テキストマイニングによる予測

これまでテキストマイニングの手法による市場予測が多く行われてきた. [3] では，ニュースの見出しを分析し，為替市場の短期予測を行った. 予め決めた単語が見出し内に含まれているかを検出し，TFIDF 法などにより重み付けを行う. 2 時間分の見出しの単語の出現パターンからその後 1 時間における為替レートの上昇，下落，不変を予測する.

[4] では，企業の年次報告書をテキストマイニングし，将来の株価の変動率を予測する実験を行った. 年次報告書のうち市場リスクに関する定性的，定量的記述のある章のみを抽出し，単語に重み付けをした上，サポートベクター回帰により，報告書が発行されてから 1 年後の株価の変動率を予測している.

*連絡先： 東京大学工学系研究科システム創成学専攻大橋研究室
〒 113–8656 東京都文京区 本郷 7–3–1
E-mail:yono@crimson.q.t.u-tokyo.ac.jp

[5] では，Twitter のつぶやきからダウ平均株価を予測した．980 万件のつぶやきを対象に感情を示したつぶやきのみを抽出し，それを 6 つの感情パラメータに変換した．そのうち,「平穏」のパラメータがダウ・ジョーンズ工業株価平均と高い相関を持った．平穏パラメータと自己組織型ファジィニューラルネットワークを用いて，外挿テストを行った結果，86.7 %という高い精度で株価の動きを予測することができた．

4 モデル

本研究のモデルは，和泉らの研究 [6] を応用したものである．その研究では，日本銀行の金融経済月報を題材に，長期的な市場動向予測を行った．
共起グラフを用いた単語抽出と主成分分析，回帰分析から 3 つのステップから成りたっている．本研究では，共起ブラフを用いた単語抽出の代わりに，TF，TFIDF，LOG1P の 3 種類の重み付けを利用した単語抽出をモデルに取り入れた．

4.1 単語抽出

まず各文章に形態素分析を施し，名詞, 動詞, 形容詞以外の単語を除去し, 単語を原形に変換した．名詞＋名詞や，形容詞＋名詞のなど順番で出てきた単語を 1 語の複合名詞に連結した．この機能を加えることで，2 語に分けられていた単語が 1 語として捉えることができ，より正確な単語抽出が可能となる．その後，以下の 3 つの方法で単語の重み付けを行った．

TF

一つのテキストにおける単語の出現頻度をテキストに含まれる全単語数で割ったもの．

$$tf_i = \frac{freq(x_i; d)}{|d|} \tag{1}$$

ここで， $freq(x_i; d)$ は単語 i の出現頻度を表す．

TFITF

TF（単語の出現頻度）と IDF（逆出現頻度）の 2 つの指標を掛け合わせたもの．文章中の重要単語を抽出する際に よく用いられる方法である．

$$tfidf = tf \times idf \tag{2}$$

$$tdf_i = log \frac{N}{|\{d : freq(x_i; d) > 0\}|} \tag{3}$$

ここで，$|\{d : freq(x_i; d) > 0\}|$ は単語 i を含むドキュメント数，N は総ドキュメント数である．

LOG1P

TF（単語の出現頻度）を対数で正規化したものである．

$$log1p_i = log(1 + freq(x_i; d)) \tag{4}$$

最後に，各ドキュメントごとに重み付けのスコア順に単語を抽出した．

4.2 主成分分析による単語のグループ化

各ドキュメントから抽出した単語のスコアを一つに統合し (表 1)，それ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．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．ここで注意してほしいのは，ここまで市場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．

主成分分析を施す前に，抽出単語のうち，一つのドキュメントのみに出現した単語を削除した．これは，主成分分析時に一つのドキュメントのみから抽出した単語がそのドキュメントのみを表すため，回帰分析の外挿予測に利用しにくいと考えたからである．

4.3 重回帰分析による市場データの動向分析

最後に，各主成分スコアの毎月の動きから月次での市場金利の動きを解析する．具体的には， 前節で述べた 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．

$$\tilde{y} = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \tag{5}$$

ここで，$\tilde{y}$ は，予測した金利，x_i は各主成分スコアを表す．

5 ツールの開発

下記のような GUI を wxpython を使い，日本語の形態素解析は mecab を, 英語は python の nltk のパッケージを利用した．また，主成分分析と重回帰分析は python と R を連携されて作成した．

単語	文章 A	文章 B	文章 C	文章 D	···
best collect project	0.00498	0	0	0	···
price-cost margin	0.00456	0	0	0	···
intervent	0.00383	0	0.00243	0.00163	···
novemb project	0.00268	0	0	0	···
januari	0.00233	0.00573	0.00237	0	···
non-eu good deficit	0.00206	0	0	0.00219	···
otherwis	0.00204	0	0	0	···
exchang rate	0.00188	0	0	0	···
fs further appreci	0.00182	0	0	0.00194	···
wedg	0.00182	0.00187	0	0	···
februari	0	0.00229	0.00321	0.00184	···
budget	0	0.00217	0.00599	0	···
march	0	0.00166	0.00528	0.00532	···
pre-budget report	0	0.00165	0.00285	0	···
···	···	···	···	···	···

表 1: 各文章から抽出した単語とスコアの例．これを主成分分析にかけることで，単語のスコアで表されていた各文章の特徴が主成分スコアとなり，低い次元に特徴が凝縮される．

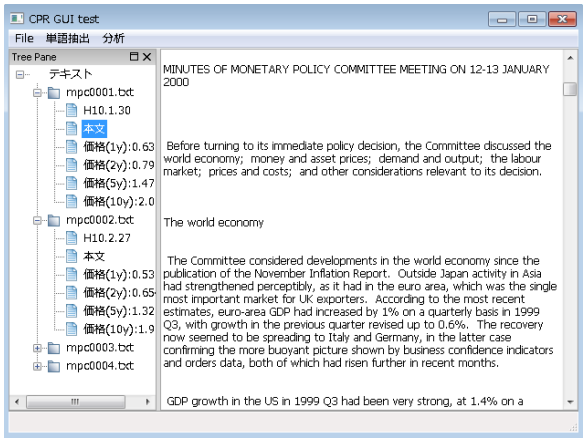


図 1: 開発したツールの画面

6 実験

6.1 使用データ

本研究では，テキストデータとして英国の中央銀行である英国中央銀行（Bank of England，BOE）の金融政策委員会（Monetary Policy Committee，MPC）が発行している議事録 [7] を対象に選んだ．

金融政策委員会は，毎月上旬に 2 日連続で開催され，政策金利変更は，2 日目の正午に発表され，市場の注目を集める．議事録はその 2 週間後に 10 ページ前後の分量で公表される．分析対象として，英国中央銀行の金融政策委員会の議事録を選んだ理由は，月次のレポートであり，文章の段落構造がある程度決まっており，時系列分析を行いやすいからである．また，英国中央銀行の金融政策委員会の議事録は，金融関係者が常に注目していおり，市場への影響力が大きいと考えられる．下記に 2009 年 1 月英国中央銀行の金融政策委員会の議事録の一部を記載しておく．

> Financial markets had been volatile since the December MPC meeting. Despite this, there had been a few encouraging signs. Sentiment in some markets had improved a little in the few days since the start of 2009. Equity prices had risen internationally on the month and the FTSE All-Share index was back close to the level prevailing at the time of the November Inflation Report. There had recently been some investment-grade corporate bond issuance which had been largely absent during the previous few months.

また，予測対象の時系列データとして議事録が発表された月の月末の英国国債金利 10 年物の金利とした．議事録が公開されるのは毎月 20 日頃であり，それが金利に反映するのに一週間ほどかかると仮定している．

6.2 実験方法

実験としては，5 年間の試験期間として，1 年間の外挿予測を行った．1999 年 1 月から 2003 年 12 月までの 60ヶ月を試験期間とし，議事録のテキストと月末の金利から回帰式を求める．その回帰式を使い，2004 年 1

月から 12 月までの外挿テストを行った. これを一年ずつずらし，合計で 6 回分のデータセットで行った.

一つの議事録からは 50 単語ずつ抽出し，TFIDF,TF, LOG 1 P のそれぞれで重み付けを行った. 試験期間における実値と回帰式による値との相関関数 (R^2)，外層期間における平均二乗誤差 (%) をそれぞれの重み付けの方法により比較した.

6.3 結果

各期間における実際の金利の変動と回帰による予測値の関係を図 2 に示す. また，各単語の重み付け方法による試験期間の相関関数と，外層期間における平均二乗誤差を表 2 にまとめた. 図 2 より，年により外挿区間の予測値と実際の値との一致度が異なることがわかる.

7 まとめ

本研究では，従来のテキストデータを用いた長期的な市場分析に新たな拡張を加えた上で，GUI のシミュレーションツールを開発した. ツールを使い，英国中央銀行の金融政策委員会を対象に英国国債金利の予測を行った. 今後は，連邦準備制度や欧州中央銀行など他の国の中央銀行英文テキストを試みる予定である.

参考文献

[1] 山口和孝: ニューラルネットと遺伝的アルゴリズムを用いた 株式売買支援システム, 2002

[2] 日本銀行: 日本銀行調査月報,1998 年 6 月号，108 ページ

[3] Desh Peramunetilleke, Raymond K. Wong: Currency Exchange Rate Forecasting from News Headlines, *The Thirteenth Australasian Database ConferenceL*(2001)

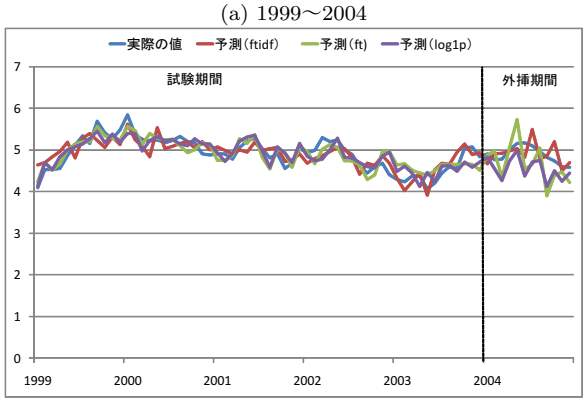
[4] Shimon Kogan, Dimitry Levin, Bryan R.Routledge, Jacob S.Sagi, Noah A.Smith: Predicting Risk from Financial Reports with Regression, *The 2009 Annual Conference of the North American Chapter of the ACL*,pages 272-280,June 2009

[5] Johan Bollen, Huina Mao, Xiao-Jun Zeng: Twitter mood Predicts the stock marcket, *arXiv*:1010.3003v1 14 Oct 2010

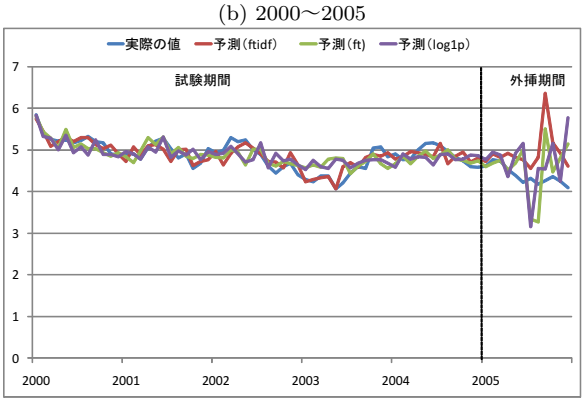
[6] 和泉 潔，後藤 卓，松井 藤五郎: テキスト情報による金融市場変動の要因分析, *2009* 年度人工知能学会全国大会

[7] イングランド銀行: 金融政策委員会が発行している議事録, http://www.bankofengland.co.uk/publications/minutes/mpc/index.htm

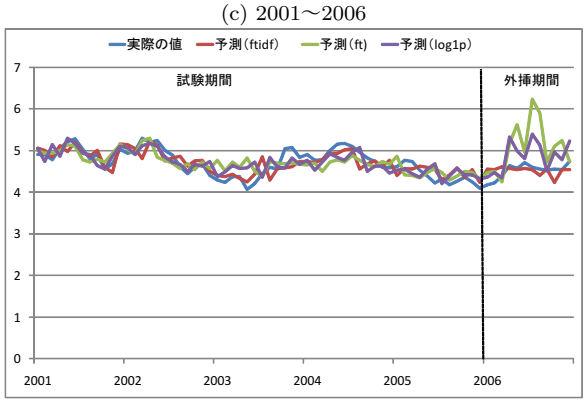
(a) 1999〜2004



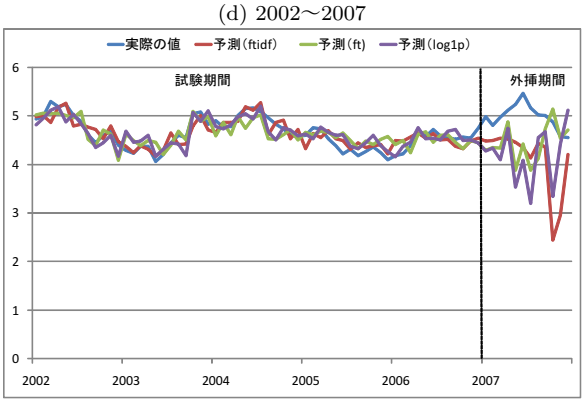
(b) 2000〜2005



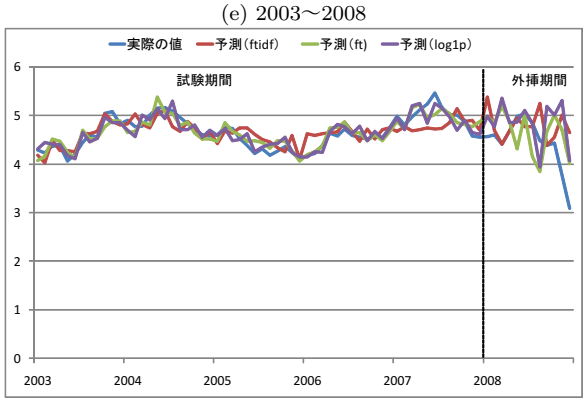
(c) 2001〜2006



(d) 2002〜2007



(e) 2003〜2008



(f) 2004〜2009

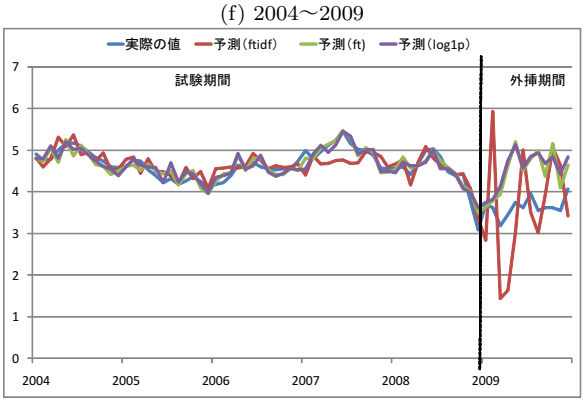


図 2: 各市場トレンドの推定．訓練期間: 1998 年 1 月〜2007 年 12 月，外挿期間: 2008 年 1 月〜12 月．

試験期間		1999-2003	2000-2004	2001-2005	2002-2006	2003-2009	2004-2008
外挿期間		2004	2005	2006	2007	2008	2009
試験期間における	ftidf	0.67	0.74	0.63	0.73	0.5	0.63
実値と回帰値との	ft	0.70	0.53	0.47	0.58	0.79	0.83
相関関数 (R^2)	log1p	0.70	0.44	0.57	0.69	0.80	0.78
外挿期間における	ftidf	4.76	17.31	4.51	21.07	15.03	34.69
平均二乗誤差 (%)	ft	8.94	15.38	16.83	14.98	13.55	26.59
	log1p	8.23	16.44	9.44	21.14	15.58	27.67

表 2: ftidf の場合