E-010

Wikipedia から作成した辞書によるブログのカテゴリ分類

Text Categorization For Blog Entries Using Wikipedia Data

田村 直之† 伊藤 直之† 西川 侑吾† 中川 修† 新堀 英二†
Naoyuki Tamura Naoyuki Ito Yugo Nishikawa Osamu Nakagawa Eiji Shinbori

１．はじめに

ブログや SNS などの CGM(Consumer Generated Media)と呼ばれる新たなメディアの普及により,誰でも手軽に自らの日常や意見を発信することが一般的になった.CGM で発信される情報は消費者の貴重な生の声でもあるため,これらの情報を有効活用することは企業のマーケティング活動において重要である.

CGM からマーケティングに有用な情報を抽出するためには,ブログを内容に応じたカテゴリ（例えば,映画カテゴリ，スポーツカテゴリ，料理カテゴリなど）に分類することが大切である．分類されたカテゴリごとに特徴語抽出や評価情報抽出をすることで，興味が似通っている一定層のブロガー集団の特徴を掴むことが可能となる．

そこで，本発表では Wikipedia から自動的に作成した名詞辞書を用いて，ブログを自動カテゴリ分類する手法を提案する．本手法の特徴は，一般的な機械学習によって分類器を作成しブログを分類する手法に比べて，人手による下準備のコストを大幅に下げた点にある．自動分類させたい各カテゴリに対応した Wikipedia の一覧項目名を管理しておけば，常に最新の用語をカテゴリ分類手法に反映可能な為，日々新たな用語が現れるブログ等の CGM テキストの分類に対応が可能である．本発表ではブログのカテゴリ分類に必要な Wikipedia から自動的に辞書を作成する手法と辞書を使って，ブログをカテゴリ分類する手法を提案する．

２．Wikipedia から作成する辞書について

本手法の特徴は,Wikipedia から名詞を大量に抽出し,作成した名詞辞書を使用してカテゴリ分類することである．ここでは,そのカテゴリ分類に必要な名詞辞書を Wikipedia から作成する手法について説明する．

2009 年 6 月現在 Wikipedia において日本語のエントリーは約 59 万のエントリーが存在し,日々追加・更新されている．Wikipedia から辞書を作成する理由は,体系化された情報が大量に存在することに加えて,最新の情報が常に誰かの手によって更新されることもある．

Wikipedia のエントリーの中でも我々が着目したのは,エントリー名が「○○一覧」のものである．この「○○一覧」エントリーの本文には,エントリー名に属する固有名詞が箇条書きで記述されていることが多く，Wikipedia から自動的に大量の固有名詞辞書を構築する為に有用なページである．この「○○一覧」の編集ページ本文に現れる Wiki 記法の法則性に着目して,大量の上位-下位概念の名詞辞書を自動構築する．

辞書構築手法の流れを図 1 に示す.まず, 小説家一覧,野球選手一覧.野菜一覧などの Wikipedia 上にある全ての「○○一覧」ページをリスト化する．このリストを用いて,リストに掲載の全ての Wikipedia 編集ページから,Wiki 記法の法則性を利用して上位-下位項目のペアを大量に抽出する．

例えば，図 1 では,上位項目として「日本の小説家一覧」下位項目として,「阿井景子」,「青野聴」などを抽出している．抽出したペアを上位項目の一覧を表す名前の辞書に登録する．

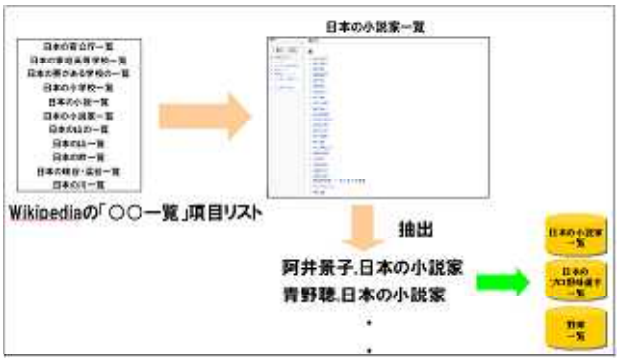

図 1 Wikipedia からの辞書作成手法

ただし,Wikipedia の「○○一覧」から抽出できる名詞は固有名詞のみである．ブログのカテゴリを判断する際に当然ながら固有名詞だけで判断するのではなく、各カテゴリに関連した固有名詞以外の名詞（以下本稿では一般名詞と記す）も判断の材料となる．この一般名詞を抽出する手法を図 2 に示す.

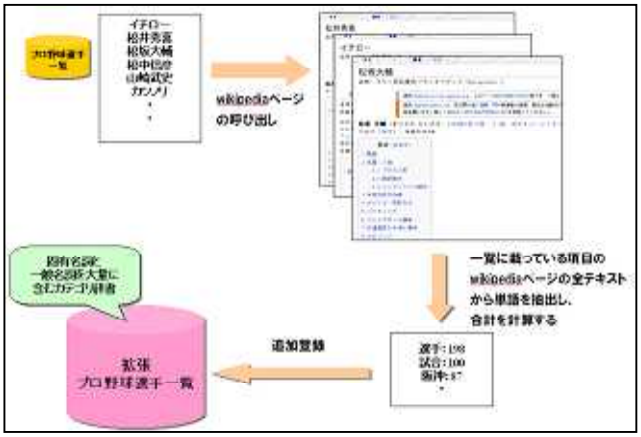

図 2　一般名詞抽出を辞書に登録する為の手法

† 大日本印刷株式会社 情報コミュニケーション研究開発センター ユビキタスメディア研究所

「○○一覧辞書」に登録されている固有名詞の全ての Wikipedia エントリー本文を形態素解析し，名詞を抽出し，集計する．ただし,名詞の連続で構成される固有名詞（例：高野豆腐,松井秀喜など）の単語を正しい区切りで抽出する為に,さらに詳しい名詞の品詞分類(chasen の品詞分類を利用)によって,隣り合っていた名詞の連結可否を判断し,名詞の区切り位置の判断を行なう工夫をして名詞の抽出を行なう．（本予稿では詳細の説明は省略）

このような方法で抽出し，集計した一般名詞の上位 20%を○○一覧の固有名詞辞書に追加登録することで関連する一般名詞を取得する．この辞書を拡張○○一覧辞書とする．

最後に,どのカテゴリにも出現する一般的な語（例えば,「毎日」「以前」など）を辞書内から削除する処理を行なう．一般的な語を排除するには,全ての拡張○○辞書を作成した後,各名詞の idf を計算し idf が閾値を越える語に関しては,辞書から削除する方法で単語の選別を行なう．

このような一般的な語は各カテゴリの特徴を表す語とはなりえない為，極力排除することが必要である．

以上の手法で Wikipedia からブログのカテゴリ分類に利用する名詞辞書を作成する．

3．提案手法

ここからは,ブログをカテゴリ分類する提案手法を説明する．提案手法の概要を図 4 に示す．

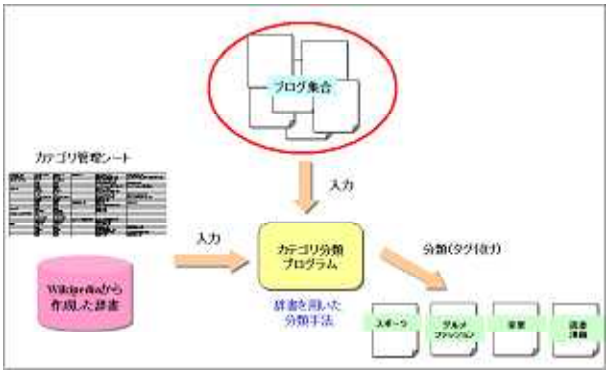

図 4 カテゴリ分類手法概要

本手法はカテゴリ分類プログラムと 3 つの入力によって構成される．入力の 1 つ目は「カテゴリ管理シート」と呼ばれるもので，分類したいカテゴリの種類と Wikipedia から作成した辞書の「○○一覧」との対応付けをする表である.この表を人手で作成する.各カテゴリにいくつの「○○一覧」を対応付けても良い．カテゴリ管理表の例を図 5 に示す．

カテゴリ名	対応する一覧(1)	対応する一覧(2)
音楽	ミュージシャン一覧（グループ）	ミュージシャン一覧（個人）
ゲーム	Wiiのゲームタイトル一覧	ニンテンドーDSのゲームタイトル一覧
本	日本の小説家一覧	小説家一覧
映画	日本の映画作品一覧	映画作品一覧
食	食材の一覧	料理と菓子の一覧
TV	日本の男優一覧	日本の女優一覧
野球	日本のプロ野球選手一覧	読売ジャイアンツの選手一覧
サッカー	日本のサッカー選手一覧	セリエA（サッカー）
ファッション	ファッションブランド一覧	男性モデル一覧

図 5 カテゴリ管理表の例

2 つめの入力は 2 章で説明した拡張○○一覧辞書である．

3 つめの入力は，カテゴリごとに分類したいブログの集合である．以上の 3 つを入力すると，自動的に各ブログのカテゴリ度を表すスコアが計算され,計算結果から各カテゴリに分類される．ただし，分類されるカテゴリは 1 つとは限らない為，タグのようにカテゴリ名はブログごとに複数個付与される．

続いて,ブログにカテゴリ名を付与するカテゴリ分類プログラムの処理について説明する．処理概要を図 6 に示す．

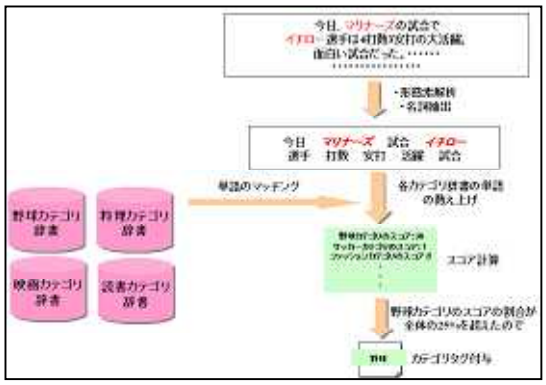

図 6　カテゴリ分類プログラムの処理概要

まず,ブログの本文を形態素解析する．形態素解析結果から固有名詞・一般名詞を抽出する． 名詞の抽出は，辞書構築の際と同様に，名詞の連結を考慮して抽出する．

そして,ブログ本文から抽出した名詞と各カテゴリ辞書に登録されている一般名詞・固有名詞をマッチングさせ、マッチした名詞数を数え,これを各カテゴリのスコアとする．固有名詞のほうが属するカテゴリにを判断する材料になりやすいため，固有名詞は 2 倍のスコアとする．

ブログごとに各カテゴリのスコアを計算し，スコアが全体の 25%を超えていたカテゴリをこのブログのカテゴリとする．全てのカテゴリのスコアが閾値以下のブログは「未分類」に分類する．

以上の手法を用いて,ブログをカテゴリ分類する．

このように本手法では,カテゴリ管理表を作成する以外に人手での作業は発生しない．カテゴリ管理表も一度作成してしまえば,ほとんど更新の必要なく Wikipedia からカテゴリ分類に必要な最新の用語辞書が自動的に作成されるため,一度管理表を作ってしまえば,長期間運用可能なブログのカテゴリ分類システムを構築可能である．

4.まとめ

Wikipedia から作成したカテゴリごとの固有名詞・一般名詞ごとの辞書を用いて,ブログのカテゴリ分類を行なう手法を提案した．本手法を用いることで,大量の教師データつきブログを用意するなどの人手によるカテゴリ分類に必要な下準備のコストを下げることができる．また Wikipedia から辞書を作成する為,新しい語にも対応したカテゴリ分類を常に行なうことが可能になった．

参考文献

[1]平野 他．日本語ブログの自動分類，社団法人情報処理学会研究報告,2005-NL-170,2005.

[2]古林 他．ブログ記事の自動分類により消費者意識の側面を捉える試み,NRI 技術開発,2006