# Webからの新店舗情報の自動収集に関する研究

樫山武浩† 田中成典‡ 木下智弘† 増満光† 西江将男‡
関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

## 1．はじめに

近年，情報処理技術の発展により，カーナビゲーションやインターネット地図などの空間情報を用いたサービスが普及[1]している．しかし，空間情報における属性情報の整備は，未だ進んでいない問題[2]がある．そのため，属性情報をWebから収集[3]する研究に注目が集まっている．既存研究では，住所録の店舗情報を基にWebページを収集し，住所録に存在しない店舗情報を収集する研究[4]が行われている．しかし，既存研究では，Tableタグのように，特定のHTMLタグ構造を持つWebページのみに対応しているため，掲示板やBlogから店舗情報を収集できない問題がある．そこで，本研究では，あらかじめ新店舗情報が含まれるWebページを教師データとして収集し，One-Class SVM（Support Vector Machine）[5]により新店舗情報が含まれるWebページの特徴を学習し，判別することで，特定のHTMLタグ構造に依存しない新店舗情報の収集手法を提案する．

## 2．研究の概要

本研究では，Webから新店舗情報を自動で収集する手法を提案する．システムの概要を図1に示す．本システムは，1）新店舗教師データ収集機能，2）新店舗教師データ学習機能，3）新店舗情報有無判定機能で構成される．入力データは，大量のWebページのテキストとし，出力データは，新店舗情報が含まれるWebページのテキストとする．

### 2．1　新店舗教師データ収集機能

本機能では，新店舗情報を含むWebページを収集する．まず，新店舗教師データとして新店舗情報の名称，出店日と出店場所を収集する．次に，新店舗の名称を検索キーワードとし，出店日までに作成や編集されたWebページを収集する．そして，収集したWebページに新店舗の名称，出店日と出店場所に関する情報が正しく含まれるか判別する．その際，出店日と出店場所については，完全一致ではなく，言い換え表現を考慮する．最後に，収集したWebページに新店舗情報が含まれるかを目視で確認する．

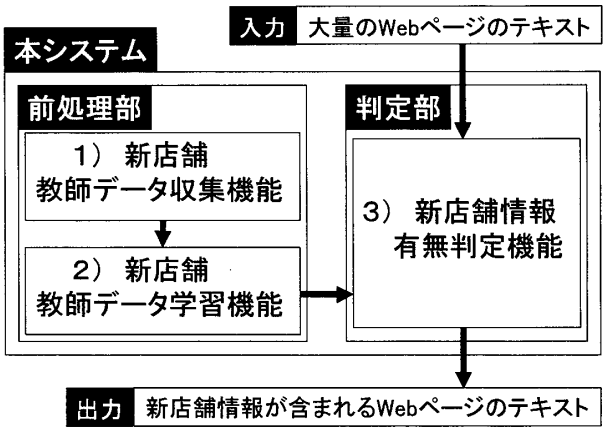


図1　システムの概要

### 2．2　新店舗教師データ学習機能

本機能では，新店舗情報を含むWebページの特徴を学習する．新店舗情報の学習として，2クラス以上の識別器を用いると，新店舗情報を含まないWebページをどのように定義するのかが問題となる．そこで，本研究では，1クラスで学習可能なOne-Class SVMを採用する．まず，文章内には，単語の言い換え表現が含まれている可能性があるため，シソーラス辞書を用いて単語を同義語ごとにまとめる．次に，Webページの特徴語を抽出するため，単語のTF-IDF値を算出する．そして，単語とWebページの共起行列に対して特異値分解を行うことで，潜在的意味情報を保持したまま次元圧縮された単語の特徴ベクトルを算出する．最後に，算出した特徴ベクトルをOne-Class SVMにより学習する．

### 2．3　新店舗情報有無判定機能

本機能では，入力された大量のWebページのテキストに新店舗の情報が含まれるか判定する．学習したOne-Class SVMから解を求め，新店舗の情報が含まれると判定された場合は，そのWebページのテキストを出力する．

Research for Collecting New Store Information on WWW
† Takehiro Kashiyama, Tomohiro Kinoshita, Hikaru Masumitsu
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho, Takatsuki-shi, Osaka 569-1095, Japan
‡ Shigenori Tanaka, Masao Nishie
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho, Takatsuki-shi, Osaka 569-1095, Japan

## 3．システムの実証実験と考察

本システムの有用性を実証するため，特定のHTMLタグ構造を持たないBlogテキストから新店舗情報を抽出できるか検証する実験を行った．

### 3．1　実証実験

まず，新店舗情報の定義を明らかにするため，出店してから何日後までの情報が新店舗の情報であるか，学生31人に対してアンケート調査を行った．アンケート調査の結果を表1に示す．アンケート調査の結果，出店してから1ヵ月後までの情報を新店舗情報として採用した．本実証実験では，グルメ情報検索サイトぐるなびが提供する店舗情報を基に新店舗情報が含まれるBlogテキスト100件と，含まれないBlogテキスト20件を収集した．新店舗情報が含まれるBlogテキストの内，90件を学習データとし，残りの10件と新店舗情報が含まれないBlogテキストを合わせ，合計30件の実験データを作成した．実験データを本システムに入力し，新店舗情報として判定したテキストを目視で確認し，正確性と網羅性の指標として適合率，再現率を算出し，総合評価として適合率と再現率の調和平均値F値を算出した．

### 3．2　結果と考察

本システムでBlogテキスト内の新店舗情報を収集した結果を表2に示す．表2より適合率よりも再現率が高いことが確認できる．これは，“毎朝9時に開店”の“開店”といった学習データの特定の特徴語に過剰反応し，実際の新店舗情報よりも多く，新店舗情報として誤抽出したためである．この問題は，IF-THENルールを採用し，誤抽出された新店舗情報を除去することで改善できると考えられる．また，実証実験の総合評価であるF値は，7割を上回る結果となった．このことから，既存研究では対応していなかったBlogテキストなどの特定のHTMLタグ構造を持たないWebページから新店舗情報を収集できたと考えられる．

新店舗情報を含むWebページから抽出した特徴語を図2に示す．ただし，特徴語における頻度傾向を分析するために，15件以上出現した名詞語句を対象とした．図2より新規に開店を表す語句の中で，“オープン”の頻出が顕著に見られた．これは，その他の“出店”や“開業”などの開店を表す語句と比較して，年齢や性別に関係なく，ユーザの視覚に直接訴えやすいからだと考えられる．

## 4．おわりに

本研究では，特定のHTMLタグ構造に依存しない新店舗情報の収集手法を提案した．そして，

表1　アンケート調査の結果

| | 票数 |
|---|---|
| 出店当日まで | 1票 |
| 3日後まで | 2票 |
| 1週間後まで | 5票 |
| 1ヵ月後まで | 23票 |

表2　実験結果

| | 値 |
|---|---|
| 適合率 | 0.6667 |
| 再現率 | 0.8000 |
| F値 | 0.7272 |

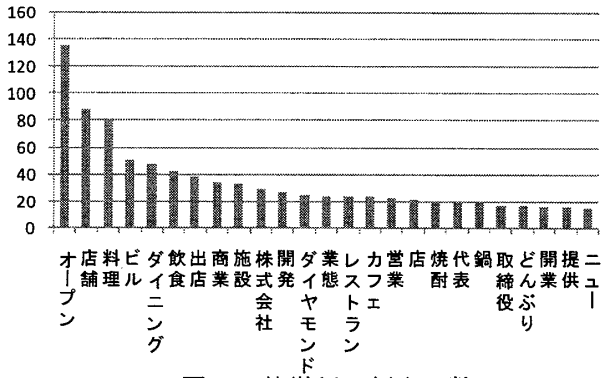


図2　特徴語の頻出回数

本システムの実証実験の結果，Blogテキストなどの特定のHTMLタグ構造を持たないWebページから新店舗情報を収集することができた．しかし，本研究では，新店舗情報の確認として，Webページのテキストに新店舗情報が存在するかを判定しているため，最終的に人の目視で新店舗の名称，出店場所などの情報を確認しなければならない．また，時系列やランドマークの特性として，新しく駅ができた場合，駅周辺に新店舗が開店するといった，新店舗情報の収集に関係する新たな条件が考えられる．今後は，空間情報の属性情報として，新店舗情報以外に必要だと考えられる属性情報を検討し，空間情報の充実を目指す．

### 参考文献


[1] 総務省：平成20年度版情報通信白書，ぎょうせい，2008.7.

[2] 物部寛太郎，田中成典，古田均，加藤佑一，野中広茂：WWW自動探索による電子地図の属性情報自動抽出システムの研究開発，土木情報利用技術論文集，土木学会，Vol.13，pp.95-102，2004.10.

[3] Arasu, A. and Garcia, H.：Extracting Structured Data from Web Pages, ACM SIGMOD Conference 2003, ACM, pp.337-348, 2003.6.

[4] 相良毅，喜連川優：Webからの効率的な新店舗の発見・登録支援手法，情報処理学会論文誌，情報処理学会，Vol.48，No.SIG 11，pp.49-57，2007.6.

[5] Manevitz, L. and Yousef, M. : Journal of Machine Learning Research, 2002.3.