原著論文

Web リソースと地理情報を活用した 新店舗情報の自動収集に関する研究[†]

寺口 敏生^{*1}・田中 成典^{*2}・西江 将男^{*1}

近年、情報関連技術の普及と発展に伴い、カーナビゲーションを代表とする地理空間情報サービスが普及している。しかし、地理空間情報サービスの基盤となる地図情報の整備は人手による現地踏査に依存しているため、新店舗の開店などの実空間の情報変化に地図情報が追従できていない問題がある。そこで、著者らは、Web上の新店舗の開店情報を対象として、自然言語から店舗名、住所や業種などの店舗情報を自動収集するテキストマイニングの研究を行ってきた。

しかし、新規開店に関連するキーワードだけを用いてマイニングを行った場合、業種別の単語出現傾向の違いや、単語の接続関係によって生じるノイズなどの自然言語の曖昧性に起因する課題に加え、情報自体の正誤を評価できないという課題に直面した。そこで、本研究では、キーワードによるマイニングに加え、マーケティング分析の指標に基づいて、店舗の業種と出店位置の地理的特性との関係を活用することで、これらの課題を克服することを目指す。実験から、地理的特性を考慮した情報を評価することによって、新店舗に関する開店情報を高精度に取得できることを実証した。

キーワード：地理空間情報サービス、地図情報の整備、新店舗情報、テキストマイニング、地理的特性

1. はじめに

インターネットの普及や情報端末の高機能化に伴い、GIS(Geographic Information System)を利用した数多くのサービスやシステムが提供されるようになった。GISは、ナビゲーションシステム[1]、防災情報の共有[2]、被災時のシミュレーション[3]などの公共サービス、店舗情報の共有サービスやマーケティング分析システムなど、様々な分野の基幹システムとして利活用されている。このような背景を受けて、政府は、地理空間情報活用推進基本計画[4]の策定を通じてGISの基盤情報である地図の普及を図ると共に、G空間プロジェクト[5]を立ち上げ、地理情報の利便性向上のための取り組みを進めている。したがって、GISを用いたシステムやサービスは、今後の発展が期待される分野であるといえる。

しかし、GISの基盤情報である地物の属性情報の更新は調査員の現地踏査に依存しており、地域によって踏査の頻度が異なるため、即時性を確保できない問題がある。GISを用いたナビゲーションシステムやマ-

ケティング分析などでは、地物の属性情報を最新に保つ必要があるが、調査員による現地踏査は高コストであり、あらゆる地域を網羅的に調査することは難しい。このため、地図の更新頻度が現実空間における地物の変化に追従できず、GISを利用したサービスの利便性を低下させる原因となっている。そこで、地物の属性情報を自動抽出する研究[6-8]が取り組まれている。しかし、これらの既存研究では、店舗名や業種などの属性情報が既知であるものに対してしか適用できないため、属性情報が不明な新店舗には適用できないという問題がある。この問題に対応する方法として、基盤情報である新店舗の属性情報を自動的に取得する研究を組み合わせることが考えられる。

新店舗の属性情報を取得する既存研究では、WebページのHTMLタグ構造の繰り返しパターンや表構造に着目して解析する手法[9-11]と新店舗情報を含むコンテンツに出現する単語に着目して解析する手法[12]が提案されている。WebページのHTMLタグ構造の繰り返しパターンや表構造に着目して解析する手法では、店舗名、業種、住所や電話番号などの情報が表形式などに整理された情報源を対象とするため、効率よく高精度に新店舗の情報を収集できるという特徴がある。しかし、これらの手法は、ブログや掲示板などのCGM(Consumer Generated Media)上に投稿される自然言語で記述された情報には適用できないため、利用できる情報リソースが限定されるという問題がある。

[†] Research on Automatic Collection of New Store Information Utilizing Web Resource and Geographic Information
Toshio TERAGUCHI, Shigenori TANAKA and Masao NISHIE

^{*1} 関西大学大学院総合情報学研究科
Graduate School of Informatics, Kansai University

^{*2} 関西大学総合情報学部
Faculty of Informatics, Kansai University

一方、新店舗情報を含むコンテンツに出現する単語に着目して解析する手法では、コンテンツ中で使用される単語の特徴を学習するため、単語に基づいた新店舗情報の取得が可能である。この手法は、自然言語で記述されたコンテンツを解析対象とするため、幅広い情報源から情報を収集できる。しかし、出現単語に着目してコンテンツを解析する場合、業種ごとにコンテンツ中で使われる単語の傾向が異なったり、前後の文脈により単語の意味的な役割が異なるといった自然言語の曖昧性により、精度が低下するという課題がある。また、自然言語で記述されたコンテンツを適切に解析できた場合でも、表記ミスや誤情報のような情報源としてのCGMの信頼性により、解析結果の評価が困難である。

以上より、WebページのHTMLタグ構造の繰り返しパターンや表構造に着目して解析する手法は、高精度に情報を取得できるが情報源とする対象が限られており、また、新店舗情報を含むコンテンツに出現する単語に着目して解析する手法は、情報源とする対象は膨大であるが取得した情報の精度が低いという問題をそれぞれ抱えている。そこで、本研究では、Web上に投稿された自然言語で記述されたコンテンツを研究対象として、幅広い情報源から高精度に新店舗情報を取得する。そして、取得した新店舗情報をGIS上に反映することで、地物の属性情報を自動的に更新する手法の実現を目指す。

2. 研究の概要

本研究では、自然言語で記述されたWebページ群から新店舗情報のみを取得することを目的とした場合の既存手法の課題を解決する。既存手法を分析したところ、自然言語で記述された文書の曖昧性とCGM上の情報の信頼性などの性質から、「業種ごとに出現する単語の傾向が異なることに起因する課題」、「文脈により単語の意味的な役割が異なることに起因する課題」と「CGMの信頼性に起因する課題」の3つの課題を抽出した。そこで、本研究では、これらの3つの課題について以下の手法を用いて解決する。

(1) 業種ごとに出現する単語の傾向が異なることに起因する課題

業種ごとに出現する単語の傾向が異なることに起因する課題に対しては、業種ごとに接続関係を考慮した複合語の出現傾向を学習することで解決する。業種ごとに学習用データを用意し特徴語を学習することで、業種ごとに単語別の出現頻度の違いを考慮できる。この時、業種を特徴づける複合語が誤って複数の単語に

分割される場合があるため、複合語用語抽出処理を行い業種ごとの特徴を強調した学習を行う。

(2) 文脈により単語の意味的な役割が異なることに起因する課題

文脈により単語の意味的な役割が異なることに起因する課題に対しては、前後の単語間における接続関係を考慮して学習することで解決する。たとえば、「開店」や「オープン」は新店舗情報を取得するために有用な単語であるが、「朝8時にオープン」などの新店舗とは関係がない状況でも用いられるため、単語の出現頻度のみを用いて学習すると学習の精度が低下する。そこで、単語の前後の接続関係を考慮し、扱われる文脈ごとに別々の特徴語として学習することで、学習の精度を向上させる。

(3) CGMの信頼性に起因する課題

CGMの信頼性に起因する課題に対しては、自然言語による記述内容だけではなく、現実空間における出店位置の地理的な特性を併せて考慮し情報を評価することで解決する。従来の既存研究では、新店舗情報の中でも住所情報に関する正確さを検証することなく利用していた。しかし、特徴語を用いて新店舗情報を含むWebページを抽出した場合、表記のミスや誤情報を掲載したWebページをも取得することがあるため、正確な情報を抽出できない課題があった。そこで、本研究では、新店舗の出店位置が地理的な特性に依存する点に着目し、地理的な特性を考慮したフィルタリングを行う。地理的特性を考慮することで、現実には存在しない新店舗の情報、記述内容の齟齬や誤情報が発生する課題を解決する。

本研究では、以上の3つの課題を解決し、幅広い情報源から新店舗情報を高精度に取得する手法を提案する。

3. 地理的特性の分析

既存の自然言語で記述されたコンテンツを解析する研究[6-11]では、現実空間における店舗の立地条件が考慮されていなかった。しかし、企業が出店位置を選定する場合、一般的に「交通条件」、「顧客層」、「競合他店や関連業種の出店状況」や「地域ごとの潜在的な特徴」などをマーケティング分析し、損益を見込んだ上で意志決定を行う。また、業種によっては、風俗営業法などの法律によって出店地域が制限されたり、風俗街や倉庫街などの競合他店との位置的な相関が見られる場合がある。このため、これらの要素を分析することで、分析対象地点における業種別出店確率を算出できると考えられる。そこで、本研究では、これらの地理的な要素に基づいて決定される地域特性を地理的特

性と定義し、自然言語で記述された情報の曖昧性を地理的特性に基づいたフィルタリングによって補正する。本研究では、「交通条件」、「対象となる顧客層」、「用途地域」、「風俗営業法」、「店舗間の相関」と「地域の潜在的な要素」の6種類の地理的特性を採用した。各地理的特性の内容とその意図を次に列記する。

(1) 交通条件

分析対象地点へのアクセスのしやすさは、全ての業種において重要となる要素である。しかし、飲食業や小売業の場合と流通業やガソリンスタンドなどの場合では、重視する交通条件は異なると考えられる。そこで、本研究では、交通条件として、地図情報や道路ネットワークから取得した電車やバスの駅からの距離、高速道路の出入り口からの距離、隣接する道路種別と幅員、および幹線道路からの距離を基に、分析対象地点の交通条件を評価した。

(2) 対象となる顧客層

分析対象地点の商圈における人口分布や顧客層は、特に飲食業や小売業などで重要な要素と考えられる。しかし、業種や地域によって商圈の定義は異なるため、機械的な分析は困難である。そこで、本研究では、顧客層を分析する指標として、区(市区町村)レベルと字レベルの両方の範囲を商圈として考慮し、数値地図や統計局の統計情報から取得した各商圈における周辺人口、世帯数、昼間人口と地価を基に、分析対象地点の顧客層を評価した。ここで、世帯数や昼間人口は、分析対象の商圈がビジネス街なのか住宅街なのかで特徴が異なると考えられるため、地理的特性として採用した。また、地域住民の経済レベルを反映する指標の一つとして地価を地理的特性に採用した。

(3) 用途地域

都市計画法により分析対象地点がどの用途地域に分類されているかは、業種を特定する上で重要な要素である。また、分析対象地点の用途地域と併せて建蔽率と容積率を考慮することによって、建物の規模を推定できる。そこで、本研究では、用途地域を取得する情報源として、たとえば大阪市であれば、大阪市計画調整局が公開するマップナビおおさか[13]のような公共サービスを基に、分析対象地点の用途地域、建蔽率と容積率を取得した。

(4) 風俗営業法

周辺にある施設との位置関係によって、特定の業種は、分析対象地点に出店が不可能となる場合がある。たとえば、風俗関連業種は、風俗営業法によって学校や病院などの保護対象施設から一定距離以内に出店することが規制されている。したがって、保護対象施設

は風俗関連業種の出店情報を補正する指標として有用であると考えられる。そこで、本研究では、風俗営業法に定められている建築規制に基づき、保護対象施設と分析対象地点における距離を考慮することで出店規制を考慮した。なお、保護対象施設は地方によって異なるため、分析対象地域ごとに風俗営業法の保護対象となる施設分類を取得した。

(5) 店舗間の相関

分析対象地点の周辺地域における同業種店舗や他業種店舗の出店状況は、マーケティング分析において重要な要素である。しかし、業種によっては特定の地域に密集して出店される場合もあり、一概に店舗間の相関を定義することはできない。そこで、本研究では、店舗間の相関を考慮する指標として、分析対象地点から一定距離内に含まれる他業種店舗および同業種店舗の数に基づき、分析対象地点の店舗間の関係を評価した。

(6) 地域の潜在的な要素

中華街やコリアタウンなどの文化的、および風土的な背景から、地域ごとに出店する業種の傾向が異なる場合がある。そこで、本研究では、地域の潜在的な要素として、地名や最寄り駅の名前、および路線名に基づき、分析対象地域の潜在的な文化的背景を評価した。

本研究では、以上の6項目を用いて、解析対象の出店位置に出店される可能性が高い業種を特定する。

4. 研究の対象と処理の流れ

4.1 本研究の対象

本研究では、Web上に投稿された自然言語で記述された新店舗情報を効率良く自動的に抽出することを目的とする。本節では、本研究において解析対象とするWebページと抽出対象とする新店舗情報について定義する。本研究で解析対象とするWebページは、住所情報を記述内容に含むものとする。これは、効率よく店舗に関する情報を収集する条件として、住所情報を含むWebページが最適であると考えたためである。また、自然言語で記述されたコンテンツを対象として新店舗情報を抽出する。これは、消費者が掲示板やブログなどのCGMに投稿した情報を取得対象に含めることで、情報源の範囲を拡張するためである。これらの定義に基づき、新店舗情報を取得する。

4.2 処理の流れ

本提案手法のシステムは、図1の処理の流れに示すように、学習部と判別部に分かれている。学習部は、地理的特性学習機能、業種別特徴語学習機能と業種別新店舗情報の有無判別モデル構築機能の3つの機能で

図1 処理の流れ

構成されている。判別部は、業種別新店舗情報の有無判別機能と新店舗情報抽出機能の2つの機能で構成されている。これらの処理を通じて新店舗の属性情報を抽出する。処理の流れを以下に示す。

学習部の地理的特性学習機能では、図2に示すように、業種別住所録、地点情報や統計情報を入力とし、各店舗が出店している地理的特性を解析することで、業種ごとに开店傾向モデルを構築する。本機能により構築した开店傾向モデルは、新店舗情報が記載されたWebページ中から住所情報を抽出し、住所が示す地点における業種別开店確率を算出する場合に利用する。

学習部の業種別特徴語学習機能では、図3に示すように、業種別に用意した店舗情報を含むWebページ群を入力する学習データとし、Webページ中の特徴語の出現傾向を学習することで、単語に基づく業種判別モデルを構築する。この時、日本語複合語用抽出手法[14]を用いて単語の組み合わせを複合語として扱い、学習を行った。これは、一単語のみに着目した場合、一般語であっても組み合わせで複合語として扱うことで業界用語になる単語を学習するためである。本機能により構築した単語情報に基づく業種判別モデルは、Webページ中で話題に上っている店舗の業種を判別する場合に使用する。

学習部の業種別新店舗情報の有無判別モデル構築機能では、図4に示すように、業種別に用意した店舗情報を含むブログ記事群を新店舗情報の有無によって分割して入力する学習データを作成し、新店舗情報の有

図2 地理的特性学習機能

図3 業種別特徴語学習機能

図4 業種別新店舗情報の有無判別モデル構築機能

無による単語の出現傾向を学習することで、業種別新店舗情報有無判別モデルを構築する。この時、文脈を考慮して特徴語を学習するために、形態素n-gram[15]を利用して特徴語の接続関係の特徴を学習する。本機能により構築した業種別新店舗情報有無判別モデルは、判別データとして入力されたWebページ中の文

図5 業種別新店舗情報の有無判別機能と新店舗情報抽出機能

書を解析し、文書に含まれる話題が新店舗情報を含む確率を算出する場合に利用する。

判別部の業種別新店舗情報の有無判別機能では、図5に示すように、判別対象Webページを解析して抽出した文書情報と住所情報を入力する。そして、文書情報からは特徴語に基づく新店舗情報の有無確率を算出する。また、住所情報からは地理的特性に基づく新店舗情報の有無確率を算出する。その後、業種ごとの特性に基づいて、特徴語の解析結果から求めた確率と地理的特性の解析結果から求めた確率を統合した確率を算出する。この統合確率が閾値以上であれば、新店舗情報を含むWebページと判別し取得する。

判別部の新店舗情報抽出機能では、図5に示すように、業種別新店舗情報の有無判別機能による判別結果を入力とし、新店舗情報を含むWebページから、店舗名、住所、電話番号、出店日と業種などの属性情報を取得する。

本研究では、これらの処理を行うことで、Webページから自然言語で記述された新店舗情報を抽出する手法を提案する。

5. 地理的特性の取得

本研究では、第3章で分析した6種類の地理的特性を用いて、判別対象の住所に出店される業種を判別する。そこで、住所録にある店舗ごとの住所情報を基に、業種ごとの出店地域の地理的特性を取得し、業種ごとの出店傾向モデルを構築する。なお、本研究では、タウンページに掲載されている施設情報や店舗情

報から、店名、業種、住所と電話番号を抽出する。そして、住所情報をジオコーディングし、算出した緯度経度を店舗の出店位置と扱い、出店位置における店舗ごとの地理的特性を取得する。本章では、6種類の地理的特性の取得方法についてそれぞれ解説する。

(1) 交通条件

本研究では、交通条件として、電車の最寄り駅からの距離、最寄りのバス停からの距離、高速道路の出入り口からの距離、幹線道路からの距離に加えて、隣接する道路の道路種別および幅員を分析する。

電車の最寄り駅からの距離の取得方法では、まず、Web上をクロールし、全駅名と住所情報を抽出する。本研究では、地図検索サイトのマピオン[16]から駅名と住所情報を取得した。次に、取得した住所情報を基に、駅の緯度経度を算出する。最後に、解析対象店舗の緯度経度と全駅の緯度経度の直線距離を算出し、最小値を最寄り駅からの距離として取得した。また、最寄りのバス停や高速道路の出入り口からの距離については、マピオンのバス停情報[17]や高速道路IC情報[18]を基に、電車の最寄り駅からの距離を求めるのと同様の手法により取得した。

幹線道路からの距離の取得方法では、国土地理院が作成した数値地図25,000(空間データ基盤)より、道路種別が幹線道路として登録されている道路の端点情報と区間情報を取得する。そして、これらの情報を基に接続した幹線道路の道路中心線と解析対象店舗との最短距離を算出し、幹線道路からの距離として取得した。なお、隣接する道路については、解析対象店舗周辺の道路端点情報と道路区間情報を基に接続した全ての道路中心線との距離を算出し、距離が最短となる道路を隣接する道路として取得した。また、隣接する道路の地理的特性として、数値地図25,000より、道路種別と道路幅員を取得した。なお、本研究で取得した道路種別は、一般道、高速道路、石段と庭園路の4種類である。

(2) 対象となる顧客層

本研究では、対象となる顧客層を分析する指標として、区レベルと字レベルの両方における周辺人口や世帯数[19]、昼間人口[20]、および地価を用いる。なお、解析対象店舗の住所における地価は、国土交通省の地価公示[21]に掲載されている地点の中から、最も距離に近いものの値を用いる。これらの情報は、統計局や地方自治体が公開している調査結果や統計を手作業で収集し、解析対象店舗の出店地域における統計情報を取得する。

(3) 用途地域

本研究では、解析対象店舗の住所における用途地域を判別するため、自治体が公開する都市計画より、地点ごとの用途地域を取得する。しかし、用途地域の情報は、画像やGISを介して提供される場合が多く、単純なクロージングでは取得できない。このため、解析対象とする地域に適した情報の取得方法を検討する必要がある。本研究では、大阪市計画調整局が公開するマップナビおおさかのように、住所を入力するとGIS上に用途地域が表示されるシステムを対象として、自動的に情報を取得するシステムを開発した。このシステムでは、まず、解析対象店舗の住所を自動的に入力し、その地点が中央に表示されるようにGIS画面を移動させる。次に、自動的にマウスを制御して、GIS画面の中央にあたる座標をクリックイベントを発生させることで、入力した住所の地点をGISにて指定する。最後に、画面に表示される解析対象店舗の住所における用途地域、建蔽率、および容積率を取得する。

(4) 風俗営業法

本研究では、風俗営業法の影響力を考慮する指標として、保護対象施設と解析対象店舗間の距離を取得する。この時、保護対象施設に該当する業種として、多くの自治体で保護対象施設に指定されている教育機関の関係業種と診療所や医療関係業種を考慮した。

(5) 店舗間の相関

本研究では、店舗間の相関を考慮する指標として、分析対象地点から一定距離内に含まれる他業種店舗および同業種店舗の数を取得した。なお、本研究では、地図上の直線距離により店舗間の距離を評価した。

(6) 地域の潜在的な要素

本研究では、業種ごとの出店傾向に影響を与える地域の潜在的な要素を考慮する指標として、本項(1)で取得した解析対象店舗の最寄り駅の名前、路線名、市区町村名と字名を取得した。これらの要素を考慮することで、地区別の出店業種の偏りを取得する。

以上の手法により取得した業種ごとの出店位置における地理的特性を分析し、出店傾向モデルの構築に用いる素性を抽出する。

6. 新店舗情報の有無判別

6.1 3つの判別モデルの構築

本節では、出店傾向モデル、業種判別モデルと業種別新店舗情報有無判別モデルを構築するための素性について解説する。各判別モデルの構築の流れを図6に示す。

図6 各判別モデルの構築の流れ

(1) 出店傾向モデルの素性

出店傾向モデルは、業種ごとの出店傾向に影響を与える地理的特性を学習した判別器である。出店傾向モデルに住所を入力することで、入力した住所の位置に出店し易い業種や出店し難い業種を取得することが可能となる。出店傾向モデルの構築に用いる素性の一覧を表1に示す。出店傾向モデルの構築時には、第5章で前述した方法で取得した6種類の地理的特性を基に、表1に示す素性をそれぞれ抽出した。

(2) 業種判別モデルの素性

業種判別モデルは、業種分けされた店舗情報を含むWebページから抽出した特徴語を用いて、業種別の単語の出現傾向を学習した判別器である。業種判別モデルにWebページを入力することで、入力したWebページに記述されている店舗の業種を判別することが可能となる。業種判別モデルを構築する素性は、業種分けされたブログ記事の記述内容を形態素に分割し、これに日本語複合語用語抽出手法を用いて抽出した複合語の出現頻度を解析することで取得する。

(3) 業種別新店舗情報有無判別モデルの素性

業種別新店舗情報有無判別モデルは、業種ごとに収集した新店舗情報を含むWebページと住所録に登録されている既知の店舗情報を含むWebページを用いて、それぞれのWebページ群における単語の出現傾向を学習した判別器である。業種別新店舗情報有無判別モデルにWebページを入力することで、入力したWebページに記述されている内容が新店舗の開店情報であるか、そうでないかを判別することが可能となる。業種別新店舗情報有無判別モデルを構築する素性は、入力したWebページの記述内容を形態素解析と日本語複合語用語抽出手法にかけ抽出した複合語用語を対象に、形態素n-gramを用いて接続関係を解析すること

表1 出店傾向モデルの構築に用いる素性一覧

地理的特性	各地理的特性から取得する素性	
(1)交通条件	最寄り駅からの距離 (徒歩 1 分～60 分) (1 分につき道路距離 80m)	60 種類
	最寄りのバス停からの距離 (徒歩 1 分～60 分) (1 分につき道路距離 80m)	60 種類
	高速道路出入り口からの距離 (車で 1 分～60 分) (1 分につき道路距離 500m)	60 種類
	隣接する道路の種類と幅員	15 種類
	道路種別	道路幅員
	一般道	1.5m 未満
	一般道	1.5m 以上 3.0m 未満
	一般道	3.0m 以上 5.5m 未満
	一般道	5.5m 以上 13.0m 未満
	一般道	13.0m 以上
	一般道	真幅道路など
	高速道路	3.0m 以上 5.5m 未満
	高速道路	5.5m 以上 13.0m 未満
	高速道路	13.0m 以上
	高速道路	真幅道路など
	石段	5.5m 以上 13.0m 未満
	庭園路	3.0m 以上 5.5m 未満
	庭園路	5.5m 以上 13.0m 未満
	庭園路	13.0m 以上
	庭園路	真幅道路など
(2)対象となる顧客層	幹線道路からの距離 (50.0m 刻みで素性を生成)	
	周辺人口	区レベルでは 5,000 人刻み、字レベルでは 500 人刻みで素性を生成
	世帯数	区レベルでは 5,000 世帯刻み、字レベルでは 500 世帯刻みで素性を生成
	昼間人口	区レベルでは 5,000 人刻み、字レベルでは 500 人刻みで素性を生成
	地価	50,000(円/㎡)刻み
(3)用途地域	用途地域	第 1 種低層住居専用地域
		第 2 種低層住居専用地域
		第 1 種中高層住居専用地域
		第 2 種中高層住居専用地域
		第 1 種住居地域
		第 2 種住居地域
		準住居地域
		近隣商業地域
		商業地域
		準工業地域
		工業地域
		工業専用地域
(4)風俗営業法	建蔽率	
	容積率	
(5)店舗間の相関	建築規制のある業種（学校関係施設、保育所、病院、診療所、その他）からの距離 (50m 以内と 100m 以内)	
	他業種店舗（関連）の数 (100m 圏内、500m 圏内、1,000m 圏内、2,000m 圏内、5,000m 圏内の各圏において、条件に合致する店舗が何件あるかを分析. 5 件刻みで素性を生成)	
(6)地域の潜在的な要素	同業種店舗（競合）の数 (100m 圏内、500m 圏内、1,000m 圏内、2,000m 圏内、5,000m 圏内の各圏において、条件に合致する店舗が何件あるかを分析. 5 件刻みで素性を生成)	
	地名（区レベルと学レベルの両方を使用） 最寄駅の名前 路線名	

で取得する。なお、入力するWebページには、新店舗情報を含むWebページと含まないWebページの両方を用いる。形態素の接続関係を用いることで、新店舗情報の有無を判別するための詳細な素性を構築できる。

(4) 各判別モデルの学習

本研究で提案する3つの判別モデルの構築では、図6の各判別モデル構築の流れに示すように、学習に用いる入力情報はそれぞれ異なるが、学習に用いる手法は、潜在的ディリクレ配分法(LDA: Latent Dirichlet Allocation) [22]を用いて次元圧縮[23]を行った各学習

データの素性をSVM(Support Vector Machine) [24]に入力するという点で共通している。SVMは一般的に高精度なフィルタリング手法であるが、本節(1)から(3)のように、膨大な素性の発生が見込まれる場合は、スパースネスの問題を回避するため、次元圧縮手法を用いて擬似的に素性の次元を圧縮する必要がある。その時、既存手法[12]では、主成分分析(PCA: Principal Component Analysis)を用いて次元圧縮を行っていたが、業種ごとにデータの分布は多様であると考えられるため、PCAを用いた場合、適切な次元圧縮を行えない可能性がある。そこで、本研究では、素性同士の意味的な近さを考慮して特徴空間を構築できるLDAを用いて次元圧縮を行った。

また、通常のSVMでは、識別結果をクラスに属するか属さないかの2値で出力する。しかし、店舗は、「フレンチ&イタリアン」といったように、一つの業種に属するとは限らず、複数の業種に属する可能性があるため、各クラスに属する度合いを算出する必要がある。そこで、SVMによって構築した各クラスの分離超平面との距離を基に、各クラスに属する確率を算出する仕組みを構築した。

これらの処理により構築した各種判別モデルを用いて、判別対象として入力されたWebページに新店舗情報が含まれる確率を算出する。

6.2 新規出店情報の有無判別

構築した3種類の判別モデルを用いて、図5に示す流れで判別対象のWebページ中に新店舗情報が含まれるか否かを判別する。本節では、業種判別処理、特徴語に基づく新店舗情報の有無確率算出処理、出店傾向解析処理と統合確率算出処理について述べる。

(1) 業種判別処理

業種判別処理では、学習部にて構築した単語情報に基づく業種判別モデルを用いて、判別対象として入力されたWebページ中で話題に上がっている店舗の業種を判別する。この時、日本語複合語用語抽出処理を用いて単語の組み合わせから複合語を構築することで、業種判別モデルの構築に用いた特徴語と同じレベルの素性をSVMに入力し、各業種に属する確率を判別結果として出力する。

(2) 特徴語に基づく新店舗情報の有無確率算出処理

特徴語に基づく新店舗情報の有無確率算出処理では、業種判別処理の判別結果と学習部にて構築した業種別新店舗情報有無判別モデルを用いて、新店舗情報の有無確率を算出する。この時、日本語複合語用語抽

出処理を用いて複合語を抽出すると同時に形態素n-gramを用いて接続関係を考慮した特徴語を構築することで、新店舗情報有無判別モデルに用いた特徴語と同じレベルの素性をSVMに入力し、新店舗情報の有無確率を出力する。また、業種判別処理で算出した各業種に属する確率に応じて、業種ごとの新店舗情報の有無確率に重みづけを行うことで、対象Webページに関連する業種の単語の出現傾向のみを考慮した新店舗情報の有無判別が可能となる。さらに、複数の業種を跨ぐような店舗の場合にも対応できる。

(3) 地理的特性に基づく新店舗情報の有無確率算出処理

地理的特性に基づく新店舗情報の有無確率算出処理では、学習部にて構築した地理的特性に基づく出店傾向モデルを用いて、各業種が判別対象のWebページ中に含まれる住所の位置に出店する確率を算出し、出力する。

(4) 統合確率算出処理

統合確率算出処理では、単語情報を手掛かりとした業種判別処理と特徴語に基づく新店舗情報の有無確率算出処理で出力した業種ごとの新店舗情報有無確率と、地理情報を手掛かりとした地理的特性に基づく新店舗情報の有無確率算出処理で出力した業種ごとの出店確率とを統合し、判別対象のWebページにおける新店舗情報の有無確率を算出する。この時、業種によっては地理情報における判別が難しい場合があるため、業種ごとに統合する時の重みを変える必要がある。そこで、本処理では、あらかじめ、出店傾向モデルと業種判別モデルの業種判別精度の比を算出しておくことで、業種ごとに地理情報と単語情報の重みを付加する。この時、SVMで出力した距離が最も近い業種を正解として判別した回数で業種判別の精度を比較すると、業種の確実な判別ではなく、絞り込みを目的とした出店傾向モデルの適切な精度を算出できない。また、判別結果の業種の順位を考慮することも考えられるが、順位のみでは、例えば、SVMで出力した距離が最も近かった業種と2番目に近かった業種でどれだけの差があったのかを考慮できず、適切な精度を算出できない。そこで、本処理では、SVMにより最も距離が近いと判別された業種と最も距離が遠いと判別された業種との間における判別対象の業種の相対的な位置を算出する。そして、最も距離が近いと判別された業種と判別対象の業種の相対的な距離を業種判別の誤差値とし、各業種の誤差値の平均を業種判別における業種ごとの判別精度とした。ある業種 C_m の誤差値の平均 T_m は、式(1)により算出する。なお、 D は、業種 m における i 番目のデータのSVMにより出力された距

離である。また、 D_{Low} は、最も距離が遠かった業種の距離である。そして、 D_{Top} は、最も距離が近かった業種の距離である。ただし、 $n = C_m$ のデータ数とする。

$$T_m = \frac{1}{n} \sum_{i=0}^n \left\{ 1 - \frac{(D_{Low_i} - D_i)}{(D_{Low_i} - D_{Top_i})} \right\} \quad (1)$$

本処理では、あらかじめ、出店傾向モデルと業種判別モデルの業種判別結果を式(1)に入力し、各業種の単語情報と地理情報の誤差平均値の比を算出しておくことで、それぞれのモデルに対して重み付けを行う。そして、重み付けした確率を統合し、得られた新店舗情報の有無確率が閾値以上である場合、新店舗情報が含まれると判別する。また、新店舗情報の有無確率が閾値を下回る場合は、新店舗情報が含まれていないと判別する。

7. 新店舗の属性情報抽出

新店舗情報が含まれると判別したWebページから、新店舗の店舗名、住所、電話番号、出店日と業種を抽出する。

まず、地名辞書を用いることで住所を取得し、正規表現によって出店日と電話番号を取得する。そして、住所、電話番号と出店日を出力する。

次に、新店舗の店舗名は複合語である可能性があるため、業種判別処理と同様に日本語複合語用語抽出処理を行い、複合語を構築する。この時、既存手法[25]では地名などの情報を用語生成におけるノイズとして除外しているが、店舗名では地名のみや地名を店名の一部にしているケースが少なからず存在しているため、本研究では地名を考慮した。その上で、チェーン店の店舗名に散見される末尾に地名が入るようなパターンについてはクリーニング手法[25]を適用して除外することとした。これは、本提案手法にて抽出した新店舗の属性情報を住所録に登録し、学習データとして再利用するための工夫である。

そして、取得した名詞句を新店舗名候補とし、新店舗名の絞り込みを行う。この時、CGMでは、各要素の省略が頻繁に起こるため、地物に係る性質のある語を手掛かりとする構文解析処理によって新店舗名を取得するのは困難である。また、新店舗名は、未知語であるため、TF-IDF法などの統計的に特徴語を抽出する手法では取得できない。そこで、本処理では、店舗名が住所、電話番号、「オープン」や「開店」といった新店舗特徴語の近くに出現することに着目し、住所、電話番号と本文中に出現するそれぞれの新店舗特徴語からの文字数を指標に用いる。なお、本文中に同じ新店

舗名候補が複数回出現する場合、住所、電話番号とそれぞれの新店舗特徴語に対して最も位置的に近い新店舗名候補の文字数を採用する。

また、新店舗名以外の不要語を除去するため、各新店舗名候補のTF-IDF値を算出する。さらに、本文中で新店舗名が「ブルーバード」を「Blue Bird」といったように、アルファベットをカタカナに言い換える場合や省略される場合を考慮し、各新店舗の店舗名候補に対して翻訳やローマ字変換を行い、文字n-gramによるテキストの類似度判定手法[26]によって新店舗名候補同士の類似度を算出した。

最後に、取得した住所からの文字数、電話番号からの文字数と各新店舗特徴語からの文字数を用いて、新店舗名候補の絞り込みを行う。m番目の新店舗名候補文字列のスコア N_m を式(2)により算出する。なお、式(2)において、 $Tfidf$ はm番目の新店舗名候補文字列のTF-IDF値、 L_a は住所からの文字数、 L_t は電話番号からの文字数、 L_i はi番目の新店舗特徴語からの文字数とする。

$$N_m = Tfidf \times \left\{ \frac{1}{\log\{L_a + L_t + (\sum_{i=0}^n L_i)\}} \right\} \quad (2)$$

そして、全ての新店舗名候補文字列を式(2)に入力し、 N_m の最も高い新店舗名候補を新店舗の店舗名として出力する。

8. 実証実験

8.1 実験の概要

本提案手法の有用性を確認するために、図7に示す実験計画に従って、実証実験を行う。なお、図7は、図5と対応しており、どの項目についての実験を行ったかを可視化したものである。図7に示すように、本実験計画は、3つの事前実験と2つの評価実験で構成されている。それぞれの実験について概説する。

事前実験1では、文書情報と住所情報の次元圧縮手法として、PCAとLDAの両次元圧縮手法の効果を比較する。事前実験2では、文書情報から抽出した単語間の接続を考慮する形態素n-gram手法において、nの最適値を算出する。事前実験3では、特徴語に基づく新店舗情報の有無判別確率と地理的特性に基づく新店舗情報の有無判別確率を統合する時に、業種によって各確率の重みを変動させる必要性を検証すると同時に、業種ごとの各確率の重みを算出する。

評価実験1では、算出された統合確率を用いた新店舗情報の有無判別の精度を評価する。評価実験2では、新店舗情報の有無判別によって正しく判別された新店舗情報を含むWebページから新店舗の属性情報を取得する精度について評価する。これらの実験を通じて、本提案手法の有用性を検証する。

8.2 学習データとテストデータ

本節では、実証実験で用いる学習データとテストデータおよびその取得方法について解説する。本実験では、大阪市全区域を解析対象地域として、学習データおよびテストデータを収集した。また、特徴を学習する対象業種として、タウンページを分析し、類似するものを独自にまとめた業種分類より、入れ替わりが激しい飲食業、小売業やサービス業など計220業種を採用した。本提案手法にて学習対象とした業種一覧を表2に示す。これらの基準に基づき、学習データを収集した。本実証実験で用いる各判別モデルの学習データ一覧を表3に示す。なお、業種別新店舗情報有無判別モデルの構築時において、2008年12月1日から2010年12月1日までの期間中に投稿された新店舗情報を含むブログ記事をYahoo! Blog検索から手作業で収集したところ、200記事以上投稿されている業種は、表2の灰色に色づけした36業種のみであった。そのため、業種別新店舗情報有無判別モデルの学習データは、上記の36業種のみを用いた。以上より、テストデータとして用いる業種も同様に36業種とする。本実験で用いる各判別モデルのテストデータ一覧を表4に示す。

図7 実験計画

表2 学習対象とした業種一覧(判別対象は灰色)

業種	大業種	業種	大業種	業種	大業種	業種	大業種	業種	大業種
食堂	飲食業	オランダ料理店	飲食業	書店	小売業	ロックア	サービス業	会館・ホール	サービス業
ファミリーレストラン	飲食業	スライス料理店	飲食業	新聞店	小売業	駐車場	サービス業	清瀬・歌劇場	サービス業
洋食レストラン	飲食業	スペイン料理店	飲食業	ホームセンター	小売業	動物病院	サービス業	ブール	サービス業
回転寿司	飲食業	ロシア料理店	飲食業	文具・事務用品店	小売業	クリーニング	サービス業	スポーツクラブ・ジム	サービス業
ファーストフード	飲食業	フランス料理	飲食業	アニメショップ	小売業	コインランドリー	サービス業	キャンプ場	サービス業
喫茶	飲食業	バー	飲食業	おもちゃ屋	小売業	写真屋	サービス業	卓球場	サービス業
アミューズメントショップ	飲食業	スナック・クラブ	飲食業	ゲームソフト販売店	小売業	書画	サービス業	ゴルフ場	サービス業
たこ焼き	飲食業	百貨店	小売業	CD・DVD・ビデオ・レコード店	小売業	結婚式場	サービス業	野球場	サービス業
日本料理・焼石・料亭	飲食業	生活協同組合	小売業	ゴルフショップ	小売業	ベトナム美容室	サービス業	パッチングセンター	サービス業
魚料理店	飲食業	コンビニエンスストア	小売業	テニスショップ	小売業	ベトナムホテル	サービス業	ゴルフ練習場	サービス業
居酒屋	飲食業	スーパーマーケット	小売業	サーフショップ	小売業	ベトナム公園	サービス業	テニスコート	サービス業
おでん料理店	飲食業	ケーキショップ	小売業	スキーショップ	小売業	病院	サービス業	モータースポーツショップ	サービス業
お好み焼き店	飲食業	パン屋	小売業	スキューバダイビングショップ	小売業	介護施設	サービス業	サーキット場	サービス業
焼き肉・ホルモン	飲食業	酒屋	小売業	スポーツショップ	小売業	老人ホーム	サービス業	インターネットカフェ	サービス業
京料理店	飲食業	食肉店	小売業	釣具・釣具店	小売業	児童館	サービス業	旗館	サービス業
丼	飲食業	鮮魚店	小売業	アウトドア用品	小売業	福祉施設	サービス業	旅館	サービス業
串	飲食業	青果物店	小売業	自動車販売	小売業	保育所	サービス業	民宿	サービス業
鍋	飲食業	くだもの店	小売業	自動車部品	小売業	大学	サービス業	公共の宿	サービス業
寿司	飲食業	菓子店	小売業	オートバイ販売・修理	小売業	高等学校	サービス業	ペンション	サービス業
鉄板焼店	飲食業	米店	小売業	オートバイ部品・用品	小売業	中学校	サービス業	ユースホステル	サービス業
てんぷら料理店	飲食業	呉服店	小売業	タイヤ販売	小売業	小学校	サービス業	ホテル	サービス業
焼鳥店	飲食業	生地店	小売業	自転車店	小売業	高等専門学校	サービス業	ラブホテル	サービス業
とんかつ店	飲食業	寝具店	小売業	家具店	小売業	養護学校	サービス業	ビジネスホテル	サービス業
ふく料理	飲食業	アウレットモール・ショップ	小売業	神具店	小売業	ろう学校	サービス業	健康ランド	サービス業
うどん	飲食業	紳士服店	小売業	オーディオ販売店	小売業	盲学校	サービス業	温泉	サービス業
そば	飲食業	ブティック・アウトリチュール	小売業	電器店	小売業	幼稚園	サービス業	銭湯	サービス業
スパゲティ店	飲食業	婦人服店	小売業	パソコンショップ・コンピュータ用品	小売業	専修学校	サービス業	風俗	サービス業
ラーメン	飲食業	子供服店	小売業	書画・骨董品	小売業	専門学校	サービス業	エステ	サービス業
ステーキハウス	飲食業	ベビー・マタニティ用品	小売業	買物	小売業	ビジネススクール	サービス業	理容店	サービス業
ハンバーグ料理店	飲食業	紳士洋品店	小売業	リサイクルショップ	小売業	法律スクール	サービス業	美容院	サービス業
ビザハウス	飲食業	婦人洋品店	小売業	雑貨店	小売業	研修センター	サービス業	公園	サービス業
カレー	飲食業	靴店	小売業	たばこ店	小売業	趣味各種学校・教室	サービス業	遊園地・テーマパーク	サービス業
中華	飲食業	アक्सサリー	小売業	喫茶店	小売業	自動車教習所	サービス業	アミューズメントパーク	サービス業
イタリア料理	飲食業	服飾	小売業	ガラス店	小売業	塾・進学塾・予備校	サービス業	ビルヤード場	サービス業
インド料理店	飲食業	コンタクトレンズ	小売業	カメラ店	小売業	道場	サービス業	ボウリング場	サービス業
朝鮮料理	飲食業	時計店	小売業	アダルトショップ	小売業	図書館	サービス業	カラオケ	サービス業
イギリス料理	飲食業	めがね店	小売業	ベトナムショップ	小売業	科学館	サービス業	映画館	サービス業
インドネシア料理店	飲食業	宝飾・貴金属店	小売業	宝くじ売り場	小売業	博物館	サービス業	パチンコ・スロット店	サービス業
ドイツ料理店	飲食業	スポーツフアングス	小売業	郵便局・郵便室	サービス業	美術館	サービス業	ゲームセンター	サービス業
タイ料理店	飲食業	金物・刃物店	小売業	レンタルCD・DVD・ビデオ・レコード店	サービス業	資料館	サービス業	ダンスホール	サービス業
台湾料理店	飲食業	日用品雑貨店	小売業	レンタカー	サービス業	水族館	サービス業	ディスコクラブ	サービス業
ベトナム料理店	飲食業	菓物・菓子店	小売業	レンタルバイク	サービス業	動物園	サービス業	遊覧・観光船	サービス業
中南米料理	飲食業	化粧品販売	小売業	レンタルサイクル	サービス業	植物園	サービス業	競艇場	サービス業
欧風料理店	飲食業	ガソリンスタンド	小売業	弁護士・弁護士・会計士・各種事務所	サービス業	公民館・集会所	サービス業	釣船	サービス業

表3 学習データ一覧

対象モデル	学習データの概要
出店傾向モデルの学習データ	住所録に記載されている大阪市の商業施設、約 140,000 件の住所情報
業種判別モデルの学習データ	業種ごとにブログ記事 400 件、合計 88,000 件のブログ記事
業種別新店舗情報有無判別モデルの学習データ	業種ごとに新店舗情報を含むブログ記事を 200 件、合計 7,200 件の新規店舗情報を含むブログ記事 (36 業種) 業種ごとに既知の店舗情報を含むブログ記事を 200 件、合計 7,200 件の既知の店舗情報を含むブログ記事 (36 業種)

表4 テストデータ一覧

対象モデル	テストデータの概要
出店傾向モデルのテストデータ	住所録から業種ごとに 50 件ずつ、合計 1,800 件の店舗情報 (36 業種)
業種判別モデルのテストデータ	業種ごとにブログ記事 50 件ずつ、合計 1,800 件のブログ記事 (36 業種)
業種別新店舗情報有無判別モデルのテストデータ	2008 年 12 月 1 日から 2010 年 12 月 1 日までの期間中に投稿された新店舗情報を含むブログ記事 50 件 (36 業種が混在して 50 件) 既知の店舗情報を含むブログ記事 450 件、住所情報を含むブログ記事 500 件 (36 業種に限らない)、合計 950 件

この時、出店傾向モデルの学習データやテストデータである店舗情報として、住所録に記載されている情報から収集した。また、業種判別モデルと業種別新店舗情報有無判別モデルの学習データやテストデータであるブログ記事の収集方法は、まず、店舗情報や新店舗の開店情報を整理した住所録やポータルサイトから、店舗名、業種、出店日、住所や電話番号の店舗情

報の組み合わせを抽出する。次に、地域名や業種名などの不要語を含む店舗名を対象にクリーニング手法 [25] を適用し、検索に適した店舗名を抽出する。最後に、店舗名、業種、出店日、住所や電話番号の組み合わせを Yahoo! Blog 検索に入力し、共起検索することで、既知の店舗および新店舗に関する記述を扱ったブログ記事を収集する。以上の流れで収集した学習データとテストデータをにより構築した各判別モデルを用いて、第 7.1 節で概説した 3 つの事前実験と 2 つの評価実験を行う。

なお、実験システムを構築するにあたり、形態素解析器として mecab [27]、係り受け解析器として cabocha [28]、SVM のライブラリとして LibSVM [29]、LDA のライブラリとして GibbsLDA++ [30]、PCA のライブラリとして Lapack [31] をそれぞれ使用した。

8.3 事前実験

8.3.1 事前実験 1：PCA と LDA の次元圧縮精度比較

(1) 実験の概要

本実験では、既存研究で用いられている PCA と LDA の両手法によって次元圧縮を行い抽出した素性をそれぞれ SVM に入力し、その判別精度を比較する。この時、LDA による次元圧縮では、次元数として抽出するトピック数を入力する必要がある。そこで、事前実験

として、出店傾向モデルのテストデータ、業種判別モデルのテストデータと業種別新店舗情報有無判別モデルのテストデータを入力とし、各判別モデルに最適なトピック数を決定する実験を行った。実験では、トピック数を8, 16, 32, 64, 128, 256, 512と変動させ、その時の出店傾向モデル、業種判別モデルと業種別新店舗情報有無判別モデルの誤差平均を検証する。なお、LDAのパラメータは初期値への依存性があると考えられるため、各トピックに付き1回から5回の次元圧縮処理を行い、各条件での誤差平均を式(1)により求めた。

事前実験によって得られた各モデルの最適なトピック数を用いてLDAにより次元圧縮した素性の優位性を検証するため、PCAにより次元圧縮した素性との比較実験を行った。比較実験には、PCAにより次元圧縮した素性とLDAにより次元圧縮した素性のそれぞれを用いて学習した3種の各モデルにテストデータを入力し、その判別精度を算出する。なお、判別精度の評価指標として、出店傾向モデルと業種判別モデルでは、正解の業種が36業種中1位になる確率、3位以内に含まれる確率、5位以内に含まれる確率と10位以内に含まれる確率の4段階で評価した。また、業種別新店舗情報有無判別モデルの判別精度の評価指標として、適合率、再現率とF値を用いた。

(2) LDAの最適トピック数の検証実験結果

事前実験であるLDAのトピック数の決定実験における実験結果を図8と図9に示す。ここで、図8はLDAのトピック数と出店傾向モデルの判別精度の関係を表すグラフであり、図9はLDAのトピック数と業種判別モデルの判別精度の関係を表すグラフである。

図8と図9から、出店傾向モデルと業種判別モデルの両方で、トピック数が増加するに従い誤差が改善され、ある一定のトピック数で収束することがわかる。出店傾向モデルではトピック数が64の時に、そして、業種判別モデルではトピック数が32の時に誤差が収束した。したがって、実験結果より、出店傾向モデルにおけるLDAの最適なトピック数は64であり、業種判別モデルの最適なトピック数は32である。なお、業種別新店舗情報有無判別モデルについても同様の実験を行ったところ、およそその業種において32トピックで収束していることが確認できた。そのため、業種別新店舗情報有無判別モデルの最適なトピック数は32である。

(3) PCAとLDAの次元圧縮精度の比較実験結果

本項(2)の事前実験の結果を用い、出店傾向モデル、業種判別モデルと業種別新店舗情報有無判別モデルについてPCAで次元圧縮した場合とLDAで次元圧縮

図8 LDAのトピック数と出店傾向モデルの誤差平均の関係

図9 LDAのトピック数と業種判別モデルの誤差平均の関係

表5 出店傾向モデルにおける判別結果

	1位の確率	3位以内の確率	5位以内の確率	10位以内の確率
PCA	0.22	0.52	0.78	0.97
LDA	0.25	0.55	0.80	1.00

表6 業種判別モデルにおける判別結果

	1位の確率	3位以内の確率	5位以内の確率	10位以内の確率
PCA	0.39	0.76	0.90	0.99
LDA	0.44	0.83	0.95	1.00

表7 業種別新店舗情報有無判別モデルにおける判別精度

	PCA	LDA
適合率	0.70	0.71
LDA	0.64	0.68
F値	0.67	0.69

した場合実験の判別精度の結果を表5から表7に示す。

表5、表6と表7から、出店傾向モデル、業種判別モデルと新店舗情報有無判別モデルのすべてにおいて、PCAによって次元圧縮した素性よりもLDAによって次元圧縮した素性を用いた方が判別の精度が高いことがわかる。以上の結果から、本提案手法で採用したLDAの次元圧縮は、PCAによる次元圧縮より多様な要素を含む学習データの次元圧縮に有用であると判断できた。

8.3.2 事前実験2：形態素n-gramのnの最適値

(1) 実験の概要

本実験では、単語の接続性を考慮する時に使用する形態素n-gramのnの値を決定するため、新店舗情報有無判別モデルのテストデータを入力とし、nの値を1、2、3、4、5と変動させ、各値により構築した新店舗有無判別モデルの判別精度を検証する実験を行った。

(2) 実験結果

nの値と新店舗情報有無判別モデルの判別精度の関係を表8と図10に示す。

図10から、形態素n-gram手法のnの値が大きくなるに従い、適合率が上昇した。一方、nの値が3よりも大きくなると、再現率が低下した。これは、形態素の接続数を増加するに従い、素性が複雑になるため、学習データ中の素性とテストデータ中の素性の一致率が低下したのが原因である。また、nの値が3よりも小

表8 nの値と新店舗情報の有無判別モデルの判別精度

	1-gram	2-gram	3-gram	4-gram	5-gram
適合率	0.71	0.75	0.80	0.85	0.90
LDA	0.68	0.72	0.82	0.56	0.38
F値	0.69	0.73	0.81	0.67	0.54

図10 nの値と業種別新店舗情報有無判別モデルの判別精度の関係

さい場合、十分な文脈を考慮することができないため、新店舗用語として誤抽出していることが分かった。このことから、新店舗情報の有無を判別する時、形態素n-gram手法を用いて素性を構築することで、誤抽出を軽減できることが分かった。なお、図10より、nの値が3の時、F値が最も高い値を示したため、本実験の形態素n-gram手法のnの値は3とした。

8.3.3 事前実験3：特徴語に基づく確率と地理的特性に基づく確率の統合時における業種別の重み付け

(1) 実験の概要

本実験では、特徴語により取得した新店舗情報の有無確率と地理的特性により取得した新店舗情報の有無確率を統合する時に、業種によって重みを変動させる必要性と、業種別の重みの値を検証するため、業種別に店出傾向モデルによって求めた業種分類の誤差平均と業種判別モデルによって求めた業種分類の誤差平均の相関を検証する実験を行った。

(2) 実験結果

事前実験1において算出した店出傾向モデルと業種判別モデルの業種判別精度の結果を用いて、単語情報と地理情報による業種判別精度の関係を図11に可視化した。図11の縦軸は業種判別モデルの業種分類の誤差平均であり、横軸は店出傾向モデルの業種分類の誤差平均である。縦横どちらの軸においても値が小さいほど、業種の判別精度が高いことになる。図11から、婦人服、ホテル、ゲームセンターや風俗といった業種は、地理情報と単語情報の両方で業種判別精度が高いことが分かった。菓子、魚料理店や焼鳥店といった業種は、地理情報と単語情報の両方で業種判別精度が低いことが分かった。また、鍋、カレー、そばやラーメンといった業種は、地理情報を用いた業種判別精度が低く、単語情報を用いた業種判別精度が高いことが分かった。一方、バー、居酒屋や弁理士・弁護士・会計士・各種事務所といった業種は、地理情報を用いた業種判別精度が高く、単語情報を用いた業種判別精度が低いことが分かった。

これらの結果から、地理情報と単語情報のスコアを統合する時、全ての業種の重みを均一とするのではなく、業種ごとに重みを変える必要があることが分かった。本実験では、図11に示した店出傾向モデルと業種判別モデルの業種判別の誤差平均の比を用いることで、各業種の重みを決定した。

以上の3つの事前実験の結果を用いて、2つの評価実験を行う。

図11 単語情報と地理情報による業種分類精度の関係

8.4 評価実験

8.4.1 評価実験1：新店舗情報の有無判別精度の評価実験

(1) 実験の概要

本実験では、新店舗情報の有無判別時における本研究の各提案項目の有用性を評価する。実験に用いるシステムについて表9に示す。表9に示すように、比較システムB, D, EとFでは、業種別の単語出現傾向の違いを考慮して学習を行った。比較システムC, DとFでは、形態素n-gram手法により、単語の接続関係を考慮して学習を行った。比較システムEとFでは、地理的特性を考慮して学習を行った。

本実験では、これら6件の比較システムを用いて評価実験を行い、それぞれの判別結果を比較した。まず、比較システムAとB, CとDを比較することで、業種別に新店舗情報有無判別モデルを構築する有用性を検証する。次に、比較システムAとC, BとD, EとFを比較することで、単語の接続関係を考慮する有用性を検証する。最後に、比較システムBとE, DとFを比較することで、地理的特性を考慮する有用性を検証する。

(2) 実験結果

各比較システムの新店舗情報の有無判別精度の比較を表10に示す。実験の結果、比較システムAとB, CとDでは、AよりもB, CよりもDの適合率、再現率とF値のすべてが優れた結果を導き出した。したがって、新店舗情報の有無判別において、業種別に学習データを作成することは、有用であることが分かった。これは、業種別に学習しなかった場合、業種が増加するに従い素性の数が膨大となり、スパースネスの問題が生じるためである。

比較システムAとC, BとD, EとFでは、AよりもC, BよりもD, そして、EよりもFの適合率、再現率とF値のすべてが優れた結果を導き出した。したがって、新店舗の有無判別において、単語の接続関係を考慮することは、有用であることが分かった。これは、単語の接続関係を考慮することで「毎朝11時にオープン」といったノイズを除去できたためである。

比較システムのBとE, DとFでは、再現率では大きな変化は見られなかったが、地理的特性を考慮するこ

表9 実験に用いるシステム

	業種別の単語出現傾向の違いを考慮	単語の接続関係を考慮	地理的特性を考慮
比較システムA	×	×	×
比較システムB	○	×	×
比較システムC	×	○	×
比較システムD	○	○	×
比較システムE	○	×	○
比較システムF	○	○	○

表10 判別精度の比較

	適合率	再現率	F 値
比較システムA	0.71	0.68	0.69
比較システムB	0.76	0.86	0.76
比較システムC	0.80	0.82	0.81
比較システムD	0.84	0.86	0.85
比較システムE	0.83	0.76	0.79
比較システムF	0.89	0.88	0.89

とで、BよりもE、DよりもFの適合率が高い結果となった。これは、出店傾向モデルにより、誤抽出を防止できたからである。判別結果の中身を確認すると、地理的特性を考慮することで、単語情報のみでは誤抽出するデータがいくつか確認できた。誤抽出の例としては、新しく創設されるパチンコサイトの広告が挙げられる。また、単語情報のみでは、「オープン」や「新規スタート」といった新店舗特徴語に反応して誤抽出するが、住所の地理的特性を解析することで、パチンコサイトを運営するオフィスの住所の位置には、パチンコ店が出店する確率が極端に低いと判別できたため、地理的特性を考慮した比較システムEとFでは誤検出しなかった。さらに、地理的特性を考慮することで、Webページに含まれる住所に関連する業種を絞り込むことができるため、業種別に学習することに加えて、不要な語を除去することができ、新店舗の有無判別性能が向上した。しかし、すべての業種において地理的特性を考慮することで新店舗有無判別の精度が向上するわけではなく、コンビニや飲食店などの業種については、出店確率の算出精度が低下するため、地理的特性の有用性は低いことが分かった。

以上の実験結果より、本研究の3つの提案である、業種別に単語の出現傾向を学習する手法、単語間の連関係を考慮して学習する手法と地理的な特性を考慮して情報を評価する手法の各種法の有用性を実証した。

8.4.2 評価実験2：新店舗情報の抽出精度の評価実験 (1) 実験の概要

本実験では、業種別新店舗情報の有無判別処理により正しく判別された新店舗情報を含むWebページを対象とし、店舗名、住所、電話番号、出店日や業種を適切に抽出することができるかについて検証し、本提案手法の最終出力を評価する。実験の方法としては、本提案手法が各要素を省略している文章から店舗名を取得できていることを検証するため、地物に関する単語を用いた構文解析手法によって店舗名を取得するシステムと比較した。これは、住所と業種に加え、店舗名を適切に取得できれば、電話番号や出店日は調査可能であると考えたためである。評価指標としては、適合率を用いた。なお、再現率とF値を評価指標としないのは、属性情報の抽出時において、網羅性ではなく情報の正確性が重要となるためである。

(2) 実験結果

属性情報の抽出精度を表11に示す。また、システムの実行結果を図12に示す。実験結果を解析すると、構文解析を用いたシステムでは、各要素が省略されている文から店舗名を抽出することができなかった。一

表11 属性情報の抽出精度

	適合率
構文解析を用いたシステム	0.55 (24/44)
提案システム	0.86 (38/44)

図12 システムの実行結果(GoogleMap APIを用いて実装)<<http://code.google.com/intl/ja/apis/maps/>>

方、住所や新店舗特徴語と共起するというシンプルなルールを用いた本提案手法により抽出した店舗名は、誤判別を1件も含まなかった。この結果より、本提案手法の店舗名抽出手法は有用であることが実証できた。しかし、店舗名を「あの店」「例の店」などの代名詞で表記したり、「お店」などの一般名詞で表記された場合は、店舗名を推測できないため、属性情報を取得することはできなかった。この問題に対しては、店舗名以外の属性情報や特徴語などを検索エンジンに入力することで補完できると考えられる。

以上の実験結果から、本提案手法は、大量のノイズ情報に対して頑健で、かつ効率的な新店舗情報の抽出が可能であることを実証した。

9. おわりに

本研究では、幅広い情報源から高精度に新店舗情報を取得することを目的とし、Web上に投稿された自然言語で記述されたコンテンツを研究対象として、自然言語を解析する手法の「業種ごとに出現する単語の傾向が異なることに起因する課題」「文脈により単語の意味的な役割が異なることに起因する課題」と「CGMの信頼性に起因する課題」の3つの課題を解決する手法を提案し、その有用性を確認した。

実証実験では、前述の提案手法を考慮した場合と考慮しなかった場合の比較実験を行い、その結果から本提案手法が有用であることを明らかにした。また、住所情報を含む膨大なWebページの中から新店舗情報を含むWebページを約9割の精度で抽出できたことか

ら、本提案手法を使用することで、実地踏査の補助的な情報を自動的に取得できたり新店舗の位置を用いた実地踏査における効率的な巡回路の提案に応用できると考えられる。このため、本提案手法は土木工事や地理情報サービスの基盤情報である地図の更新への貢献が期待できる。

その一方で、本提案手法では、住所情報が完全である情報源のみを対象としているが、CGM上には膨大な量の住所情報を含まない新店舗情報があると考えられる。そこで、今後は、住所情報を含まない情報源に擬似的に住所情報を付与する手法について研究し、地理的特性を考慮する本提案手法の有用性の向上を図る予定である。また、地理的特性の一般的な有用性については検証ができたが、各地理的特性間の相関や地理的特性ごとの有用性の詳細な検証についても行う予定である。

謝辞

本論文は、私立大学戦略的研究基盤形成支援事業(2008年～2012年選定プロジェクト「セキュアライフ創出のための安全知循環ネットワークに関する研究」)より助成を受け、その成果を公表するものである。ここに記して、感謝の意を表する。

参考文献

- [1] 吉成雄一郎：カーナビゲーション、写真測量とリモートセンシング、日本写真測量学会、Vol.43, No.2, pp.13-15, 2004. 5.
- [2] 越智繁雄：防災情報の共有と活用のあり方、地質と調査、全国地質調査業協会連合会、No.2, pp.2-7, 2010.6.
- [3] 堀宗朗、田中謙吾、Sobhaninejad, G., 市村強、小国健二：自然災害シミュレーションのためのGISから都市モデルへの汎用的データ変換、土木学会論文集A, 土木学会、Vol.66, No.1, pp.1-12, 2010.1.
- [4] 国土地理院：地理空間情報活用推進基本計画、国土交通省、2008.4., < <http://www.gsi.go.jp/common/000003539.pdf> >, (入手 2010.12.1)
- [5] 経済産業省商務情報政策局：地理空間情報に関する政策パッケージ～G空間プロジェクト～：経済産業省、2008.7., < http://www.meti.go.jp/press/20080703007/02_seisaku_p.pdf >, (入手 2010.12.1)
- [6] 物部寛太郎、田中成典、古田均、岡崎真也：Web情報からの地物名自動抽出システムの研究開発、地理情報システム学会講演論文集、地理情報システム学会、Vol.14, pp.281-284, 2005.9.
- [7] 物部寛太郎、田中成典、古田均、岡崎真也：Webリソースを利用した地物間の関連度評価に関する基礎的研究、地理情報システム学会講演論文集、地理情報システム学会、Vol.14, pp.285-288, 2005.9.
- [8] 小林卓弥、大島裕明、小山聡、田中克己：レビューアの生活圏の特定に基づくCGM評判情報評価、データベースシステム研究会研究報告、情報処理学会、Vol.2007, No.65, pp.221-216, 2007.7.
- [9] 村山紀文、南野朋之、奥村学：メタデータ付与のための住所録自動生成、自然言語処理研究会研究報告、情報処理学会、Vol.2004, No.73, pp.41-47, 2004.7.
- [10] 矢野宏実、目良和也、相沢輝明：嗜好を考慮した評判情報検索手法、自然言語処理研究会研究報告、情報処理学会、Vol.2004, No.108, pp.165-170, 2004.11.
- [11] 相良毅、喜連川優：Webからの効率的な新規店舗の発見・登録支援手法、情報処理学会論文誌 データベース、情報処理学会、No.48, SIG11 (TOD34), pp.49-57, 2007.6.
- [12] 樫山武浩、田中成典、木下智宏、増満光、西江将男：Webからの新店舗情報の自動収集に関する研究、情報処理学会全国大会講演論文集、情報処理学会、Vol.71, No.1, pp.591-592, 2009.3.
- [13] 大阪市計画調整局都市再生振興部情報化施策担当：施設情報マップ、マップナビおおさか、大阪市計画調整局、< <http://www2.wagamachi-guide.com/osakacity/index.asp?dtp=1> >, (入手 2010.12.1)
- [14] 小山照夫：日本語テキストからの複合語用語抽出、情報知識学会誌、情報知識学会、Vol.19, No.4, pp.306-315, 2010.2.
- [15] 森信介、山地治：日本語の情報量の上限の推定、情報処理学会論文誌、情報処理学会、Vol.38, No.11, pp.2191-2199, 1997.11.
- [16] マピオン：Mapion, < <http://www.mapion.co.jp/station/27/> >, (入手 2010.12.1)
- [17] マピオン：バス停情報、Mapion, < <http://www.mapion.co.jp/s/q=%E3%83%90%E3%82%B9%E5%81%9C/t=spot/?cate=M12001&area=27126> >, (入手 2010.12.1)
- [18] マピオン：高速道路IC情報、Mapion, < <http://www.mapion.co.jp/s/q=%E9%AB%98%E9%80%9F%E9%81%93%E8%B7%AF%E9%BC%A9%E9%BC%A3/t=spot/?cate=M08> >, (入手 2010.12.1)
- [19] 大阪市：町丁目別人口・世帯、平成22年国勢調査(速報)、< http://www.city.osaka.lg.jp/shisei_top/category/1756-7-1-0-0.html >, (入手 2010.12.1)
- [20] 大阪市：大阪市の区別概況、平成17年度国勢調査、< <http://www.city.osaka.lg.jp/keikakuchosei/cmsfiles/contents/0000015/15569/H17-kokuseityousachomokutyukan-0shi.xls> >, (入手 2010.12.1)
- [21] 国土交通省：都道府県地価調査、国土交通省地価公示、< <http://www.land.mlit.go.jp/landPrice/AriaServlet?MOD=0&TYP=0> >, (入手 2010.12.1)
- [22] Blei, M. D., Ng, Y. A. and Jordan, I. M.: Latent Dirichlet Allocation, Journal of Machine Learning Research, MIT Press, Vol.3, pp.993-1022, 2003.1.
- [23] 正田備也、喜安千弥、宮原末治：文書クラスタリングのための潜在的ディリクレ配分法による次元圧縮、データベースシステム研究会研究報告、情報処理学会、Vol.2007, No.65, pp.381-386, 2007.7.
- [24] Cortes, C. and Vapnik, V.: Support-Vector Networks, Machine Learning, Springer, Vol.20, No.3, pp.273-297, 1995.9.
- [25] 相良毅、喜連川優：住所情報を用いた店舗名称のクリーニング手法、電子情報通信学会論文誌、電子情報通信学会、Vol.J91-D, No.3, pp.531-537, 2008.3.

- [26] 澁木猛, 秋山祐樹, 柴崎亮介: デジタル地図と電話帳データの時空間統合による店舗及び事業所の立体分布変動モニタリング手法, 日本建築学会計画系論文集, 日本建築学会, Vol.73, No.626, pp.789-793, 2008.4.
- [27] 工藤拓: MeCab: Yet Another Part-of-Speech and Morphological Analyzer, <http://mecab.sourceforge.net/>, (入手 2011.2.10.)
- [28] 工藤拓: CaboCha: Yet Another Japanese Dependency Structure Analyzer, <http://chasen.org/~taku/software/cabocha/>, (入手 2011.2.10.)
- [29] Chih-Chung, C. and Chih-Jen, L.: LIBSVM-A Library for Support Vector Machines, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>, (入手 2011.2.10.)
- [30] Xuan-Hieu, P. and Cam-Tu, N.: GibbsLDA++, <http://gibbslda.sourceforge.net/>, (入手 2011.2.10.)
- [31] Univ. of Tennessee, Univ. of California, Berkeley, Univ. of Colorado Denver, and NAG Ltd: LAPACK-Linear Algebra Package, <http://www.netlib.org/lapack/>, (入手 2011.2.10.)

(2010年12月21日 受付)
(2011年6月7日 採録)

【問い合わせ先】

〒569-1095 大阪府高槻市霊仙寺町2-1-1
関西大学大学院総合情報学研究科
寺口 敏生
TEL: 072-690-2151
E-mail: k102036@edu.kut.ac.jp

著者紹介

てらぐち としお
寺口 敏生 [非会員]

2007年関西大学総合情報学部卒業。2009年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了。現在、関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程在学中。修士(情報)。WebデータマイニングおよびGISに関連する研究に従事。

にしえ まさお
西江 将男 [非会員]

2009年関西大学総合情報学部総合情報学科卒業。2011年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了。修士(情報)。WebデータマイニングおよびGISの研究に従事。

たなか しげのり
田中 成典 [正会員]

1986年関西大学工学部土木工学科卒業。1988年関西大学大学院工学研究科土木工学専攻博士課程前期課程修了。同年、(株)東洋情報システム(現在、TIS)に入社。人工知能に関する研究受託開発業務に従事。1994年関西大学総合情報学部専任講師として着任。1997年助教授、2004年教授、2006年から学生センター副所長、現在に至る。2002年8月から1年間、カナダのUBCにて客員助教授。博士(工学)。専門は知識工学と社会基盤情報学。CAD/CG、GIS/GPS、画像処理及びWebソリューションビジネスに関する研究に従事。2000年(株)関西総合情報研究所を起業、設立当初から現在まで取締役会長。2006年(株)フォーラムエイトの顧問に就任。現在に至る。建設省土木研究所CAD製図基準検討委員会委員長、土木学会土木情報システム委員会幹事長、同委員会土木CAD小委員会委員長、ISO/TC184/SC4国内委員などを歴任。現在、国土交通省日本建設情報総合センター社会基盤情報標準化委員会委員、同委員会CAD/データ連携小委員会委員長、土木学会情報利用技術委員会副委員長。主に、ISOに準拠したCAD製図基準とCADデータ交換基盤の開発に従事。

Research on Automatic Collection of New Store Information Utilizing Web Resource and Geographic Information

by

Toshio TERAGUCHI, Shigenori TANAKA and Masao NISHIE

Abstract :

With diffusion and development of information-related technologies in recent years, geographic space information services, typified by car navigation, are widespread. However, provision of map information, which is the basis of geographic space information services, depends on on-site investigations conducted manually ; therefore, there is a problem that the map information cannot catch up with the changes in the real-space information, such as the opening of a new store. Therefore, the authors have been conducting a study on text mining to automatically collect store information such as store names, addresses, and the industry types from the natural language, with focus on the information about the opening of new stores on the Web.

However, when conducting the text mining using only the keywords related to new store opening, we were faced with two problems. One is that the differences in the word appearance tendency by industry type or the combination of words before and after one another become a noise, which causes the accuracy of the collected store information to fall. Another problem is that since the precision of the information sources that may contain typing errors or false information is not guaranteed, it is difficult to judge whether the data is correct or not if there is inconsistency in the contents. This study aims to overcome these problems by mining using the keywords, and by considering the fact that there is a strong correlation between the industry type of a store and the geographic characteristics of the store location. By experimenting, we verified that this study considering the geographic characteristics of the store location was effective to collect the new store information.

Keywords : geographic space information service, provision of map information, collection of new store information, text mining, geographic characteristics

Contact Address : **Toshio TERAGUCHI**

Graduate School of Informatics, Kansai University
2-1-1, Ryozenji-cho, Takatsuki City, Osaka 569-1095, JAPAN
TEL : 072-690-2151
E-mail : k102036@edu.kutc.kansai-u.ac.jp