様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　１６　日現在

機関番号：　３４５０６
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３０３６３

研究課題名（和文）リアルタイム検索を基盤とした時空間テキストマイニング

研究課題名（英文）Spatio-temporal text mining based on real-time information retrieval

研究代表者
関　和広（Seki, Kazuhiro）

甲南大学・知能情報学部・准教授

研究者番号：３０４４４５６６

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：本研究では，ユーザが検索した時点での情報の価値を重視したリアルタイム検索を基盤技術として確立することを第一の目的とし，その後，時空間的なテキストマイニングに取り組んだ．具体的な成果としては，話題の時間的変化を考慮したマイクロブログ検索モデルの提案，および再帰的ニューラルネットワークを用いたニュース速報からの金融指標（株価）の短期的な変動予測手法の提案があげられる．

研究成果の概要（英文）：This research project first studied realtime information retrieval systems, which consider temporal properties of users' information needs, as a building block of intelligent information systems, and then investigated spatio-temporal text mining. The main outcome of the project is two-fold: one is to devise a realtime microblog retrieval model modeling trends of topics, and the other is the development of a framework to predict short-term stock price movements using recursive neural networks analyzing breaking news headlines.

研究分野：　知的情報システム

キーワード：　リアルタイム検索　株価動向予測

１．研究開始当初の背景

インターネットやモバイルコンピューティングの大衆化によって，CGM（Consumer Generated Media），特に Twitter に代表されるマイクロブログのユーザが急速に増加している．Twitter は，これまでのテキストマイニングが対象としてきたメディアと比較し，まさに今そこで起きている事象に対して生成されるコンテンツであり，非常にリアルタイム性が高いという特徴がある．例えば，エジプトの民主化運動やオリンピックのような時間とともにダイナミックに状況が変化する事象に関して，その発生や盛り上がりと連動して多くの投稿がなされる．このような特徴から，マイクロブログは，任意の事象に関して刻一刻と変わる状況を，その当事者や目撃者・聴衆から直接取得するための貴重な情報源となりうる．しかしながら，Twitter には 1 日に約 3 億 4 千万のツイート（ユーザが発信するメッセージ）が投稿され，かつ，一つひとつの投稿が 140 文字と短文であるため，そもそも特定の話題についてのツイートを正確に取得することが難しい．言い換えると，ある話題をクエリとして Twitter を検索しても，多くの無関係なツイートが取得されるため，検索精度が低いという問題がある．このため，マイクロブログから特定の話題に関する情報を取得・利用するためには，まずマイクロブログを対象とした情報検索の精度を向上させることが急務である．特に，マイクロブログ検索では検索した時点での情報の価値が重要であり，これを考慮した検索は「リアルタイム検索」と呼ばれる．一方，このようなリアルタイム性の高いテキスト情報を分析することで，実世界のイベントの発生を検知する研究が国内外で盛んに研究されている．

２．研究の目的

以上のような背景から，本研究では次の 2 つのテーマを扱った．

①時間的な語の出現・共起の傾向に着目したリアルタイム検索．
研究代表者らの予備実験により，ある話題に関連する語は，その話題と類似の時間的推移をもってマイクロブログで使われることが多いことが明らかになっている．この特徴をユーザの検索クエリの自動拡張に利用することで，リアルタイム検索の精度向上を図る．

②実世界のイベント予測
マイクロブログ等のメディアはリアルタイム性が高く，現実の事象を強く反映している．そのため，これらのテキストを時間的に分析すれば，盛り上がっている話題の検出や特定の対象に対する世論の変化を追跡することができる．本研究では，マイクロブログと同様にリアルタイム性が高いながらもノイズが少ないニュース速報から抽出される種々の事象と経済指標の変動の因果関係を大量のデータを分析することで発見し，これを基に経済指標の予測を試みる．

３．研究の方法

①マイクロブログ検索には，単語を用いた疑似適合フィードバックによるクエリ拡張が有効である．しかし，単語は意味的・時間的な曖昧性を持つため，単語を用いたクエリ拡張は有効に機能しない場合がある．そこで本研究では，単語や 2 語以上の単語の組合せであるコンセプトを用いた疑似適合フィードバックによるクエリ拡張手法を提案した．さらに，検索クエリと同時期に出現するコンセプトの頻度の時間遷移に関する情報を疑似適合フィードバックに組み入れることで，マイクロブログサービスのリアルタイム性を考慮した．

②ニュース記事などのテキスト情報が企業の株価動向に影響を与えると仮定し，その影響を予測するためのモデルとして，自然言語文に対して構文構造を表現できる Recursive Neural Network（RNN）の感情分析モデルを援用し，テキスト情報の構文構造を用いて，株価動向に与える影響を予測する手法を提案した．なお，RNN モデルの学習を行うためには，構文木の各ノードに相当する単語に株価変動の動向を表すラベルを付与する必要があるため，過去の新聞記事とその配信時の株価の変動データから，自動的にラベルを付与して訓練データを生成する枠組みを提案した．

４．研究成果

①本提案手法は，疑似適合フィードバック手法の考えを用いて，コンセプトが語彙情報と時間情報の 2 つの異なる情報源から成り立つと考え，語彙情報としての疑似適合文書中でのコンセプトの頻度，時間情報としての疑似適合文書中でのコンセプトの頻度の変化を言語モデルに基づく情報検索の枠組みに導入している．TREC2011 と 2012 のマイクロブログトラックのデータセットを用いた実験から，従来の語彙情報だけを用いる適合フィードバックよりも本提案手法が複数の評価指標で優れた検索性能を示すことが分かった．特に，コンセプトの語彙情報と時間情報を用いる手法は，非常に適合する文書の検索に対して優れた検索性能を示すことが分かった（表 1）．

②本研究では，従来研究でよく行われている 1 日後の株価動向予測の問題点を挙げ，この問題に対処するために記事発行後の最初の株価動向を予測対象とした．そして，記事が

表 1. 単語の語彙情報と時間情報を使ったモデル（wTRM）とコンセプトの語彙情報と時間情報を使ったモデル（cTRM）の実験結果.

	allrel			*highrel*		
Method	**AP**	**P@30**	**bpref**	**AP**	**P@30**	**bpref**
wTRM	**0.3726**	**0.4660**$^{\beta}$	**0.3872**	0.2580	0.2094	0.2361
cTRM	0.3644	0.4485	0.3825	**0.2694**	**0.2101**	**0.2527**

株価に影響与える期間をパラメータとして設定し，その期間内で株価が変動した見出しのみを用いて実験を行い，結果を評価した．その結果，記事が株価に影響を与える期間を 15 分間と設定したとき，精度で 60.41%，F 値で 0.60 となり，最も良い結果が得られた．また，その期間に対して構文構造を考慮していないモデルである bag-of-words や VecAvg との比較実験を行った結果，RNN が最も高い予測性能を示した．また，テストデータの中で記事数上位 10 銘柄に注目したところ，bag-of-words や VecAvg に対して，F 値で有意な結果向上が確認された（表 2）.

表 2. 株価動向予測結果（F 値）.

	RNN	bag-of-words	VecAvg	ratio
パナソニック	**0.53**	0.29	0.38	0.38
東洋証券	**0.40**	0.17	**0.40**	0.45
あおぞら銀行	**0.67**	0.25	0.38	0.38
大和証券	**0.46**	0.41	0.41	0.38
いちよし証券	**0.68**	0.58	0.35	0.38
沖電気工業	**0.73**	0.53	0.69	0.39
東芝	**0.52**	0.41	**0.52**	0.35
昭和電工	**0.65**	0.55	**0.65**	0.40
サントリー	0.68	0.67	**0.73**	0.44
NEC	**0.55**	0.48	0.43	0.36

また，これらの研究課題に関連して，米国 NIST が主催する評価型ワークショップ TREC の Knowledge Base Acceleration トラックへの参加などを通し，情報抽出に関する研究を行った.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 10 件）

① Akira Yoshihara, Kazuhiro Seki, and Kuniaki Uehara. Leveraging Temporal Properties of News Events for Stock Market Prediction. *Artificial Intelligence Research*, Vol. 5, No. 1, pp. 103-110, January 2016. 査読あり

② Shun Kawahara, Kazuhiro Seki, and Kuniaki Uehara. Detecting Vital Documents Massive Data Streams. *Open Journal of Web Technologies*, Vol. 2, No. 1, pp. 16-26, 2015. 査読あり

③ Kazuhiro Seki. Hypothesis Discovery Exploiting Closed Chains of Relations. *Transactions on Large-Scale Data- and Knowledge-Centered Systems*, Vol. 22, pp. 145-164, December 2015. 査読あり

④ 東山翔平，関和広，上原邦昭．医療用語資源の語彙拡張と診療情報抽出への応用．自然言語処理，Vol. 22，No. 2，pp. 77-106，June 2015．査読あり

⑤ Shohei Higashiyama, Mathieu Blondel, Kazuhiro Seki, and Kuniaki Uehara. Cost-Sensitive Structured Perceptron Incorporating Category Hierarchy for Named Entity Recognition. *Journal of Information and Communication Technology (JICT)*, Vol. 14, pp. 1-20, May 2015. 査読あり

⑥ 宮西大樹，関和広，上原邦昭．コンセプト追跡を用いたマイクロブログ検索. 情報処理学会論文誌：データベース，Vol. 7, No. 2, pp. 1-10, June 2014. 査読あり

⑦ 宮西大樹，関和広，上原邦昭．マイクロブログ文書の選択による適合フィードバックを用いた疑似適合フィードバックの検索性能改善. 情報処理学会論文誌，Vol. 55，No. 5，pp. 1585-1594，May 2014．査読あり

⑧ 関和広，上原邦昭．三段論法的パターンに着目した解釈容易な仮説の生成規則獲得と順位付け．情報処理学会論文誌，Vol. 55，No. 4，pp. 1428-1437，April 2014．査読あり

⑨ 東山翔平, ブロンデルマチュー, 関和広, 上原邦昭. カテゴリ階層を考慮した構造化パーセプトロンによる固有表現抽出. 情報処理学会論文誌:数理モデル化と応用, Vol. 6, No. 3, pp. 43-52, December 2013．査読あり

⑩ Mathieu Blondel, Kazuhiro Seki, and Kuniaki Uehara. Block Coordinate Descent Algorithms for Large-scale Sparse Multiclass Classification. *Machine Learning*, Vol. 93, No. 1, pp. 31-52, October 2013. 査読あり

〔学会発表〕（計 16 件）

① Shun Kawahara, Kazuhiro Seki, and Kuniaki Uehara. Detecting Vital Documents Using Negative Relevance Feedback in Distributed Realtime Computation Framework. In *Proceedings of the 2015 Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pp. 101-108, May 20, 2015. Bali（インドネシア）．査読あり

② 秋田諒，吉原輝，関和広，上原邦昭．再帰的ニューラルネットワークによる感

情分析モデルを用いた株価動向予測. 第 29 回人工知能学会全国大会（JSAI2015）, 1J4-OS-13a-5, 4 pages, 2015 年 5 月 30 日. はこだて未来大学（北海道・函館市）. 査読なし

③ 吉原輝, 関和広, 上原邦昭. ニュース記事の時間的特性を考慮した株価動向予測. 第 102 回数理モデル化と問題解決研究発表会, Vol. 102, No. 4, pp. 1-6, 2015 年 3 月 3 日. 島原文化会館（長崎県・島原市）. 査読なし

④ 川原 駿, 関和広, 上原邦昭. 分散ストリーム処理基盤 Storm と言語モデリングによる新情報を含む文書の検出. 第 7 回データ工学と情報マネジメントに関するフォーラム（第 13 回日本データベース学会年次大会）(DEIM2015), E4-3, 7 pages, 2015 年 3 月 3 日. 磐梯熱海ホテル華の湯（福島県・郡山市）. 査読なし

⑤ Akira Yoshihara, Kazuki Fujikawa, Kazuhiro Seki, and Kuniaki Uehara. Predicting the Trend of the Stock Market by Recurrent Deep Neural Networks. In *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI-2014)*, pp. 759-769, December 5, 2014. Gold Coast（オーストラリア）. 査読あり

⑥ Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. Time-Aware Latent Concept Expansion for Microblog Search. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*, pp. 366-375, June 2, 2014. Michigan（アメリカ）. 査読あり

⑦ Shohei Higashiyama, Mathieu Blondel, Kazuhiro Seki, and Kuniaki Uehara. A Cost-Sensitive Approach to Named Entity Recognition with Category Hierarchy. In *Proceedings of the International Conference on Computer and Information Sciences (ICCOINS2014)*, CD-ROM, 6 pages, June 4, 2014. Kuala Lumpur（マレーシア）. 査読あり

⑧ 吉原輝, 藤川和樹, 関和広, 上原邦昭. 深層学習による経済指標動向推定. 第 28 回人工知能学会全国大会（JSAI2014）, 3H3-OS-24a-5, 4 pages, 2014 年 5 月 14 日. ひめぎんホール（愛媛県・松山市）. 査読なし

⑨ 藤川和樹, 関和広, 上原邦昭. 言語情報を用いた経済指標の予測と分析. 第 28 回人工知能学会全国大会（JSAI2014）, 3L4-OS-26b, 4 pages, 2014 年 5 月 14 日. ひめぎんホール（愛媛県・松山市）. 査読なし

⑩ Sayaka Kitaguchi, Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. Interactive Disaster Information Search System for Microblog by Minimal User Feedback. In *Proceedings of the 9th Asia Information Retrieval Societies Conference (AIRS 2013)*, pp. 476-487. December 9, 2013. シンガポール. 査読あり

⑪ 宮西大樹, 関和広, 上原邦昭. コンセプト追跡を用いたマイクロブログ検索. 第 6 回 Web とデータベースに関するフォーラム（WebDB Forum 2013）, CD-ROM, 8 pages, 2013 年 11 月 27 日. 京都大学（京都府・京都市）. 査読あり

⑫ Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. Improving Pseudo-Relevance Feedback via Tweet Selection. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, pp. 439-448, October 29, 2013. San Francisco（アメリカ）. 査読あり

⑬ Shohei Higashiyama, Kazuhiro Seki, and Kuniaki Uehara. Developing ML-based Systems to Extract Medical Information from Japanese Medical History Summaries. In *Proceedings of the First Workshop on Natural Language Processing for Medical and Healthcare Fields*. October 18, 2013. 名古屋国際会議場（愛知・名古屋市）. 査読あり

⑭ Koji Kumanami, Kazuhiro Seki, and Kuniaki Uehara. Agglomerative Co-Clustering for Synonymous Phrases Based on Common Effects and Influences. In *Proceedings of the IEEE Big Data 2013 Workshop on Scalable Machine Learning*, pp. 87-94, October 8, 2013. Santa Clara（アメリカ）. 査読あり

⑮ Kazuhiro Seki, and Kuniaki Uehara. Supervised Hypothesis Discovery Using Syllogistic Patterns in the Biomedical Literature. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, pp. 1663-1669, August 9, 2013. 北京（中国）. 査読あり

⑯ 宮西大樹, 関和広, 上原邦昭. マイクロブログ文書の選択による擬似適合フィードバック. 情報処理学会研究報告 データベース・システム（DBS）, 2013-DBS-157(15), pp. 1-6, 2013 年 7 月 23 日. 北海道大学（北海道・札幌市）. 査読なし

６．研究組織

(1)研究代表者

関 和広（SEKI, Kazuhiro）
甲南大学・知能情報学部・准教授
研究者番号：30444566