概要

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．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．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．

本研究では，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ことに注目した．まず通常の学習および分類を行い，ヒントの可能性がある文のうちいくらかを分析する．分析を行ったデータはヒントか否かの情報が得られ，この正例と負例のデータを SVM の学習データに追加して再学習し，残りの分析対象の再分類を行うという能動学習の手法を提案する．

本研究では SVM の学習データとして江ノ島，三陸海岸，若狭湾のブログデータ 12,044 文を用い，テストデータとして糸魚川のブログ文 3,222 文を用いた．これらのデータに対し SVM を使用せず全ての文の分析を行った場合，SVM による分類を 1 回のみ行った場合，SVM による分類を 2 回すなわち再学習を行った場合の 3 通りの手法に対し，性能を比較する実験を行った．

その結果，分析すべき文の量を削減しさらにヒントの含有率を高めることに成功した．これにより，能動学習を用いることでブログ記事の分析性能が向上することが確認された．しかしながら，再学習前と再学習後にどの程度の割合の文を分析すべきかを求めることはできていないため，今後の課題はこれらの割合を分析を行う前に求めることである．

目 次

図 目 次

表 目 次

第1章　はじめに

近年，都道府県等の地方公共団体において観光立県宣言がなされる例があるように，観光開発の重要性に対する認識が高まっている．観光開発とは，観光客が訪れることが少ない観光地の利用の促進のために旅行関係施設の配置や整備を行うことである．このような地域では，訪れる観光客が持つ親睦，休養，見物といった様々な目的を満たすことが可能で，かつその地域の特色を生かすことができるような適切な観光開発案が求められており，このような案は実際に旅行を行った観光客の旅日記から開発に関するヒントを発見することができると考えられる．

このような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自動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，この方法における自動抽出の精度を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．そこで本研究では，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．

本研究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．このような発想を生んだ文は開発のヒントとなった文である．

本研究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．本研究における観光開発のヒントとはこのような文であり，以降このような文をヒント文と呼ぶことにする．

第 2 章ではこれまでに行われた観光開発の研究およびこの研究に取り入れる能動学習に関する研究の説明を行う．第 3 章では，能動学習を観光開発に取り入れる手法の説明を行う．第 4 章では，分析対象となるブログデータの利用法に関する説明を行う．第 5 章では，実際にどのように自動抽出および性能評価を行うかの説明を行う．第 6 章では，従来手法と提案手法の評価結果の比較を行う．第 7 章では評価結果からどのように能動学習が有効かの考察を行う．第 7 章ではまとめを行う．

第2章　関連研究

2.1　観光開発案のに繋がるヒント文の自動抽出

2.1.1　特徴度と情緒推定の利用による観光地分析

徳久らは，観光開発の支援のために観光ブログから開発のヒントとなる文を抽出する手法の提案を行った．抽出は「観光地のブログ記事と一般のブログ記事を比較することによる観光地の特徴語」,「感情の原因・状態・表出を表す言語表現」という 2 点の要素を用いて行い，3 地域の観光ブログ文からヒントの抽出を行った．これによって閲覧すべき文の量を 3 分の 1 に減らし，かつヒントの含有率を 19%から 27%へと高め，観光開発の発想支援に繋がる一つの手法を示した [1].

2.1.2　SVM の利用による観光地分析

徳久らは，SVM を用いてブログ記事がヒントであるか否かを判定する手法を提案した．観光ブログのある程度の文に対してその文がヒントであるか否かの注釈を付与したものを学習データとし，残りの文に対して SVM を用いることで，ブログ記事の抽出結果におけるヒントの含有率を向上させた．ブログ記事は [1] と同様のデータを用い，2 地域を学習データ，1 地域をテストデータとして抽出を行ったところ，性能は F 値で 25.2 となった [2].

2.1.3　先行研究における自動抽出の手法

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．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．この分析結果がこの手法における出力となる．

図 2.1 にこの手法による動作の図を示す．図 2.1 におけるクラスとは「ヒント文 (+1)」と「非ヒント文 (-1)」の 2 値のことであり，スコアとは，SVM による分類で算出される値である．このスコアはヒントであるか否かの可能性を示しており，スコアが高くなるほどその文がヒントである可能性が高くなる．

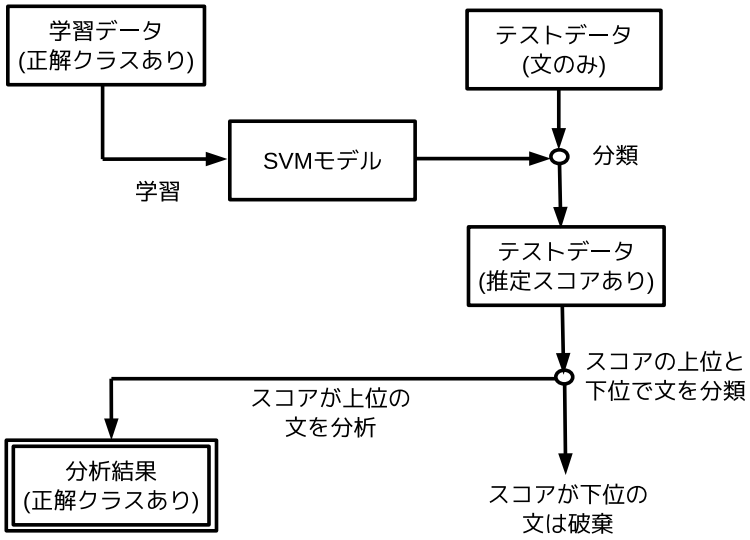

図 2.1 SVM の利用による自動抽出

2.2 能動学習

能動学習とは質問学習とも呼ばれ，専門家に質問しながら分類器の学習を行う手法である．この学習法により，分類器が判断に迷っている情報に対し修正を行うことができるため，分類器の性能の向上が期待できる．[3] では能動学習の利用により固有表現における学習コストを 3 分の 1 に軽減させることに成功した．

第3章　手法の提案

3.1　能動学習を用いたヒント文の自動抽出手法

この章では，能動学習を観光地分析に取り入れる手法についての説明を行う．まずは [2] で行われた，SVM を利用することによる基本的な手法についての説明を行う．

図 3.1 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，スコアでソートされた文に対し，スコアが高い文からその文がヒントであるかどうかの判別を分析者が行う．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の分析を行う．再学習前に分析したものと再学習後に分析したものすなわち図 3.1 における二重四角の部分を合わせたものがこの手法の出力となる．

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．

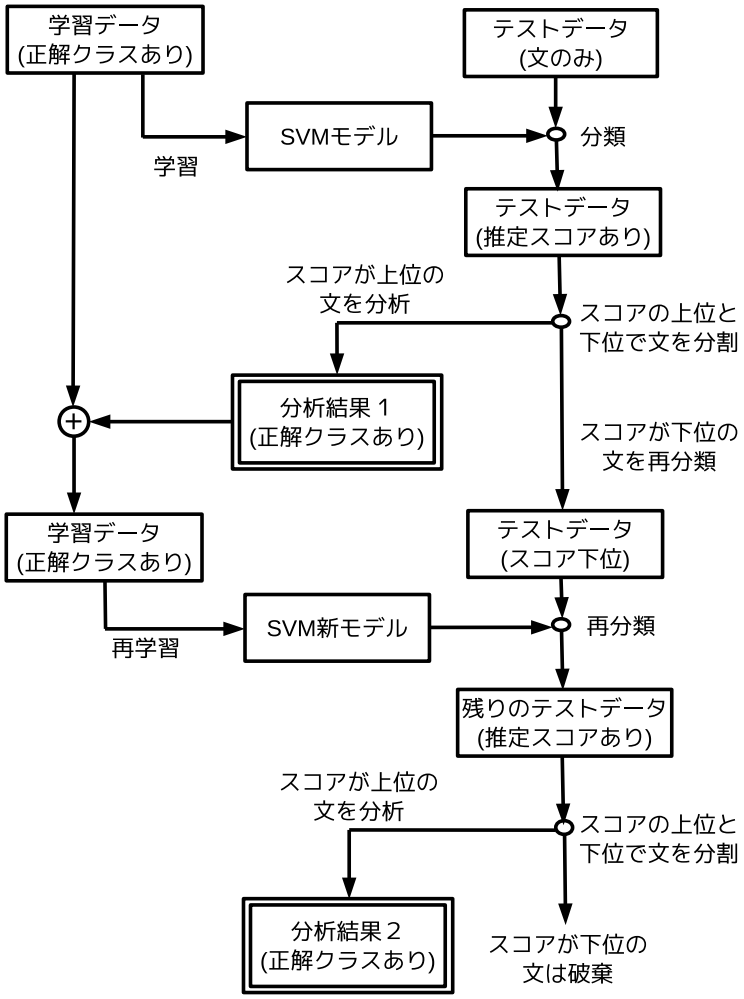

図 3.1 能動学習を用いた手法

3.2 ヒントの付与

学習データから SVM モデルを作成するために，ブログ文へ人手でヒントの付与を行う．ヒントがある文にはヒントあり (+1)，ヒントがない文にはヒントなし (-1) のクラスを付与し，ヒントありの文にはさらにヒントカテゴリを付与する．

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 分類のことであり [1]，全てのヒント文にはこれらのうち一つが付与される．

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．

E00001/-1/ヒントなし/昨日１１日、鉄ちゃんの後そのまま帰るのも芸が無いので、江ノ島に行ってみる事にした。
E00002/-1/ヒントなし/藤沢より江ノ電に乗り一路江ノ島駅へ …
E00003/-1/ヒントなし/江ノ島駅からは江ノ島まで歩いて１５分ほど途中小田急の駅により帰りのロマンスカーの時間確認と切符を買いに行き、１７時に出るロマンスカーの切符を入手。
E00004/-1/ヒントなし/出発時間まで散策時間２時間３０分歩き通しの散策に出発。
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。
E00009/+1/スポーツ・アウトドア/江ノ島を出た段階でまだ１時間以上時間が有った事から片瀬江ノ島海岸をぶらぶら、夏は海水浴で賑わう同海岸ですがオフシーズンの今は多くのサーファーと波打ち際で遊ぶ家族連れ、高校生たち、散歩をする犬など平静な光景が広がっていた。
E00010/+1/飲食/江ノ島についてから休憩も無く歩き通しで来た為、喉が渇いたのとお腹も減った事から少し早いが夕食をと思いまずは喉を潤す為生ビールそして食事はしらす丼を食べお腹を満たし時間も丁度よくなったので片瀬江ノ島駅よりロマンスカーに乗り帰路に着いた。

図 3.2 ヒントを付与したブログデータの例

3.3 使用する素性一覧

本研究ではブログデータから抽出した素性を SVM の学習および分類に用いる．本研究に用いる素性は以下の通りである．これらの素性が抽出された例を表 3.1 に示す.

- 各種品詞

  ブログ文に対して形態素解析を行うことにより各種品詞を抽出し，それを素性とする．素性として用いる品詞は，記号，名詞，動詞，形容詞，形容動詞，副詞，接続詞，感動詞，接辞，助詞とする．

- 各単語の特徴度

まず，BM25 を用いて各単語の特徴度 [1] を求め，その中から特徴度が 2 以上の単語の集合（以下，特徴語を呼ぶ）を求める．次に，ブログ文中に特徴語 w が含まれていた場合，$C:w$ の形で素性としてデータに追加する．また，一文中に特徴語となる複数含まれている場合は，全てを素性として追加する．特徴度は [1] と同様の手法を用いて算出する．

4.360423:ヒスイ
4.281887:糸魚川市
4.281887:フォッサマグナ
4.217718:高浪
4.170839:明星山
(中略)
2.056053:トレッキング
2.056053:たぐ
2.056053:くらげ
1.970173:浪
1.970173:野天風呂
1.970173:豊科
(中略)
-8.071920:こと
-8.076501:ん
-8.107385:の

図 3.3 各単語の特徴度の例

- **情緒推定による情緒**

まず，各文に対して情緒推定 [1] を行う．出力された「喜び」,「好ましい」,「恐れ」,「嫌だ」,「怒り」,「期待」,「驚き」,「悲しみ」,「なし」の 9 分類の情緒 a を $E:a$ の形で素性としてデータに追加する．また，情緒推定により複数の情緒が出力された場合はそれらを全て追加する．情緒推定は [1] と同様の手法を用いて行う．

E00001:なし/昨日１１日、鉄ちゃんの後そのまま帰るのも芸が無いので、江ノ島に行ってみる事にした。
E00002:なし/藤沢より江ノ電に乗り一路江ノ島駅へ …
E00003:嫌だ/江ノ島駅からは江ノ島まで歩いて１５分ほど途中小田急の駅により帰りのロマンスカーの時間確認と切符を買いに行き、１７時に出るロマンスカーの切符を入手。
E00004:嫌だ/出発時間まで散策時間２時間３０分歩き通しの散策に出発。
E00005: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:なし/老若男女ものすごい人手だ。
E00007:なし/朱の鳥居を超え階段を登り参拝、江ノ島大師、奥津宮ををを経て島の南端、稚児ヶ淵に到達。
E00008:なし/岩屋洞窟を見学の後来た道を戻った。
E00009:なし/江ノ島を出た段階でまだ１時間以上時間が有った事から片瀬江ノ島海岸をぶらぶら、夏は海水浴で賑わう同海岸ですがオフシーズンの今は多くのサーファーと波打ち際で遊ぶ家族連れ、高校生たち、散歩をする犬など平静な光景が広がっていた。
E00010:好ましい/江ノ島についてから休憩も無く歩き通しで来た為、喉が渇いたのとお腹も減った事から少し早いが夕食をと思いまずは喉を潤す為生ビールそして食事はしらす丼を食べお腹を満たし時間も丁度よくなったので片瀬江ノ島駅よりロマンスカーに乗り帰路に着いた。

図 3.4 各文の情緒推定の例

表 3.1 抽出された素性の例

ブログ文	抽出された素性
蓮華温泉は例年、雪が降る頃の１０月中旬に閉鎖されます。	蓮華温泉/は/例年/、/雪 が/降る/頃/の/１０月/中旬/に/閉鎖する/れる/ます/。/C:蓮華温泉/E:悲しみ
糸魚川に入ると「フォッサマグナ」と言う文字を良く見ます	糸魚川/に/入る/と/「/フォッサマグナ/」/言う/文字/を/良く/見る/ます/C:糸魚川/C:フォッサマグナ/E:好ましい/E:嫌だ/E:期待/E:恐れ

第4章　実装

本研究は，[1],[2] で使用されたプログラムを参考に作成した各種プログラムを用いて行った．開発に使用したプログラム言語はRuby，シェルスクリプトである．この章では，実験を行うにあたって作成した様々なプログラムの紹介を行う．

4.1　ブログデータの収集

ブログデータの収集は [1] で使用されたプログラムを参考に作成した各種プログラムを用いて行った．以下にブログデータの収集に用いたプログラムの一覧および収集したブログデータの例を示す．収集したブログデータについてもその例を示す．

表 4.1　ブログデータ収集プログラム一覧

プログラム名	概要
geturl_yahoo.rb	入力したキーワードからブログの URL を抽出するプログラム
yahoo_article_extract.rb	URL リストからブログ文を，ID 番号を付与して収集するプログラム
border.rb	ブログ文を記事単位に分割するプログラム

I000000 確か２年ぶりの晴山ゴルフ場。
I000001 会社関係で２０人弱でのコンペ。
I000002 初めてコースを周る初心者が数名いるので、このコースは距離が短いので最適かな。
I000003 東京から距離も近いし地方から転勤で来た人も観光地の軽井沢を案内できるので、何かと便利。
I000004 乗用カートは無いので、手引きカート。

図 4.1　ブログデータの例

4.2 素性の抽出

本研究では，ブログデータから素性を抽出し，ベクトル化することで SVM による学習および分類を行う．以下に素性の抽出で使用したプログラムの一覧および各データの例を示す．また，プログラム名が括弧で囲まれているものは [1],[2] で既に用意済のプログラムである．

表 4.2 素性抽出プログラム一覧

プログラム名	概要
(MorphAnalyzer.rb)	ブログデータの形態素解析を行うプログラム
(ExtractFeatures.rb)	形態素データから各種品詞を抽出するプログラム
op1.sh,op2.sh	[1] のプログラムを使用し，特徴度の算出を行うプログラム
op3.sh	[1] のプログラムを使用し，情緒推定を行うプログラム
addfeature.rb	素性に特徴度を追加するプログラム
addemotion.rb	素性に情緒を追加するプログラム
tool.sh	上記のプログラムを用いてブログ収集および素性の抽出を行うプログラム

I000000/確か/２/年/振り/の/晴山/ゴルフ場/。/C:晴山/E:なし
I000001/会社/関係/で/だ/２０/人/弱/の/コンペ/。/E:なし
I000002/初めて/コース/を/周/る/初心者/が/数/名/居る/要る/ので/のだ/、/この/は/距離/短い/最適/か/。/E:喜び
I000003/東京/から/距離/も/近い/し/地方/転勤/で/来る/た/人/観光/地/の/軽井沢/を/案内する/ので/のだ/、/何/か/と/便利/。/E:喜び
I000004/乗用/カート/は/無い/ので/のだ/、/手引き/。/C:乗用/E:なし

図 4.2 抽出した素性の例

4.3 SVMによる分類の動作

上記の素性データをベクトル化することにより SVM モデルを作成し，学習および分類に使用する．分類後のデータはスコア順にソートを行う．表 4.3 に使用したプログラム

一覧を示す．図 4.3 には SVM によるヒントの分類結果の例を示す．このデータは，SVM による分類スコア，文 ID，ブログ文で構成される．

表 4.3 作成した主なプログラム一覧

プログラム名	概要
(MySVM.rb)	TinySVM を使用し，素性データの学習および分類を行うプログラム
sort.rb	スコアデータのソートを行うプログラム
var.rb	SVM の分類結果から適合率，再現率，F 値を求めるプログラム
solveroc.rb	カテゴリ再現率（後述）を求めるプログラム
makemodel.sh	学習モデルを作成するプログラム
auto_var.sh	これらのプログラムを使用し，能動学習を行うプログラム

3.73424/I000471/「静岡～糸魚川を結ぶ日本列島構造線上に、二本の温泉断層がユニークにかさなった地点、地下１２００ｍの断層破砕帯中で、緑色凝灰岩が結晶した、石英閃緑岩の岩盤から湧出した植物性な透明薄緑色 (エメラルド湯)」
2.35589/I001759/世界各地で見られる多くの絶景ポイントがそうであるように、この景観も川の流れが長い年月をかけて作り出したものです。
2.33397/I000213/石畳の両側に紅殻格子のお茶屋が並び江戸時代の雰囲気が残っていますが、今も営業しているお茶屋は少なく、多くは茶房や和風雑貨屋や金箔を使った器やアクセサリーを扱うお店になってしまっています。
1.90724/I000214/「志摩」や「懐華楼」など、江戸時代そのままに残されたお茶屋建物もあり見学できます。
1.90374/I003178/戦国時代、織田と上杉の間に起こった凄惨な籠城戦で有名な所だ。(中略)
-4.27671/I001372/そんな思いでホテルを取ったまでは良かったのですが、いざチェックインして外へ行ってみるも、先ほどの富山ブラックラーメンの威力がまだ残っているようで、全然お腹が空きません。
-4.29781/I003196/ＧＷツー四日目、５月５日は富山から白馬を目指しました。
-4.55163/I000038/車両は 681 系の 9 両編成ですが、もしかしたら後ろ 3 両は北越急行の 683 系「スノーラビット」かもしれません。
-4.63064/I000534/富山での用事が済んだので青森へ帰るのですが、富山にどれぐらい滞在することになるか事前にははっきりわからなかったので、帰りの足は確保しないままでの富山入りでした。
-4.67727/I002193/１１月１０日（水）富山は雨‥。

図 4.3 SVM によるヒントの分類結果の例

第5章　実験

5.1　実験データ

本研究で行う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．

5.2　能動学習に関する実験パラメータ

提案手法（3.3 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．m, n はテストデータ総文数を分母とする．

本研究では以下の条件に当てはまる m, n の値の組において，それぞれ評価を行う．

$m = 0, 10, ..., 100$

$n = 10, 20, ..., 100$

$0 \leq m + n \leq 100$

また，図 5.1 に m, n の値を変えることによって求めた評価結果の例を示す．

```
m n F-measure
0 10 0.24
0 20 0.36
0 30 0.45
0 40 0.50
0 50 0.53
0 60 0.56
0 70 0.59
0 80 0.62
0 90 0.64
0 100 0.67
10 10 0.40
```

図 5.1 評価結果の例

5.3 比較手法

以下に，比較のための 2 通りの手法を提案する．

- **比較手法 1**

  この手法は自動分析使用せず，与えられた全ての文の分析を行う手法である．

- **比較手法 2**

  この手法は先行研究の手法（2.1.3 節）を用いてヒントの可能性の高いものから順番にある程度の文を分析者に提示し，分析を行う手法である．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．

5.4 評価基準

まず，本研究における評価の基準の説明を行う．評価基準は通常の情報抽出にならい，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．

$$P = \frac{|O \cap A|}{|O|} \tag{5.1}$$

$$R = \frac{|O \cap A|}{|A|} \tag{5.2}$$

$$F = \frac{2PR}{P+R} \tag{5.3}$$

$$R_\theta = \frac{1}{|C|} \sum_{c \in C} f(O, A_c; \theta) \tag{5.4}$$

$$f(O, A_c; \theta) = \begin{cases} 1 & (if \quad |O \cap A_c| > \theta \cdot |A_c|) \\ 0 & (otherwise) \end{cases} \tag{5.5}$$

$$F_\theta = \frac{2PR_\theta}{P+R_\theta} \tag{5.6}$$

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．

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．

観測した評価値を以下に示す．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)．

5.5 各評価基準における値

表 5.1 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 5.2 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 5.1, 表 5.2, 表 5.3, 表 5.4, 表 5.5 に適合率 P，再現率 R，F 値，カテゴリ再現率 R_θ とカテゴリ F 値 F_θ をそれぞれ求めた結果を示す．表 5.4 における閾値は $\theta = 0.2$ を使用する．

表 5.3 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 5.4 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									

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，$m = 0\%, n = 30\%$ の欄より 0.6 であり，提案手法では $m = 10\%, n = 20\%$ の欄と $m = 20\%, n = 10\%$ の欄より 0.66 と 0.67 であった．

表 5.5 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									

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．

5.6 評価のまとめ

表 5.3 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．

表 5.5 より，カテゴリ再現率を考慮した場合においては，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る．

このことにより，再学習前に 20%，再学習後に 20% 分析を行うことで分析量を削減し，かつ様々なカテゴリを網羅できるということが分かった．

また，F 値およびカテゴリ F 値のどちらの場合の評価においても，能動学習も用いた

場合は，m, n の値に関わらず比較手法 2 において同じ量だけ分析した場合すなわち分析量を $(m+n)\%$ にした場合と比較して性能が上昇することが分かった．このことから，どのような場合においても能動学習を用いない場合よりも用いる場合がよいということが分かる．

しかし，今回の実験で求めた $m = 20\%, n = 20\%$ という値の組み合わせは本実験コーパスに依存するものであり，別のコーパスにおいては最も性能が良くなる m, n の値は変化すると考えられる．現在分析前にこの値を知ることはできないため，この値の組み合わせを分析前や途中に求めることができるような方法の考案が今後の課題として挙げられる．

第6章　おわりに

本研究は，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

本研究を進めるにあたり，種々の御助言を頂きました鳥取大学大学院工学研究科情報エレクトロニクス専攻知能情報工学コース計算機工学Ｃ講座の村田真樹教授に心から御礼申し上げます．本研究を進めるにあたり，御指導を頂きました村上仁一准教授に心から御礼申し上げます．徳久雅人講師には，本研究の発案に始まり研究の仕方，論文の書き方など，研究全般において細部にわたって御指導を頂きました．ここに深く感謝いたします．

その他様々な場面で御助言を頂いた計算機工学Ｃ講座研究室の皆様に感謝の意を表します．

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.