経済テキストからの市況分析コメントの自動生成

Automatic generation of market analysis comments from financial articles

酒井 浩之 [1*]　坂地 泰紀 [2]　和泉 潔 [2]　松井 藤五郎 [3]　入江 圭太郎 [4†]
Hiroyuki Sakai[1]　Hiroki Sakaji[2]　Kiyoshi Izumi[2]　Tohgoroh Matsui[3]　Keitaro Irie[4]

[1] 成蹊大学 [1] Seikei University [2] 東京大学 [2] The University of Tokyo
[3] 中部大学 [3] Chubu University
[4] 三菱 UFJ 国際投信 [4] Mitsubishi UFJ Kokusai Asset Management

Abstract: 本研究では，経済新聞記事などの経済テキストから、日経平均株価などの市況について言及している文書のみを抽出し，それらの内容を自動的に要約することにより，ファンドの運用報告書における市況分析コメントを自動生成する手法の開発を行う．本手法では，まず経済新聞記事から深層学習により日経平均株価の市況について言及している記事を抽出する．次に抽出された記事の中から例えば「ギリシャへの金融支援協議が難航していることや、中国・上海株の値動きへの警戒感から、投資家のリスクオフの動きが強まった。」のような日経平均が大幅に変動した理由について言及している文を抽出する．そして，抽出された文を時系列順に並べることで市況分析コメントを自動生成する．

1　はじめに

近年，証券市場における個人投資家の比重が増大しており，個人投資家に対して投資判断の支援をおこなう技術の必要性が高まっている．そのため，人工知能分野の手法や技術を，金融市場における様々な場面に応用することが期待されており，例えば，膨大な金融情報を分析して投資判断の支援を行う技術が注目されている．その一例として，日本銀行が毎月発行している「金融経済月報」や経済新聞記事をテキストマイニングの技術を用いて解析し，経済市場を分析する研究などが盛んに行われている [1][5].

本研究では，毎月のファンドの運用報告書に記載される市況分析コメントを自動生成する手法を提案する．市況分析コメントとは，例えば「8 月の国内株式市況は、中国の景気減速懸念が台頭したことなどを背景とした世界的な株安を受けて大きく下落しました。」のような，その月における株価が大きく変動したイベント（例えば,「人民元の基準値切り下げ」）について述べ，株価が変動した理由を分析した文書である．以下に，2015 年 8 月のファンド運用報告書に記載された市況分析コメントの一部を示す.

> 8 月の国内株式市況は、2015 年度第 1 四半期決算で好業績を発表した企業への期待などを背景に上昇して始まりました。しかしながら中国人民銀行が人民元の対米ドルでの基準値切り下げを実施すると中国経済への減速懸念が広がり、国内株式市況は下落しました。さらに、中国経済の減速が世界景気へ及ぼす影響などを警戒して投資家がリスク回避姿勢を強めると世界的に株式市況は急落しました。…

上記のような市況分析コメントを記述するために，ファンド運用の担当者は①日経平均株価が大きく動いた記事を調べ，②その前後にあったイベントを確認し，③その記事の中から株価が変動した理由について述べている文を選択し，④まとめる，という作業を毎月，行う必要がある．現在のところ，市況分析コメントの作成はファンドごとに運用担当者が行っており，ファンドの特色に従ってファンドごとに異なる．しかし，①〜③については共通化しＡＩで自動化できれば，ファンドの運用担当者の負担を減らすことが可能である [1].

そこで，本研究では，上記の①日経平均株価が大きく動いた記事の判定，②その前後にあったイベントの確認，③その記事の中から株価が変動した理由について述べている文を抽出，といった処理を自動化し，③

*連絡先：成蹊大学 理工学部 情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: h-sakai@st.seikei.ac.jp
†本論文に示された所見は著者自らのものであり，所属する組織の公的な立場を代表するものではない．

[1]④についてはファンドの特色に合わせ，ファンド担当者がまとめてもよい．

の処理によって抽出された文を時系列順に並べることで市況分析コメントを自動生成することを目的とする．

2 関連研究

関連研究として，決算短信を対象として様々な情報を抽出する研究がある．酒井らは決算短信から業績要因を含む文（例えば「半導体製造装置の受注が好調でした。」）を抽出する手法を提案している [8][9]．坂地らは決算短信から原因・結果表現を抽出する手法を提案している [10]．北森らは決算短信から業績予測文 (今後の業績予測に関する情報が記述されている文) を抽出する手法を提案している [2][3]．いずれの研究も抽出対象である情報を抽出するために有効な手がかり表現に着目し，それらの表現をブートストラップ的に自動的に獲得，もしくは，人手にて用意している．また，深層学習を用いて情報を抽出している手法もあるが，手がかり表現を使用して抽出したデータを学習データとすることで，学習データの自動生成を試みている．本研究では，③その記事の中から株価が変動した理由について述べている文を抽出する処理のために手がかり表現を用いるが，その手がかり表現の獲得のために酒井らの手法 [8] を用いている．ただし，最初の入力する初期手がかり表現は，本研究における抽出対象にあわせ「で買い」「が買い」「で売り」「が買い」に変更している．

本研究は，複数の記事から 1 つの要約を生成する複数文書要約とみなすことができる．複数文書要約に関しては多くの研究があり，例えば酒井らは，ユーザが知りたい情報を「要約要求」と定義し，要約要求を反映した要約を生成するために，ユーザとのインタラクションを導入した複数文書要約システムを提案している [7]．複数文書要約では，入力として複数の記事が与えられ，それらの記事から重要な文を抽出し，重要な文同士の冗長性を排除してまとめるという処理を行い，入力された記事の内容をまんべんなく含むような要約を生成することが求められる．それに対して，本研究で与えられる記事には市況分析コメントを生成するためには不要な記事も含まれており（すなわち，日経平均株価の市況について述べてはいるが，大きく動いた日ではない記事），そのような記事を排除する必要がある点が異なる．さらに，市況分析コメントの重要文として株価が変動した理由について述べている文を抽出している点も異なる．

3 市況分析コメント自動生成手法

3.1 手法概要

本手法では，①日経平均株価が大きく動いた記事の判定，②その前後にあったイベントの確認，③その記事の中から株価が変動した理由について述べている文を抽出，の順で処理を行い，市況コメントを生成する．手法の概要を以下に示す [2]．

Step 1: ある期間の日経平均について言及している記事から，日経平均が大きく変動したことについて述べた記事（以降，分析記事と定義）を深層学習により抽出．

Step 2: ある期間の日経平均について言及している記事集合から，その期間における重要なキーワード（以降、重要キーワード）を抽出（例 : 人民元，中国人民銀行）．

Step 3: Step 1 で抽出した分析記事集合より，日経平均が大きく変動した要因について述べた文（以降，要因文と定義）を抽出．

Step 4: Step 3 で抽出した要因文集合と，Step 2 で抽出した重要キーワード集合を使用し，重要な文を判定

Step 5: Step 4 で判定された重要な文を時系列順にならべ，市況分析コメントとする

3.2 分析記事の抽出

市況分析コメントを自動生成するための情報源として日経平均株価について言及している記事を使用する．しかし，日経平均株価について言及している記事は，ほぼ毎日 1 つは存在するため，それらの記事全てを市況分析コメント自動生成のための情報源として使用すれば重要ではない情報も混ざる．日経平均株価の実データを使用して記事を選別する方法でもいいが，その場合は日経平均株価の変動とその日に対応する記事をセットで用意する必要があり，入力フォーマットが複雑になる．そこで，日経平均株価が大幅に変動したことに言及した記事（分析記事）を深層学習にて自動的に抽出する．以下に分析記事の一部を示す．

[2]Step 1 が①に，Spte 2 が②に，Step 3 が③に該当する．

１２日の東京株式市場で日経平均が大幅に続落し、下げ幅は一時３００円を上回った。中国の景気減速への警戒感が広がり、運用リスクを減らす動きが優勢になっている。人民元切り下げが発表になった午前１０時１５分すぎから先物に海外勢からとみられる大口の売りが断続的に出て、日経平均は下げを加速した。…

3.2.1 学習データの自動生成

深層学習のための学習データは，１４年分の日本経済新聞記事のタイトルに「日経平均」が含まれている記事から以下の手法で自動生成した（2772 記事を生成）．

正例： 第 1 文に「日経平均株価は大幅」、「日経平均株価が大幅」が含まれている記事

負例： 第 1 文に「日経平均」がない

上記の手法により，以下のような学習データが自動生成される．

十日の東京株式市場で日経平均株価が大幅続落し、終値で一万一〇〇円台に下落した。バブル経済崩壊後の安値を再び更新し、一九八四年八月以来の水準となった。主力のハイテク株や通信株が相場の下げを主導。銀行、鉄鋼、不動産など幅広い銘柄に売りが膨らんだ。…

負例としては，インタビュー記事などが抽出される．

株安が直ちに一九九八年のような金融システム不安を引き起こすことはない。大手銀行の自己資本比率は一〇％を超えており、株価下落による比率低下は限定的。 …

3.2.2 素性選択

自動生成された学習データから入力層の要素となる語（素性）を選択する．具体的には，自動生成された学習データにおいて正例の記事に含まれる内容語（名詞，動詞，形容詞）に対して，以下の式 1 にて重みを計算する．

$$W_p(t, S_p) = TF(t, S_p) \times H(t, S_p) \tag{1}$$

ただし，

S_p**:** 学習データにおいて正例に属する記事集合

$TF(t, S_p)$**:** 記事集合 S_p において，語 t が出現する頻度

$H(t, S_p)$**:** 記事集合 S_p における各記事に含まれる語 t の出現確率に基づくエントロピー

$H(t, S_p)$ が高い語ほど，正例の記事集合に均一に分布している語であることが分かる．$H(t, S_p)$ は次の式 2 で求める．

$$H(t, S_p) = -\sum_{s \in S_p} P(t, s) \log_2 P(t, s) \tag{2}$$

$$P(t, s) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)} \tag{3}$$

ここで，$P(t, s)$ は記事 s における語 t の出現確率を表し，$tf(t, s)$ は記事 s において語 t が出現する頻度を表す．

次に，負例の記事に含まれる内容語（名詞，動詞，形容詞）に対しても，同様に重みを計算する．

$$W_n(t, S_n) = TF(t, S_n) \times H(t, S_n) \tag{4}$$

ただし，S_n は学習データにおいて負例に属する記事の集合である．

ここで，ある語 t の正例における重み $W_p(t, S_p)$ が負例における重み $W_n(t, S_n)$ の 2 倍より大きければ，その語 t を素性として選択する．もしくは，語 t の負例における重み $W_n(t, S_n)$ が正例における重み $W_p(t, S_p)$ の 2 倍より大きければ，その語 t を素性として選択する．上記の条件を課すことで，正例，負例における特徴的な語のみを素性として選択し，正例，負例，ともによく出現するような一般的な語を素性から除去する．上記の手法により，61,678 文の学習データから 4,549 語が素性として選択された．以下に選択された素性の一部を示す．

平均, 市場, 株価, 株式, 投資, 相場, 証券, 売り, 買い

3.2.3 深層学習による分析記事の抽出

深層学習のモデルについて以下に述べる．入力は，2772 記事の学習データから素性として抽出された 4,252 語を要素，語 t における $W_p(t, S_p)$，もしくは，$W_n(t, S_n)$ の大きいほうを要素値としたベクトルとする．モデルの入力層のノード数を入力ベクトルの次元数と同じ 4,252 とし，隠れ層は，ノード数 1,000 が 3 層，ノード数 500 が 3 層，ノード数 200 が 3 層，ノード数 100 が 3 層の計 12 層とする．出力層は 1 要素である．また，活性化関数として，ReLU を使用した．学習されたモデルにより，市況コメントを生成する期間の日経新聞記事から分析記事を抽出する．テストデータは，記事のタイトルに「日経平均」が含まれる記事である．

3.3 分析記事からの要因文の抽出

分析記事を並べただけでは市況コメントとして長すぎるため，分析記事から日経平均が大幅に変動した理由について言及している文（要因文）を抽出する．例えば，以下のような文を抽出する．

> ・ギリシャへの金融支援協議が難航していることや、中国・上海株の値動きへの警戒感から、投資家のリスクオフの動きが強まった。
> ・米消費者物価指数の上昇やイエレン米連邦準備理事会議長が年内の利上げに前向きな姿勢を示し、為替相場が円安に向かった。

要因文の抽出は,「強まった」のような手がかり表現や,「ギリシャ」「金融支援」「イエレン米連邦準備理事会議長」といった，その期間における重要なキーワードを使用して抽出する．しかし，期間ごとの重要キーワード，有効な手がかり表現は数多く，全て人手で用意することは困難である．そこで，手がかり表現と重要キーワードを自動獲得する．

関連研究でも述べたが，要因文の抽出は酒井らが決算短信から業績要因文を抽出するために開発した手法 [8] を使用する．手がかり表現は以下の手法で獲得される．

Step 1: 少数の手がかり表現（具体的には,「が買い」,「で売り」の 2 表現を用いる）を人手で与え，それに係る節を取得する．

Step 2: 取得した節の集合から，その中で共通して頻繁に出現する表現（「警戒感」,「国際優良株」など）を共通頻出表現と定義し，抽出する．

Step 3: 共通頻出表現が係る節を取得し，その中から新たな手がかり表現を抽出する．

Step 4: 獲得した手がかり表現から，それに係る節を取得する．

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す（図 1 を参照）．□

Step 2 において，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 5 で求め，その値が，ある閾値以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \qquad (5)$$

ここで，分析記事の集合において，

$S(e)$: 共通頻出表現 e が係る手がかり表現の集合．

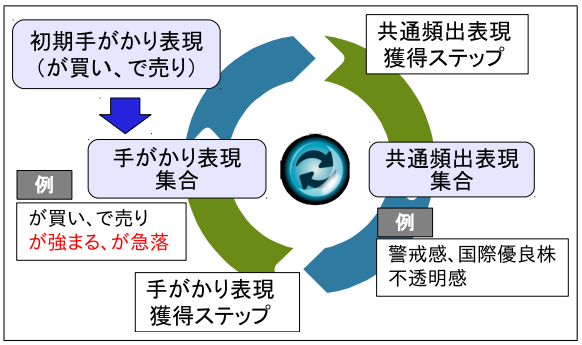

図 1: 共通頻出表現・手がかり表現自動獲得手法の概要

$P(e,s)$: 共通頻出表現 e が手がかり表現 s に係る確率．

同様に，Step 3 において，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定に基づき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを求め，その値が，ある閾値以上の手がかり表現を選別する．

以上の手がかり表現，共通頻出表現の選別処理を行うことで，例えば以下のような適切な手がかり表現を獲得する．

> が上昇，が強まる，が膨らんでいる，が急落，が堅調だった，が大幅下落した，で大幅続伸，が根強い，が急伸した，が後退，が加速した，が先行している

期間ごとの重要キーワードの抽出は，期間 t の分析記事における名詞 n に対して，以下の式 6 で重み $W(n, S(t))$ を計算することで行う．

$$W(n,S(t)) = (0.5 + 0.5 \times \frac{tf(n,S(t)}{\max tf(n,S(t)}) \times H(n,S(t)) \times \log_2 \frac{N}{df(n,N)} \qquad (6)$$

ここで，

$S(t)$: 期間 t の分析記事の集合．

$tf(n,S(t))$: $S(t)$ において，名詞 n が出現する頻度．

$H(n,S(t))$: $S(t)$ の各分析記事である d に名詞 n が出現する確率 $P(n,d)$ に基づくエントロピー．

$df(n,N)$: 記事集合 N で名詞 n を含む記事の数．

N: 日経平均について言及している記事の総数．

ここで，N の日経平均について言及している記事とは，2000 年から 2014 年までの日経新聞記事において，タイトルに「日経平均」を含む記事であり，その総数は 14,325 記事である．

表 1: 期間ごとの重要キーワードの例

期間	重要キーワード
1999 年 11 月（ＩＴバブル）	情報通信，情報通信関連株，年初来高値
2008 年 9 月（リーマンショック）	米政府，金融安定化，リーマン・ブラザーズ，総合金融安定化策，破綻
2015 年 7 月（ギリシャ問題）	ギリシャ，離脱，金融支援，中国株，国民投票，欧州連合

$W(n, S(t))$ は，情報検索で一般的な $tf \cdot idf$ 値を 1 つの期間の分析記事の集合を 1 つの文書とみなして求め，さらに，その期間の分析記事集合においてまんべんなく出現している場合に高い値をとる尺度を組み合わせたものである．表 1 に，上記の手法によって，期間ごとの分析記事から抽出された重要キーワードをいくつか示す．分析からの要因文の抽出は，分析記事から手がかり表現と重要キーワードがともに含まれている文を抽出することで行う．

3.4 市況分析コメントの自動生成

分析記事から抽出した要因文を時系列順にならべることで，市況分析コメントを生成する．ここで，分析記事から抽出した要因文を全て採用すると，市況分析コメントとして長すぎる場合がある．そのため，要因文にふくまれる重要キーワードのスコアの和を分析記事のスコアとし，スコアが上位の分析記事に含まれる要因文を時系列順に並べて，市況分析コメントとする．2015 年 8 月のタイトルに「日経平均」を含む 33 記事から本手法にて生成された市況分析コメントを以下に示す．

> 今月の国内株式市況は、◆１１日の日経平均株価は前日の米国株高を手掛かりに一時１３８円高を付けたが、人民元切り下げで一転して売りが優勢となり、２２６円安まで値下がりする場面があった。人民元安は米国の利上げ観測から軟調に推移していたアジア通貨にも波及。アジアの主要な株式市場でも売りが広がった。◆１１日の東京株式市場で日経平均株価は一時、年初来高値を上回る水準に上昇した。しかし、中国人民銀行による人民元の実質切り下げを機に下落に転じた。人民元の切り下げの影響は日本にとどまらず、欧米やアジアなどの株式市場にも及んだ。◆中国の景気減速から始まった世界市場の動揺がいったん収まり、株式や原油などリスク資産の買い戻しが活発になっている。２８日の日経平均株価は３日続伸し、直近安値の２５日に比べて７％上げた。一方、一部の新興国通貨への売り圧力はなお強く、混乱再燃の懸念はくすぶっている。

表 2: 評価結果

期間	本手法	Baseline	運用担当者
1 月	**0.2**	0.12	0.48
2 月	**0.1**	0.05	0.51
3 月	**0.2**	0.06	0.27
4 月	**0.13**	0.1	0.49
5 月	**0.27**	0.1	0.59
6 月	0.10	**0.13**	0.41
7 月	0.18	**0.2**	0.38
8 月	**0.35**	0.22	0.58
9 月	**0.14**	0.06	0.43
１０月	**0.17**	0.1	0.4
１１月	**0.25**	0.15	0.25
１２月	**0.21**	0.1	0.36

4 評価

本手法の評価を行うため，本手法を実装した．実装にあたり，形態素解析器として MeCab[3]，係り受け解析器として CaboCha[4] を使用した．評価方法は，実際に運用担当者が作成した市況分析コメント（2015 年 1 月〜12 月）と，その期間の日経新聞記事を使用して本手法により自動生成された市況分析コメントとを比較して行った．具体的には，ある期間における運用担当者が作成した市況分析コメントと自動生成された市況分析コメントとの類似度を，その文書の名詞を要素，要素値として TF・IDF 値を使用したベクトル空間モデルで求める．そして，その期間における運用担当者が作成した市況分析コメントとの類似度の平均を評価値とした．ここで，比較手法として以下の手法と比較した．

Baseline: 入力された記事の第一文を連結して生成

運用担当者: 運用担当者が作成した市況分析コメントの 1 つを選択

運用担当者は理想的な結果に基づく評価値となる．結果を表 2 に示す．表の太字は，本手法と Baseline とを比較し，大きいほうを示す．

[3] http://taku910.github.io/mecab/

5 考察

本手法と Baseline とを比較すると，本手法のほうが概ね高い類似度を達成している．しかし，理想的な結果である運用担当者とは肉薄している月もあるが，まだ大きな差があることが分かる．5月，8月は重要な発言や株価に影響が大きなイベントがあったこともあり，高い類似度を達成しているが，6月は大きなイベントがなく（その月の日経平均の高値から安値を引いた変動幅が最も小さかった），そのような場合は類似度が低くなってしまう．

本研究における評価は，正解データである実際に運用担当者が作成した市況コメントとの類似度を測ることで行っている．しかし，本評価手法では，語の一致する割合が多いと類似度が高くなる傾向になるため，内容の冗長性や網羅性を評価できているわけではない．そのため，Baseline では冗長性が高い市況コメントが作成されているにもかかわらず，類似度が高くなることもある．テキスト自動要約では，ある文書から人間が作成した要約と自動生成された要約の間で一致する N グラムの割合で評価値を求める ROUGE[6] という評価手法が一般的によく用いられている．しかし，本タスクの場合，人間が作成した要約（運用担当者が作成した市況分析コメント）は，本研究でいうところの分析記事をもとに作成しているわけではないので，ROUGE で評価することが妥当ではなかった．理想的には，運用担当者が作成した市況分析コメントを評価者が読んで自動生成された市況分析コメントと比較し，どの程度，内容が一致しているか，冗長性が除かれているかを人手にて評価すべきであるが，それを行うにはある程度の専門知識が必要であることと再現性が困難であることから，今回は行えなかった．今後の課題として，本タスクにおける評価手法の確立が必要であると考える．

6 まとめ

本稿では，経済新聞記事などの経済テキストから、例えば日経平均株価などの市況について言及している文書のみを抽出し，それらの内容を自動的に要約することによりマーケットレポートにおける市況コメントを自動生成する手法について述べた．本手法では，まず経済新聞記事から深層学習により日経平均株価が大幅に変動したことについて言及している記事を抽出し，次に抽出された記事の中からその理由について言及している文を抽出した．そして，抽出された文を時系列順に並べることで市況コメントを自動生成した．評価の結果，入力された記事の第一文を連結して生成した文書より概ね高い類似度を達成した．

参考文献

[1] 和泉潔, 後藤卓, 松井藤五郎: 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011).

[2] 北森詩織, 酒井浩之, 坂地泰紀: 決算短信 PDF からの業績予測文の抽出, 電子情報通信学会論文誌 D, Vol. J100-D, No. 2, pp. 150–161 (2017).

[3] Kitamori, S., Sakai, H. and Sakaji, H.: Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning, *IEEE Symposium on Computational Intelligence for Financial Engineering & Economics*, pp. 67–73 (2017).

[4] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).

[5] 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志: 新聞記事のテキストマイニングによる長期市場動向の分析, 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296 (2013).

[6] Lin, C.-Y. and Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, *the 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology*, pp. 150–157 (2003).

[7] 酒井浩之, 増山繁: ユーザの要約要求を反映するためにユーザとのインタラクションを導入した複数文書要約システム, 日本知能情報ファジィ学会誌, Vol. 18, No. 2, pp. 265–279 (2006).

[8] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀: 企業の決算短信 PDF からの業績要因の抽出, 人工知能学会論文誌, Vol. J98-D, No. 5, pp. 172–182 (2015).

[9] 酒井浩之, 松下和暉: 決算短信からの業績要因文の抽出, 第 11 回テキストアナリティクス・シンポジウム, pp. 87–91 (2017).

[10] 坂地泰紀, 酒井浩之, 増山繁: 決算短信 PDF からの原因・結果表現の抽出, 電子情報通信学会論文誌 D, Vol. J98-D, No. 5, pp. 811–822 (2015).