サポートベクトルマシンを用いた関西空港の利用行動に関するブログページの収集

鷹尾　和享[1]

[1]正会員　（社）システム科学研究所（〒604-8223 京都市中京区新町通四条上ル小結棚町428 新町アイエスビル）
E-mail: takao@issr-kyoto.or.jp

近年ではブログとして多くの文章が作成されており、書き手の心理状態に関する率直な記述が含まれていることが期待できる。したがって、交通行動に関しても、ブログを分析することで、ヒトの「生の声」を捉えられれば、交通空間の満足度の向上を図るうえで有用であると考えられる。本稿では、ブログに比較的書かれやすい話題として、関西空港の利用行動に着目し、利用者目線の記述が書かれたブログページを収集することを目指す。しかし、ブログには数多くの話題があり、必要なのはそのうちのごく少数である。本稿は簡単に必要ページを収集する方法として、検索エンジンの検索結果から正例・負例に特徴的な語句パターンを用意し、必要なページをフィルタリングする方法について述べる。テストの結果、良い結果が得られた。

Key Words: *travel behaviour, data collection, blog, text data, text mining, natural language processing*

1. はじめに

(1) 背景と目的

筆者らはヒトの交通行動の心理的な側面に目を向けた研究を行ってきた。従来、交通行動に関する心理的な要素は、物理的に観測可能なデータから間接的に分析されることがしばしばあった。それに対し、筆者らのこれまでの研究では、テキストデータから、より直接的に経路選択プロセスを捉えることを試みてきた[1),2),3)]。

一方、近年ではブログとして多くの文章が作成されており、一般の人々の「生の声」[4)]が含まれていることが期待できる。そこで、ブログを分析することによって、交通空間に関するヒトの気持ちを捉えることを考えた。これはテキストデータからのデータマイニングであり、「テキストマイニング」と呼ばれる。

本稿では、ブログに書かれやすい話題として、関西空港の利用行動に着目し、お客様である利用者の気持ち・印象・評価といった「生の声」の記述をブログから抽出できると考えた。

関西空港に関して言えば、経営や財務の視点からは多くの議論がなされているのに対し、「お客様」であるはずの利用者の視点からの議論は十分なされているとは言えない。さらに、東アジアにおいては、空港に関する国際競争の激化に伴い、単純に需要を予測するのではなく、利用者の満足度を向上させ、利用を増進させるような戦略的な計画が必要とされる時代になっており、利用者の率直な気持ちを捉える必要性が高まっているのも関西空港の利用行動に着目したもう一つの理由である。

ブログには数多くの話題があり、本稿で必要なのはそのうちのごく少数である。関西空港の話題に関して言えば、経営的、政治的な視点の記述も数多くあり、必ずしも利用者の声とは言えない記述も多く存在する。したがって、単純に検索エンジンで「関西空港」を検索するだけでは不十分であり、「生の声」の抽出のためには、まず、必要ページの収集を行う必要がある。また、今日ではWeb上に極めて多くの情報が存在し、その中から本当に必要なものを効率よく取り出す研究に取り組むことは極めて有意義なことと考えられる。

そこで、本稿では、関西空港の利用行動に関する利用者目線の記述が書かれたブログページを収集することを目的とし、検索エンジンを活用して収集した後、必要ページをフィルタリングする方法について報告する。

(2) ブログテキストの特徴の整理

ブログのテキストは次の点で有益であると考えられる。第一は、ブログの多くは一般の人々によって書かれている点である。したがって、管理者や企業や行政の公式見解からは得られないような、旅行者本人あるいはエ

ンドユーザーの率直な気持ちを豊富に含んでいることが期待できる。第二は、従来の調査方法と比較して新鮮なデータを得ることが可能である点である。従来の訪問調査や電話インタビュー、郵送調査といった方法は綿密な計画が必要であり、手間やコストもかかるので、気軽に行うのは困難であり、そのため、一旦収集したデータを長期間使い続けることになりがちである。一方、Web上のデータはタイムラグなしにいつでもアクセス可能である。第三は、ブログは自発的に書かれている点である。したがって、従来のアンケート調査等の方法では調査者が予め用意した設問の視点で回答が求められるのに対し、ブログは書き手本人が着目した視点に基づいて本当に言いたいことが書かれることが期待できる。第四は、ブログは相手の顔を気にせずに遠慮なく記述できる点である。電話調査等の会話による方法では、日本人の民族性かも知れないが、面と向かって本音を言わない場合があることが報告されている[5]。それに対し、ブログは匿名であり、会話の相手もいないので、印象を悪くする心配をせずに済み、本音が記述されることが期待できる。

一方、次のような短所もある。まず、母集団の不透明性や代表性の欠落は奥村(2007)[4]が指摘している。また、匿名であるが故に、嘘を書いても責任を問われないため、信憑性が問題になる場合もある。

しかし、ブログは、本節で述べたような、他の調査方法では得られないような内容を含むことが期待される宝の山であり、また、Web上に書かれた内容が口コミで広まっていく場合もあり得るので軽視できないデータであり、ブログテキストをマイニングする価値は十分にあると思われる。

2. 関連研究

交通行動に関する印象を複数の尺度で捉えようとする研究例としてZhang(2009)[6]が挙げられる。しかし、この方法は予め用意した尺度でデータ収集を行う点が本稿と異なっている。

Webページを収集する方法は大きく分けて(i)自前でクローラーを開発する方法[7,8]、(ii)検索エンジンを利用する方法[9]、(iii)特定のサイトの記事を収集する方法[10]の３つがある。(i)は膨大な数のページをクロールすることになるため、本格的な体制を立ち上げて本格的なシステムを構築する必要に迫られることになる。また、(iii)は商品レビューのように、必要とする体系で予めカテゴリ分類されている場合は効率的であるが、本稿の場合にはふさわしい方法ではない。したがって、(ii)の検索エンジンを利用する方法で行う。

ブログページのフィルタリングに関しては、類似の研究例として文書分類というアプローチがある。たとえば、橋本ら(2008)[11]の方法では、手掛かり語を人手で与えた後、大量のテキストを用いて特徴語を学習させる方法が紹介されている。しかし、彼らの方法はあらゆるブログを限られた種類のカテゴリに分類するのに対し、本稿では除去すべき不要なページにどのような内容のテキストが出現するのか一般的な予測が困難である。また、本稿では特定の固有名詞に関するブログを扱うため、大量のテキストを用いる方法は難しい。本稿の目的は汎用的な分類器を作成することではなく、特定の固有名詞に関する必要なページを効率的に収集することである。したがって、もっと手軽に、また、支配的なケースを場当たり的手法で解決することが洗練された実装に拘り続けるよりも良い結果になる[12]と考え、本稿では、大量のテキストに代えて人間の予備知識を活用し、簡単にフィルタリングを実現することを目指す。

3. 収集の基本方針

(1) 収集対象

本稿では、少量の手作業で簡単に収集することを狙いとし、検索エンジンを利用し、その中から必要なものを振り分ける方法をとる。

ブログの話題は多岐にわたるため、関西空港に関する記述であっても、利用行動だけでなく、経営的・財務的・政治的な視点からの記述も多く存在する。したがって、検索エンジンによるキーワード検索の結果だけでは不要なページが多く混入する。そのようなページを除去せずにマイニングを行うと、不適切な語が多く混入することが予想される。したがって、まず、利用行動を記述したページをフィルタリングする必要がある。

本稿で収集対象とする（正例）ブログページは次のようなものである。

・関西空港の利用行動であること。
・利用者目線の立場の記述であること。

実際に気持ちや声が書いてあるかどうかは次の段階の課題であり、ここでは問わないこととし、利用行動の記述であれば収集することとした。

また、除外すべき（負例）ページは次のようなものである。一過性のイベント等は、特殊な状況下での特殊な気持ちが抽出される恐れがあるので、除外することとした。

・時事ネタ、ニュースの転載、正規のニュース配信
・「今日は○○の日」の転載
・イベント（参加側、出演側）
・グルメネタ
・鉄道ネタ

・何らかの告知、企業のＰＲ

(2) 語句パターン

次に、正例と負例を振り分ける方法について述べる。正例と負例を眺めると、使われている文言がそれらの間で大きく異なることがわかる。たとえば、交通行動に関する記述は「関西空港へ」「関西空港から」「○○へ行った」のような文言を含みやすく、一方、時事問題に関する記述は「関西空港会社は」「関西空港は○○と発表した」のような文言を含みやすい。人間がブログページを見た場合、そのような語句の特徴を見て、利用行動の記述であるか、時事的な視点の記述であるか等を、一目見て判断できると思われる。

そこで、このような特徴的な語句パターンを含むか否かによって正例・負例の判別ができると考えた。すなわち、正例ページに特徴的な語句パターン（正例パターン）と、負例ページに特徴的な語句パターン（負例パターン）を用意し、正例パターンを多く含むページは収集対象、負例パターンを多く含むページは除去対象とすれば、効率的に振り分けることができる。

多くの語句の中からそれぞれに特徴的な語句を拾い出す作業は、既存の研究例では、大量の文書集合を用いて体系的に行っている場合もある。しかし、本稿では、少数のサンプルを用意して人間が見れば、それに含まれる語句が交通行動に特徴的な語句かどうかを容易に判断できることから、大量の文書集合を用いるまでもなく、人間の持っている経験的知識を活用し、簡単に人手でリストアップすることとする。

(3) 尤度

単純に人手で語句パターンをリストアップするだけだと、予期していない文言が記述される可能性があるため、正負判別の手がかりとしては不適切な語句パターンになっている場合がある。そこで、各語句パターンがどの程度有効に働くかを見極めるため、学習セットを用意して、語句パターンの尤度を算出する。学習セットとは、判別対象ページの一部をランダムに抽出し、正解となる正例ページ・負例ページの別を人手で与えたものである。語句パターンiの尤度L_iは、決定リストの尤度[13]を参考にして、次のように定めた。

$$L_i = \log \frac{P(i|true) + \alpha}{P(i|false) + \alpha} \quad (1)$$

$$P(i|true) = \frac{N_{i,true}}{N_{true}} \quad (2)$$

$$P(i|false) = \frac{N_{i,false}}{N_{false}} \quad (3)$$

$P(i|true)$は語句パターンiが正例ページに登場する確率、$P(i|false)$は負例ページに登場する確率である。N_{true}は学習セットの正例ページ総数、N_{false}は負例ページ総数である。$N_{i,true}$、$N_{i,false}$はそれぞれ語句パターンiを含む学習セットの正例ページ数、負例ページ数である。したがって、語句パターンが正例でよく使われるほどL_iは＋になり、負例でよく使われるほどーになる。また、絶対値が0に近いものは判別の手がかりにはならないことを意味する。これらの式は、学習セットの正負のページ数の偏りに影響を受けないように定めた。また、全く登場しない語句パターンの場合、尤度は0となる。

αは経験的に定める緩和パラメータで、本稿では0.01とした。これは$N_{i,true}$または$N_{i,false}$が0または小さな数の場合にL_iが±∞または極端に絶対値の大きな数にならないようにするためのものである。たとえば、ある語句パターンが正例にのみ1回だけ登場した場合、それが判別の手がかりとしてどれだけ確からしいかを考慮して定める。この場合、αが0だとLは∞となり、確実に正例である手がかりとして扱うことになる。また、αが大きな値だとLは0に近くなり、あまり手がかりにはならないとして扱うことになる。

(4) ブログページの正負分類

各ブログページについて、出現する語句パターンの尤度を合計したものをそのブログページの得点とする。ここで、正例パターンの＋得点と負例パターンのー得点を別々に合計し、２軸上にプロットするようにした。すなわち、収集対象の話題について記述している場合は＋が蓄積され、除外対象の話題について記述している場合はーが蓄積されることになり、ブログに書かれている話題の様子を良く表すことができる。

各ブログページが収集対象かどうかを判定する問題は、この２軸平面上に正負判定用の境界線を、学習セットを用いて定める問題となる。つまり、データの分類問題となり、自然言語処理で広く用いられている手法を利用することができる。本稿では、サポートベクトルマシンを用いた。サポートベクトルマシンとは、空間を超平面で分割することにより２つの分類からなるデータを分類する手法である[14]。

4. 収集の実際

(1) 検索エンジン

収集手順は次の通りである。まず、検索エンジンでキーワード検索を行い、ブログページのリストを取得し、各ブログページのダウンロードを実行する。本稿では、Google blog検索を用い、"関空" "関西空港" "関西国際

空港"のいずれかのキーワードを含み、投稿日の範囲が2009年7月1日～9月30日のものを1000件検索した。

(2) 学習セットの用意

次に、1000ページの中から200ページを取り出して学習セットとした。学習セットには正例か負例かを人手で与え、それと同時に、その中の文章を見て語句パターンのリストアップを行った。ここで与えた正例・負例の判別情報は、語句パターンの尤度の計算、および、サポートベクトルマシンの学習に用いる。

ただし、特定の文字列を含むURLは除外し（/news/等）、また、文字化けがあるページも除外した。たとえば、UTF-8の１文字のバイト並びの途中でちょん切っている場合が見受けられた。

なお、正例か負例かをはっきり判別しがたい場合（行ってはいるが写真紹介がメインのもの等）や、文章がほとんどないものはグレーゾーンとし、学習やテストの対象外とした。そのようなページは、手がかりが乏しく、判別が困難なのに対し、誤ったマイニング結果が抽出される原因となる恐れは低いと言える。

また、その際、学習セットを参照して、本文エリアを切り出すロジックを作成し、本文のみを用いるようにした。これは、リンク一覧等の部分に、当該ページの話題とは直接関係のない語句が記述されている場合が多いためである。たとえば、Webページ中のHTMLのコメントとして「記事本文ここまで」という文字列があれば、それ以降は切り落とす処理を行う。

(3) 語句パターン

語句パターンは、学習セットのページを参照し、正負それぞれに特徴的なものを人手で列挙して正規表現として整理した。たとえば、着陸の動作だけを拾いたい場合、「着陸」だけだと「着陸料」を拾うので、「着陸(。|し|する|で|なの)」と工夫する。多くの研究例では、体系的に行う場合、形態素と呼ばれる、語の最小単位に解析する処理を行うことが多いが、本稿の方法では、形態素解析をせずに簡単に済ませるとともに、人間の予備知識・ノウハウを活用する。

なお、後の「生の声」の抽出時に偏りが生じないよう、「声」そのものを表す語句はなるべく避け、行動や施設、言い回し等の語句を主に用いた。また、重複（部分文字列となる場合）をできるだけなくすように配慮したが、共起の問題を考慮すると複雑になるため、厳密でなくてよいとした。

次に、列挙した語句パターン候補の尤度を学習セットで算出する。実際の語句パターンを**表-1**・**表-2**に示す。「行く」「(に到着$|に到着。)」「乗り継」といった、幅広い行動を表す語句は行動を表す特徴的な語句になると予想して列挙していたが、実際にはあまり役に立たなかった。

表-1　正例パターン（尤度順上位のみ）

語句パターン	正例数	負例数	尤度
チェックイン	7	0	2.64
搭乗(ゲート|口)	6	0	2.49
(出発。|出発$)	6	0	2.49
無事	10	1	2.24
飛行機に乗	9	1	2.14
搭乗(。|し|する|で|なの)	4	0	2.13
感じです	4	0	2.13
リムジン	4	0	2.13
ラウンジ	8	1	2.03
バス(に乗|を待)	8	1	2.03

表-2　負例パターン（尤度順上位のみ）

語句パターン	正例数	負例数	尤度
ステージ	0	10	-2.46
撤退	0	9	-2.37
ニュース	0	9	-2.37
需要	0	8	-2.26
参加(した|しました|し、|してき|でき)	0	8	-2.26
[^ド](ライヴ|ライブ)	0	8	-2.26
(Festival|フェスティバル)	0	8	-2.26
発表(した|され|する|で)	0	7	-2.14
知事	0	7	-2.14
赤字	0	7	-2.14
周年	0	7	-2.14
されている。	0	7	-2.14

(4) サポートベクトルマシン

学習セットのページの得点分布を**図-1**に示す。図を見ると、左上の原点付近から斜め右下方向へ境界線を引けば適切であることがわかる。サポートベクトルマシンはTinySVM[15]を用いた。

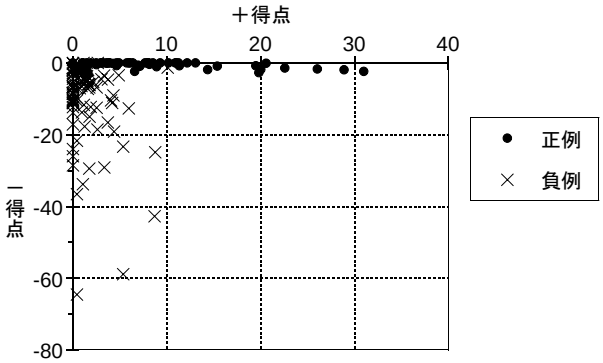

図-1　学習セットのページの得点分布

学習を行ったサポートベクトルマシンを用い、最終的に、923ページに対して判別を実行し、プレーンテキストのサイズ（本文エリアのみ）で約1.8MBのうち、664KBを収集対象の正例と判別した。

(5) 性能評価テスト

以上で得られた語句パターン・尤度・サポートベクトルマシンを用いて、どの程度の性能で判別できるかをテストした。学習セットとは別にテストセットとして100ページを用意し、人手で正解の正/負/グレーを与えた。そして、実際に自動判別を行い、再現率・適合率を調べた。その結果を**表-3**に示す。再現率(4)は高いほど漏れが少ないことを意味し、適合率(5)は高いほどゴミの混入が少ないことを意味する。通常、両者はトレードオフの関係となる。正解率とは、判別結果の正負が人手で与えた正解の正負と一致しているかどうかを表す。

$$\text{再現率} = \frac{\text{正しく正例と判定されたページ数}}{\text{正例ページ総数}} \quad (4)$$

$$\text{適合率} = \frac{\text{正しく正例と判定されたページ数}}{\text{正例と判定されたページ総数}} \quad (5)$$

表によると、おおむね良い結果が得られていることがわかる。

表-3 テスト結果

評価項目	値
再現率	0.838
適合率	0.861
正解率	0.866

(6) 失敗の原因の分析

次に、失敗の原因について分析する。失敗には(i)正例なのに取りこぼした場合、(ii)負例なのに誤って抽出してしまった場合の２通りある。

まず、(i)正例なのに取りこぼした場合であるが、利用行動の記述ではあるが、時事的な話題にも触れている場合があった。本稿ではページ単位の収集を行っており、話題の転換点で分割することは行っていない。このような正負両方の要素を含むページの扱いは今後の検討課題である。また、途中で話題の変わるページは通常長文であり、**図-1**の原点から遠く離れた右下方向にプロットされるはずであるが、学習セットにはこのような長文ページは存在しておらず、適切な境界線が引かれなかった点も一因である。

また、ブログの「ランキングに参加しました」を拾ってしまい、イベント等に「参加しました」と見なしてしまった場合があった。学習セットではランキングに「参加しています」という表現しか登場しなかったため、学習不足と言える。

また、(ii)負例なのに誤って抽出してしまった場合には、関空快速の利用行動であるが関空そのものの利用ではないケース、旅行関係の企業のＰＲページで「チェックイン」「ツアー」等の語句を多く含むケース、イベント（出演側）の記述で利用客っぽい文体であり、手がかりとなる負例パターンが少ないケースがあった。

5. 結論と今後の課題

本稿では、関西空港の利用者の「生の声」を取り出すための、利用行動に関するブログページの収集方法について述べた。簡単に実現することを目指し、既存の検索エンジンを活用し、その検索結果の中から必要なページを取り出す方法を採用した。その方法として、正例・負例に特徴的な語句パターンを用意し、学習セットを用いてそれらの尤度を計算し、ブログページの得点を定義し、サポートベクトルマシンを用いて収集対象のページかどうかを判別する方法について述べた。その際、全て自動で行おうとせず、少量の人手を活用する方法を採用した。テストの結果、ほどよい結果が得られた。

本稿の方法では、一旦語句パターンを用意して学習をさせれば、大量の検索結果の中から必要ページの収集を自動的に行うことができる。また、尤度を適切に設定すれば、かな漢字変換のユーザ登録単語のように簡単に語句パターンの補充をすることが可能であり、柔軟性の高い方法と言える。

一方、本稿では学習セットとテストセットで同じ時期のデータを用いたが、ある時期に突発的な事象が発生した場合、ブログに登場する語句パターンの傾向が突然変化する可能性もある。したがって、異なる時期のページを収集する場合の再学習の必要性の有無については残された検討課題である。

また、本稿では本文エリアのみを用いたが、本文エリアの切り出しは相当面倒であり、その効率的な方法は今後の課題である。

次の課題は、収集したブログページから実際に利用者の「生の声」をマイニングすることであるが、既存のツールは通常、きちんとしたテキストを用いて作成されており、ブログのくだけたテキストに関しては性能が良くないことがわかっている。これについては、稿を改めて報告する予定である。

参考文献

1) Takao, K. and Asakura, Y. : Extraction of cognition

results of travel routes from open-ended questionnaire texts, *Journal of the Eastern Asia Society for Transportation Studies*, Vol.6, pp.1943-1955, 2005

2) Takao, K. and Asakura, Y. : Description of route choice behaviour conforming to elimination-by-aspects by extracting choice strategy with words, *Conference CD-ROM of 11th International Conference on Travel Behaviour Research* (IATBR 2006), Kyoto, Japan, "Kazutaka Takao.pdf", 2006.
3) 鷹尾和享, 朝倉康夫：選択肢が選択または排除されるきっかけの理由をElimination-By-Aspectsで捉える, 自然言語処理, Vol.14, No.3, pp.61-80, 2007.
4) 奥村学：blogを対象とした言語処理とその応用 ― 現在, 未来 ―, 言語処理学会第13回年次大会ワークショップ「大規模Web研究基盤上での自然言語処理・情報検索研究」論文集, W2-42.pdf, 2007.
5) http://internet.watch.impress.co.jp/cda/event/2005/08/02/8657.html（2011年3月7日参照）.
6) Zhang, J. : Subjective well-being and activity-travel behavior analysis - Applying day reconstruction method to explore affective experience during travel -, *Proc. of the 14th HKSTS International Conference,* Vol.2, pp. 439-449, 2009.
7) 南野朋之, 鈴木泰裕, 藤木稔明, 奥村学：blogの自動収集と監視, 人工知能学会論文誌, Vol.19, No.6, pp. 511-520, 2004.
8) 赤峯享, 加藤義清, 河原大輔, レオン末松豊インティ, 新里圭司, 乾健太郎, 黒橋禎夫, 木俵豊：Web情報分析のための大規模Webページの収集・選択・検索, 言語処理学会第16回年次大会発表論文集, D2-2.pdf, 2010.
9) 廣嶋伸章, 山田節夫, 奥雅博： 概念ベースを用いたWebページからの評価項目の自動抽出, 言語処理学会第11回年次大会発表論文集, C2-4.pdf, 2005.
10) 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一： 意見抽出のための評価表現の収集, 自然言語処理, Vol.12, No.3, pp.203-222, 2005.
11) 橋本力, 黒橋禎夫：基本語ドメイン辞書の構築と未知語ドメイン推定を用いたブログ自動分類法への応用, 自然言語処理, Vol.15, No.5, pp.73-97, 2008.
12) 伊藤直也：はてなで利用している言語処理技術, 言語処理学会第16回年次大会 チュートリアル資料, T1-c.pdf, 2010.
13) Yarowsky, D. : Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French, *Proc. of the 32nd Annual Meeting of ACL*, pp.88-95, 1994.
14) 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均：SENSEVAL2J辞書タスクでのCRLの取り組み―日本語単語の多義性解消における種々の機械学習手法と素性の比較―, 自然言語処理, Vol.10, No.3, pp.115-133, 2003.
15) http://chasen.org/~taku/software/TinySVM/（2011年7月6日参照）.