観光開発のヒントをブログ記事から得るための支援技術 ~能動学習を用いる場合~

謝花 博[*1]　徳久 雅人[*2]　村田 真樹[*2]　村上 仁一[*2]
[*1]鳥取大学 工学部 知能情報工学科
[*2]鳥取大学 大学院 工学研究科 情報エレクトロニクス専攻
{s082025, tokuhisa, murata, murakami}@ike.tottori-u.ac.jp

1　はじめに

観光地開発のヒントを得るために，ブログ記事を分析する研究が行われている [1]．しかし，ブログ記事の全てが観光開発のヒントとなるわけではないため，分析者の負担を軽減するためにブログ文からヒントとなる文を機械的に抽出できることが望まれる．

その抽出方法の 1 つとして SVM(Support Vector Machine) を用いる方法がある [2]．しかし，抽出された文集合におけるヒントの含有率をさらに高めることが課題となっている．ここで，ブログ記事のヒント分析を進めると，自然に正例と負例が得られるので，これを SVM の学習データに追加して再学習し，残りの分析対象を再分類するという手法が対策として考えられる．

そこで本稿では，能動学習の手法を用いることにより分析精度を向上させ，分析者の負担を軽減させることを目的とする．

2　ヒント分析の概要

2.1　ヒントを得るとは

本稿におけるヒント分析とは，分析者がある観光地 A の開発案を考えるために観光地 B に関するブログを分析することである．これにより新しい発想を得ようとしている．

例えば,「山陰海岸」の観光開発を行う時に，類似の観光地である「三陸海岸」に関するブログを分析するとしよう．その結果「遊歩道から断崖絶壁を登った」という文があった場合，三陸海岸では遊歩道を整備することで観光客の満足度を高めることができたと解釈される．こうした良い開発を山陰海岸においても行うべきだという発想が生まれる．

発想を生んだ文は開発のヒントとなった文である．以降では，単にヒント文と呼ぶことにする．

2.2　分析支援とは

本稿における分析支援とは，このような観光開発の発案に繋がる文（ヒント文）を自動抽出するということである．具体的には，ある程度のブログ文を抽出し，その中から観光開発のヒントである文とそうでない文を自動的に分類する．その中からヒントであると推測される文を分析者に提示することで，ヒントではないと思われる文，すなわち読む必要のない文を削減する．こうして分析者が分析する文の量を減らし，負担を軽減することができる．

3　ヒント文の自動抽出の手法

3.1　基本的な手法

まず，ある程度の量の観光ブログ文書を用意する．その各文に対し，人手でヒント文か否かを判定し，それを SVM の学習データとする．次に，分析すべきブログ文をテストデータとして SVM による分類を行うことで各文がヒントとなるかどうかの判定を行う．ここで，学習データおよびテストデータの素性は，記号，名詞，動詞，形容詞，形容動詞，副詞，接続詞，感動詞，接辞，助詞，BM25 による特徴度区間ラベル，および，情緒推定による情緒とする [2]．最後に，SVM による分類結果からいくらかを分析者に提示する．ここまでが自動抽出である．その後，分析者は，提示された文を読みながらヒント分析を行う．

図 1 にこの手法による動作の図を示す．この図における 3 地域データとは江ノ島，三陸海岸，若狭湾のブログデータのことであり，学習データとする．新地域データとは糸魚川のブログデータのことであり，テストデータとする．これらのデータについての詳細は 4.1 節で述べる．また，クラスとは「ヒント文 (+1)」と「非ヒント文 (-1)」の 2 値のことである．スコアとは，SVM による分類で算出される値である．

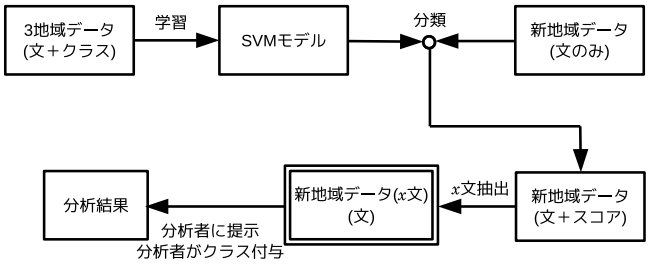

図 1 基本的な手法
※二重四角が本手法の出力である．二重四角の後はヒント分析の過程である．

3.2 能動学習を用いた手法

図 2 に能動学習を用いた手法による動作の図を示す．まず，基本的な手法と同様に学習および分類を行う．次に，スコアの降順で抽出した文（x 文）に対し，その文がヒントであるかどうかの判別を分析者が行う (図 2 の (a))．その結果を元の学習データに追加して再学習を行う．その後，残りの文を再分類し，再分類結果により抽出した文（y 文）の分析を行う (図 2 の (b))．

ここで，再学習のために抽出する手法は幾通りか考えられる．例えば，[3] ではスコアの絶対値が小さいものを優先的に抽出していた．しかし，その手法では，ヒントになりにくい文を分析者に提示することになる．本稿では，観光開発のヒントを得るための分析を主としており，能動学習は，その分析作業の副産物として機能するものとしたい．したがって，本稿では，スコアの高いものから順に抽出するという手法を選択する．

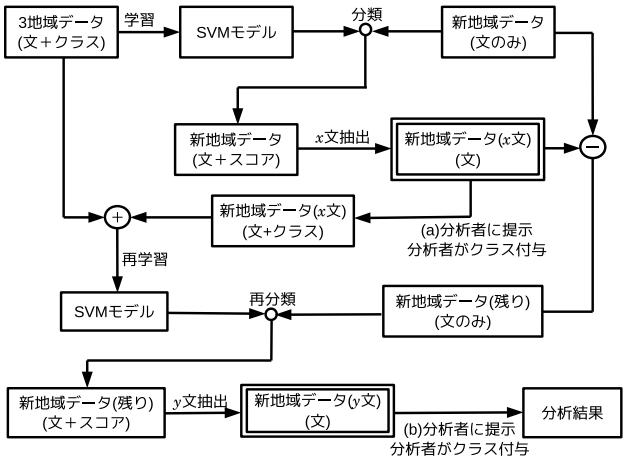

図 2 能動学習を用いた手法
※二重四角が本手法の出力である．二重四角の後はヒント分析の過程である．分析者には $x+y$[文] が提示されている．

4 実験条件

この節では，ヒント分析支援の評価実験を行う．以下に実験条件を示す．

4.1 使用するデータ

実験には以下のデータを使用する．

- **3 地域データ:江ノ島，三陸海岸，若狭湾の観光ブログデータ**
  このデータは [1],[2] で使用したものであり，既に人手によるヒントの有無の判別およびヒントのカテゴリ（後述）の付与が完了している．実験ではこのデータを学習データとして使用する．

  このデータは Yahoo!ブログの「旅行」の項目に登録されたブログから,「江ノ島海岸」,「三陸海岸」,「若狭湾」をそれぞれ検索キーとして記事を検索して得られた 444 記事，12,044 文である．検索は 2010 年 7 月 16 日に行われた．

- **新地域データ:糸魚川の観光ブログデータ**
  このデータは実験を行うにあたって新しく用意したデータである．実験の正解データを作成するためにまずこのデータに人手でヒントの有無を付与する．さらに，ヒントであるものにはヒントのカテゴリ（後述）を付与する．実験ではこのデータをテストデータとして使用する．

  このデータは Yahoo!ブログの「旅行」の項目で「糸魚川 観光」という検索キーで得られた 95 記事，3,222 文である．検索は 2011 年 10 月 19 日に行われた．

以下にこれらのブログデータの一部を示す．このデータは ID 番号，ヒント文 (+1) か非ヒント文 (-1) のクラス, ヒントのカテゴリ，および，文で構成する．

ブログデータの例

E00005/-1/ヒントなし/江ノ島海岸をひだりに見ながら江ノ島弁天橋を渡り江ノ島に入ると両側を土産物屋に挟まれた江島神社参堂に入るのだが、人々々 … 。
E00006/-1/ヒントなし/老若男女ものすごい人手だ。
E00007/+1/神社仏閣/朱の鳥居を超え階段を登り参拝、江ノ島大師、奥津宮を経て島の南端、稚児ヶ淵に到達。
E00008/+1/自然散策/岩屋洞窟を見学の後来た道を戻った。

ヒントカテゴリとは「自然散策」,「動植物」,「文化歴史」,「神社仏閣」,「街並み」,「施設」,「温泉」,「飲食」,「買い物」,「行事」,「交通」,「スポーツ・アウトドア」,「釣り」,「音楽」,「交流」,「産業」,「その他」の 17 分類のことである．

4.2 実験上でのヒント分析の手法

観光開発に向けて分析者がブログを分析する実験上の手法は，次の 3 通りが設定できる．

- **比較手法 1**

  全ての文を分析者が分析する手法とする．すなわち自動抽出がない手法である．

- **比較手法 2**

  基本的な手法（3.1 節）を用いてヒントの可能性の高い文から順番に $n\%$ の文を提示し，分析者が分析を行う手法とする．なお，スコアが負値となっても分析者に提示することができる．

- **提案手法**

  能動学習を用いた手法（3.2 節）を用いてヒントの可能性の高い文から順番に分析者が分析を行う手法であり，分析者は再学習前にテストデータの内 $m\%$ の文を分析し，再学習後にテストデータの内の $n\%$ を分析することとする[1]．

4.3 評価基準

通常の評価基準にならい，適合率 P，再現率 R，および F 値を使用する．

ここで，ヒント文の自動抽出においては，分析者に必ずしも全てのヒント文を提示する必要はない．たとえば,「遊歩道の整備」というアイデアは 1 度得られれば十分であり，同じ開発案を発想させるヒント文は何度も自動抽出で提示される必要はない．

そこで，カテゴリ再現率 R_θ という評価基準がある [1]．これは，ヒント文の網羅性を評価する代わりに，ヒントカテゴリの網羅性を評価することで，実践的な評価に近づけるものである．ヒントカテゴリに属する文のうちの一定割合 θ 以上が自動抽出により提示されれば良しとする評価基準である．ただし，同一の発想かどうかまでを評価するのではなく，同一のヒントカテゴリであるかどうかを考慮するという近似的な評価である．また，F 値に相当する評価基準として，適合率 P と R_θ の調和平均である F_θ（カテゴリ F 値と呼ぶことにする）が考えられる．

以上より，本稿では，R_θ および F_θ も使用する．以下に，各評価基準を求める式を示す．

$$P = \frac{|O \cap A|}{|O|}$$

$$R = \frac{|O \cap A|}{|A|}$$

$$F = \frac{2PR}{P+R}$$

$$R_\theta = \frac{1}{|C|}\sum_{c \in C} f(O, A_c; \theta)$$

$$f(O, A_c; \theta) = \begin{cases} 1 & (if \quad |O \cap A_c| > \theta \cdot |A_c|) \\ 0 & (otherwise) \end{cases}$$

$$F_\theta = \frac{2PR_\theta}{P+R_\theta}$$

ここで，$|X|$ は集合 X の要素数，C はヒントカテゴリの集合，O は分析者に提示された文の集合，A は分析者に提示されるべき文（正解文）の集合，A_c はヒントカテゴリ c に対応する正解文の集合をそれぞれ表す．

5 実験結果

提案手法では，再学習のために分析者に提示する文の数（図 2 における x）および再分類後に分析者に提示する文の数（図 2 における y）が定められていない．本実験では，これらのパラメータの設定値を変更しながら，評価値を観測する．

観測した評価値を表 1〜5 に示す．パラメータ m は，新地域ブログの総文数に対する割合であり，再学習のために提示する文数の比率である ($x = m \cdot$ 総文数)．同じく n は，総文数に対する割合であり，再分類後に提示する文数の比率である ($y = n \cdot$ 総文数)．

5.1 表の読み方

比較手法 1 は，全ての文を分析者に提示する手法なので，$m = 0\%, n = 100\%$ の欄から評価値を読み取る．比較手法 2 は，再学習が無いので，$m = 0\%$ の行において，n の設定値ごとの評価値を表から読み取る．提案手法は，ある程度の再学習を経るので，$m > 0\%$ の行において，n の設定値ごとの評価値を表から読み取る．

適合率によると，分析者が無駄なくヒント文を読むことができたかどうかが分かる．比較手法 1 では，0.5 なので約半分がヒント文であった．総文数の 30%を提示する条件下では，比較手法 2 では，0.6 であり[2]，提案手法では 0.66 と 0.67 であった[3]．

カテゴリ再現率によると，分析者が新たな発想に至る文を読んだかどうかが分かる．たとえば，カテゴリ再現率は，$m = 0\%, n = 20\%$ の欄において 0.76 であるが，$m = 10\%, n = 10\%$ の欄において 0.71 である．総文数の 20% を提示したとしても，前者の方が幅広い発想をしたと言える．

[1] m, n はテストデータ総文数を分母とする．

[2] $m = 0\%, n = 30\%$ の欄

[3] $m = 10\%, n = 20\%$ の欄と $m = 20\%, n = 10\%$ の欄

5.2 一般の評価

以下に適合率 P，再現率 R，F 値を求めた結果を示す．

表 1 適合率 P

m\n	10	20	30	40	50	60	70	80	90	100
0	0.71	0.63	0.60	0.56	0.53	0.51	0.50	0.50	0.50	0.50
10	0.70	0.66	0.53	0.59	0.55	0.53	0.51	0.51	0.50	
20	0.67	0.65	0.62	0.59	0.55	0.53	0.51	0.50		
30	0.65	0.53	0.60	0.57	0.54	0.52	0.50			
40	0.60	0.60	0.57	0.54	0.52	0.50				
50	0.57	0.57	0.54	0.52	0.50					
60	0.55	0.54	0.52	0.50						
70	0.53	0.52	0.50							
80	0.51	0.50								
90	0.50									

表 2 再現率 R

m\n	10	20	30	40	50	60	70	80	90	100
0	0.14	0.25	0.36	0.45	0.65	0.53	0.71	0.80	0.90	1.00
10	0.28	0.40	0.50	0.60	0.67	0.74	0.83	0.91	1.00	
20	0.40	0.52	0.62	0.71	0.77	0.85	0.93	1.00		
30	0.52	0.63	0.73	0.80	0.86	0.93	1.00			
40	0.61	0.72	0.80	0.87	0.94	1.00				
50	0.69	0.80	0.87	0.94	1.00					
60	0.77	0.87	0.94	1.00						
70	0.85	0.94	1.00							
80	0.93	1.00								
90	1.00									

表 3 F 値

m\n	10	20	30	40	50	60	70	80	90	100
0	0.24	0.36	0.45	0.50	0.53	0.56	0.59	0.62	0.64	0.67
10	0.40	0.50	0.55	0.59	0.61	0.62	0.63	0.65	0.67	
20	0.50	0.58	0.62	0.64	0.64	0.65	0.66	0.67		
30	0.58	0.63	0.66	0.67	0.66	0.66	0.67			
40	0.60	0.54	0.67	0.67	0.67	0.67				
50	0.63	0.66	0.67	0.67	0.67					
60	0.64	0.67	0.67	0.67						
70	0.56	0.67	0.67							
80	0.66	0.67								
90	0.67									

表 3 より，F 値で比較を行うと $m = 30\%, n = 40\%$ もしくは $m = 40\%, n = 30\%$ とした場合が最も性能がよく，かつ文の分析量が最も少なくなる組み合わせであることが分かる．

比較手法 1 と比較すると同じ性能で分析量を 30%削減しており，比較手法 2 で同じ量だけ分析を行った場合 ($m = 0\%, n = 70\%$) と比較すると性能が F 値で 0.08 向上していることが分かる．

しかしながら，これでは分析量が多く，同じような内容の文ばかり抽出されている可能性もあるため，次にカテゴリ再現率を考慮した評価を行う．

5.3 カテゴリ再現率を用いた評価

以下にカテゴリ再現率 R_θ とカテゴリ F 値 F_θ を求めた結果を示す．閾値は $\theta = 0.2$ を使用する．

表 5 より，再学習による分析を行う場合は $m = 20\%, n = 20\%$ のとき，すなわち全体の 2 割を再学習前に分析し，もう 2 割を再学習後に分析するという手法が最も効率が良く，比較手法 2 で同じ量だけ分析を行った場合 ($m = 0\%, n = 40\%$) と比較すると性能はカテゴリ F 値で 0.09 上昇するということが分かった．

表 4 カテゴリ再現率 R_θ

m\n	10	20	30	40	50	60	70	80	90	100
0	0.24	0.76	0.94	0.94	1	1	1	1	1	1
10	0.71	0.94	1	1	1	1	1	1	1	
20	0.94	1	1	1	1	1	1	1		
30	1	1	1	1	1	1	1			
40	1	1	1	1	1	1				
50	1	1	1	1	1					
60	1	1	1	1						
70	1	1	1							
80	1	1								
90	1									

表 5 カテゴリ F 値 F_θ

m\n	10	20	30	40	50	60	70	80	90	100
0	0.36	0.69	0.73	0.70	0.70	0.68	0.67	0.67	0.66	0.67
10	0.70	0.78	0.77	0.75	0.71	0.69	0.68	0.67	0.67	
20	0.78	0.79	0.76	0.74	0.71	0.69	0.68	0.67		
30	0.78	0.77	0.75	0.73	0.70	0.68	0.67			
40	0.75	0.75	0.72	0.70	0.68	0.67				
50	0.73	0.72	0.70	0.68	0.67					
60	0.71	0.70	0.69	0.67						
70	0.69	0.69	0.67							
80	0.68	0.67								
90	0.67									

6 おわりに

本稿は，SVM を用いてブログ記事から観光開発のヒントを得る手法 [2] に能動学習の手法 [3] を取り入れることによって，分析性能を向上させる手法を提案した．この手法により，能動学習を使用しない手法と比較して F 値で 0.08，カテゴリ F 値で 0.09 分析性能が向上するということが分かった．また，分析量が同じ場合，能動学習を使用する時は m, n の値をどのように設定しても能動学習を使用しない場合と比較して性能が向上するということも分かった．

しかしながら，実験により求めた $m = 20\%, n = 20\%$ という値は本実験コーパスに依存するものであるため，使用するデータが変わった場合，最も性能がよくなる m, n の値の組み合わせは変化すると考えられる．そのため，ヒントを分析する前や分析の最中に m, n の値を決定するような手法の考案が今後の課題として挙げられる．

謝辞 本研究は，科学研究費補助金（若手研究（B）：22700100）のもとで行いました．

参考文献

[1] 徳久雅人, 奥村秀人, 村田真樹: “観光開発のためのブログ記事からの評判分析”, 観光と情報, Vol.7, No.1, pp.85-98, 2011.

[2] 徳久雅人, 村田真樹: “観光開発のヒントをブログ記事から得るための支援技術～SVM を用いる場合～”, 第 8 回観光情報学会全国大会発表概要集, pp.44-45, 2011.

[3] 齋藤邦子, 今村賢治: “タグ信頼度に基づく半自動自己更新型固有表現抽出”, 自然言語処理, Vol.17, No.4, pp.3-21, 2010.