関西空港の利用行動に関する利用者の立場のブログページの簡単な収集方法

鷹尾 和享
システム科学研究所

1　はじめに

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」[1]が含まれていることが期待できる。

筆者らは交通行動の心理的な側面に目を向け、経路選択プロセスをことばによって分析する研究を行ってきた[2]。交通行動に関しては、物理的に観測可能な要素については盛んに研究が行われているのに対し、心理的な側面については必ずしも十分とは言えない。

そこで、筆者は、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。

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。また、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。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簡単に必要ページを収集する方法について報告する。

2　関連研究

ブログページのフィルタリングに関する研究例として、橋本ら(2008)[3]や河野ら(2008)[4]の研究例が挙げられる。彼らの方法では、手掛かり語を人手で与えることが紹介されているが、大量のテキストを用いて IDF 値を算出する方法をとっているのに対し、本稿では人間の予備知識を活用してもっと簡単に実現することを目指す。

また、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)[5]が挙げられる。しかし、この方法は予め用意した尺度で分析するのに対し、本稿ではブログから抽出することにより、今まで予想していなかった未知の要素も捉えることが可能であることが期待できる。

3　収集の基本方針

3.1 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。前述のように、検索エンジンの結果だけでは不要なページが多く混入する。そのうち、本稿で収集対象とするブログページ（正例）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。

実際に印象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、除外すべきページ（負例）は次のようなものである。一過性のイベント等は、誤った印象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

3.2 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。正例のページと負例のページには使われている語句のパターンに特徴があると見込まれる。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、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、後は自動的に振り分ける方針を採用する。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、形態素解析を行ったうえで、大量の文書集合を用いて IDF を算出する場合もあるが、本稿では、人間の持っている予備知識を活用し、簡単に人手でリストアップすることとする。

3.3 尤度

用意した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、各語句パターンの尤度を算出する。尤度は決定リストの尤度[6]を参考にして、次のように定めた。

$$L = log \frac{N_{true} + \alpha}{N_{false} + \alpha}, \quad \alpha = 1.0$$

N_{true} は当該語句パターンを含む正例ページ数、N_{false} は負例ページ数である。したがって、語句パターンが正例に偏って出現するほど＋になり、負例に偏って出現するほどーになる。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。

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し、所定の閾値より得点が高いものを収集対象とする。閾値は学習セットを用いて見極める。

4　収集の実際

4.1 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。

4.2 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、正例か負例かを人手で与えた。学習セットは語句パターンの尤度の計算、および、得点閾値の判断に用いる。ただし、その際、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、誤った印象が抽出される危険性は低いと言える。

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簡単なロジックを作成し、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。

4.3 語句パターン

次に、学習セットのページを参照し、正負それぞれに特徴的な語句パターンを人手で列挙する。語句パターンは正規表現として整理した。この方法では、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ことができる。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。

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避け、行動や施設、言い回し等の語句を用いた。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	2.08
搭乗(ゲート|口)	6	0	1.95
(出発。|出発$)	6	0	1.95
飛行機に乗	9	1	1.61
搭乗(。|し|する|で|なの)	4	0	1.61
感じです	4	0	1.61
リムジン	4	0	1.61
バス(に乗|を待)	8	1	1.50

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
ニュース	0	11	-2.48
ステージ	0	10	-2.40
撤退	0	9	-2.30
廃止	1	17	-2.20
需要	0	8	-2.20
参加(した|しました|し、|してき|でき)	0	8	-2.20
[^ド](ライヴ|ライブ)	0	8	-2.20
(Festival|フェスティバル)	0	8	-2.20

4.4 閾値

次に、収集対象とみなすのに必要な得点（尤度の合計）の閾値を、学習セットを用いて見極める。正例パターンを多く含むページほど、また、負例パターンが少ないほど得点が高くなるので、閾値を高く設定するほど、確実に利用行動と言えるページだけが収集されることになる。

図１は得点の閾値と再現率・適合率の関係をグラフで示したものである。閾値を-1.33 に定めると、再現率がおおむね 95%、適合率がおおむね 90%となる。また、閾値を-0.23 に定めると、再現率がおおむね 90%、適合率がおおむね 95%となる。

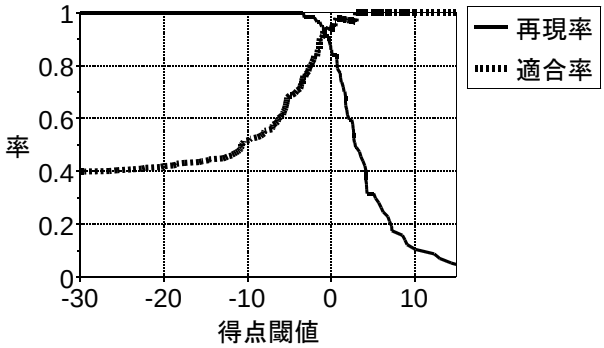

図１：閾値と再現率・適合率

５　評価テスト

5.1 テスト結果

以上で得られた語句パターン・尤度・閾値を用いて、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。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。

表によると、おおむねほどよい結果が得られていることがわかる。閾値-1.33 では再現率が 0.9 を超えているのに対し、適合率は 0.810 であり、混入した２割程度の負例を人手で除去する使い方ができると思われる。また、閾値-0.23 なら再現率・適合率ともおおよそ 0.9 弱であり、自動判別だけで済ませてしまう使い方が期待できる。

表３：テスト結果

評価項目	閾値-1.33 の場合	閾値-0.23 の場合
再現率	0.919	0.865
適合率	0.810	0.889
正解率	0.866	0.890

5.2 失敗の原因の分析

ここでは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、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。

６　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、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、再現率・適合率とも、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すれば、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。なお、本稿の方法を本文でないエリアを切り落とさずにページ全体で行うと、１割程度成績が下がった。

筆者の次の課題は、収集したブログページから実際に利用者の「生の声」を抽出することである。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］奥村学（2007）: 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.

［2］鷹尾和享, 朝倉康夫（2007）: 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.

［3］橋本力, 黒橋禎夫（2008）: 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.

［4］河野洋志, 柴田知秀, 黒橋禎夫（2008）: ブログ記事の商品カテゴリへの自動マッピング,言語処理学会第 14 回年次大会発表論文集 , A4-5.

［5］Junyi Zhang（2009）: Subjective well-being and activity-travel behavior analysis - Applying day reconstruction method to explore affective experience during travel -, in *Proceedings of the 14th HKSTS International Conference*, Vol.2, pp. 439-449.

［6］David Yarowsky（1994）: Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French, in *Proceedings of the 32nd Annual Meeting of ACL*, pp.88-95.