Wikipediaカテゴリを用いたブログ著者の得意分野プロファイリング

野田 陽平* 清田 陽司† 中川 裕志‡

概 要

本研究では，ブログ投稿者の得意分野のプロファイリングを行う．ブログ投稿者の中には，特定の分野に精通し，非常に有用なブログ記事を投稿する者も存在する．しかし，膨大なブログサイトの中から特定分野に特化したブログサイトを的確に抽出することは，明確な分類体系が存在しないブログにおいては困難である．また，投稿者の投稿記事がひとつの分野のみに特化しているとは限らないため，投稿者を単一の分類にマッピングすることは困難である．そこで，非常に有用な集合知である Wikipedia の構造に直接ブログ記事をマッピングし，投稿者の投稿行動の傾向を観察することで，あらゆる分野の専門家をブロガーの中から発見することを目指す．

1 はじめに

総務省の調査によると，2008 年 1 月現在，ブログサイトは約 1,690 万件であり，総記事数は 13 億 5,000 万件にのぼる [1]．これらの膨大なブログ記事からユーザーが興味のあるブログ記事やブログサイトの情報を得るためには，情報抽出技術が必要不可欠である．本研究では，特定の分野に注力して記事を投稿しているブロガーを発見することを目的とする．ある特定の分野に関する記事を大量に投稿している専門性の高いブロガーを提示することで，ユーザーが興味を持つ分野のブログを発見することができる．

ブログなどの文書の分類は，あらかじめ用意した辞書を用いた手法が一般的であり，辞書に含まれた各単語をベクトルの組成として使用し，TF/IDF 値などで重みづけを行って分類をしている．この手法ではひとつの記事はひとつのカテゴリに一対一対応で分類されるが，一人のブロガーが複数のカテゴリに跨ってブログ記事を投稿している場合もあるため，一人のブロガーがひとつのカテゴリに割り当てられるのは適当ではない．本研究では Wikipedia[2] の項目とカテゴリのグラフ構造に直接ブログ記事をマッピングすることで，ブロガーがどの分野に注力してブログ記事を投稿しているかを観察するという手法をとる．日本語版 Wikipedia の項目数は 2008 年 6 月 25 日に 50 万項目を超えており，多くの編集者により日々更新が続けられている非常に有用な集合知であり，各分野に関する網羅性も高い．各項目とブログ記事との紐付けは，Wikipedia のタイトルの情報を固有表現として用い，ブログ記事を検索する方法で行う．紐付けられた Wikipedia カテゴリなどの構造をもとに，各項目に関するエキスパート性のあるブロガーを発見する．

2 関連研究

本研究に関連する研究として，テキストのカテゴリ分類に関する研究と，「アルファブロガー」と呼ばれるエキスパート性を備えたブログ投稿者の発見に関する研究がある．

橋本らは，あらかじめ Domain Dictionary [3] を用意し，ブログ記事の分類を行っている [4]．Domain Dictionary は，あらかじめ設定した SPORTS，MEDIA などの 20 個の各ドメインに対して，そのドメインに属する特徴的な用語を少数設定し，それを用いてその他の 30,000 語の日本語の基本単語を各ドメインに割り当てて作成している．ブログのカテゴライズは記事に出現する単語の IDF を計算し，Domain Dictionary と対応させることで行っている．ブログ記事中に存在する未知語に関しては，Web 検索のスニペットや Wikipedia の記事を参考にしてドメイン予測を行っている．また，Timothy Weale ら [5] は Wikipedia のカテゴリ情報を用いてテキストの分類を行っている．Weale らは，Wikipedia のカテゴリ情報を利用して，positive/negative のタグが付与された文章を SVM と決定木学習により分類している．

*東京大学大学院学際情報学府 noda @ r.dl.itc.u-tokyo.ac.jp
†東京大学情報基盤センター図書館電子化研究部門 kiyota @ r.dl.itc.u-tokyo.ac.jp
‡東京大学情報基盤センター図書館電子化研究部門 n3 @ dl.itc.u-tokyo.ac.jp

アルファブロガーの発見に関する研究には，Nakajima ら [6] の研究がある．Nakajima らは，重要なブロガーを，Agitator と Summarizer の 2 種類に分類し，リンク構造や流行トピックに関する言及のタイミングなどの情報から重要なブロガーを発見している．また，松永らの研究 [7] はある特定のキーワードに対して言及したブロガーではなく，あるコミュニティにおいての話題を網羅的に扱うブロガーを発見することを目的としている．松永らの研究では，ニュース性のあるキーワードを抽出し，それらを基にキーワードクラスタを作成し，そのクラスタとブログを紐付けることにより，ブログコミュニティを抽出している．

3 システム構成

本研究においては，辞書に含まれた単語を素性とした単語ベクトルを用いた文書の分類は行わない．本研究では，Wikipedia のカテゴリ関係を用い，Wikipedia に含まれる各項目に関連するブログを紐付ける手法をとる．また，Wikipedia にはニュース性のあるキーワード以外にもさまざまな分野に関するキーワードが含まれているため，網羅的にエキスパートブロガーの発見を行うことができると考えられる．

本研究において構築したシステムは，下記のとおりである．

1. ブログ記事の収集
2. Wikipedia 項目とカテゴリ情報の取得
3. ブログ投稿者の得意分野の抽出

3.1 ブログ記事の収集

各ブログ記事については，ブログサービスの新着記事 RSS を定期的に解析し，タイトルや本文，投稿時刻，ブロガーなどの情報を取得し，HyperEstraier[9] を用いて N-gram インデックスを作成した．

ブログ記事は 2008 年 7 月 14 日から収集し，ブログ記事数は 2008 年 9 月 5 日現在で 280 万記事，ブロガー数は 213,964 である．

3.2 Wikipedia 項目とカテゴリ情報の取得

Wikipedia の全データは Wikipedia のダウンロードページから XML ファイルとしてダウンロードすることができる [10]．本研究では 2008 年 5 月 17 日付けの Wikipedia 日本語版のデータを用い，Wik-IE[11] を用いてカテゴリやタイトルのグラフ構造を表すデータを作成した．Wik-IE が出力するデータファイル (edge ファイル、node ファイル) は TSV 形式であり，NLP 若手の会有志によって実装された汎用シソーラス探索ライブラリ [12] のフォーマットに準拠している。本研究で使用したデータには，48,833 カテゴリ，491,858 項目が含まれている．edge ファイル、node ファイルの抜粋例を表 1，表 2 に示す．

表 1: edge ファイル

id-from	id-to	relation
52220	680877	hypernym
54817	52220	redirect
362802	52220	redirect
410798	52220	redirect

表 2: node ファイル

id	title	kind
52220	IPod	leaf
54817	Ipod	redirect
362802	アイポッド	redirect
41078	IPod photo	redirect
680877	category:IPod	node

また，表 1, 表 2 を加工し，一つのデータにまとめ，MySQL に保存した．保存したデータの例を，表 3 に示す．このようにして，Wikipedia のカテゴリとタイトルを表現した．

表 3: カテゴリ，タイトル間関係

id-from	category	id-to	title
680877	category:IPod	52220	IPod
680877	category:IPod	54817	Ipod
680877	category:IPod	362802	アイポッド
680877	category:IPod	41078	IPod photo

3.3 ブログ投稿者の得意分野の抽出

本研究では，Wikipedia のタイトル名と各ブログ記事とのマッピングを行った．表 3 の title 列をキーにし

て，3.1 で作成した記事インデックスに対して検索を行うことで，各タイトル項目が含まれているブログ記事を抽出した．集計は一日一回インデックスの作成が終了した際に行った．集計項目の一覧を下記に示す．

- Wikipedia 項目 ID
- Wikipedia 項目タイトル
- ブログ記事 URL
- ブロガー ID
- ブロガー名
- 集計日付

以上のようにブログ記事を Wikipedia の項目をキーにして集計することで，各ブロガーの各項目ごとのブログ記事の投稿頻度の特徴を得ることができる．

たとえば，図 3.3 のようなカテゴリ構造があった際に，Wikipedia の項目である IPod や Zune などをキーとしてブログ記事を検索する．検索にヒットしたブログ記事を，それぞれの Wikipedia の項目にマッピングし，上記の集計項目のデータを作成する．これにより，Wikipedia の各カテゴリに対するブログ投稿者の投稿活動の注力度合を取得することができ，たとえば「アップルコンピュータ」や「携帯型音楽プレーヤー」を得意分野とするブロガーを発見することができる．

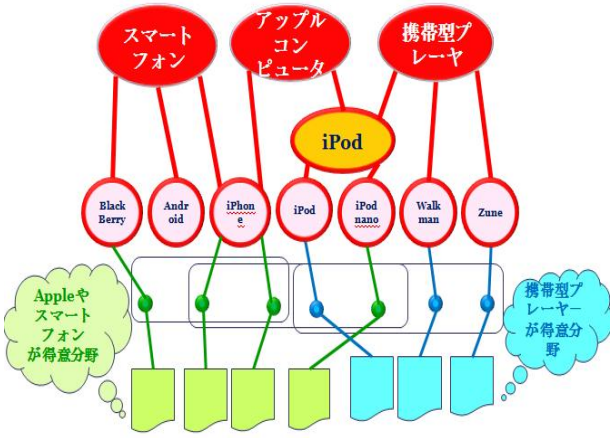

図 1: Wikipedia タイトルとブログ記事の紐付けの例

4 アプリケーション

本研究では，Wikipedia のグラフ構造を用いた検索システムを開発した．ユーザーは Wikipedia の項目名による検索を行い，その項目に関する記事を多く投稿しているブロガーの情報を得ることができる．検索窓には suggest 機能を付与し，ユーザーが Wikipedia の項目名を容易に表記揺れなく検索できるようにした．検索結果には検索項目が含まれる上位カテゴリが表示され，そのカテゴリに含まれる項目に関するブログ記事を横断的に投稿しているブロガーに関する情報を得ることができる．また，検索項目と Wikipedia 内のグラフ構造上でつながりのある項目も提示している．

たとえば，Wikipedia の項目である”iPhone”を検索項目として検索すると，”iPhone”に関連したブログ記事を多く投稿しているブロガーの情報が表示される．”iPhone”が含まれる上位カテゴリには，”Category:IPod”と”Category:スマートフォン”が存在する．”Category:スマートフォン”について注力して記事を投稿しているブロガーを検索したい場合は，カテゴリ名である”Category:スマートフォン”をクリックすると，カテゴリに含まれているすべてのタイトルに関連したブログ記事を多く投稿しているブロガーの情報が表示される．

カテゴリにより差はあったものの，各カテゴリに含まれる項目について注力してブログ記事を投稿しているブロガーを抽出することができた．しかし，近年 bot を使用して大量に投稿されているスプログが多数含まれており，スプログの排除は非常に大きな課題の一つである．

5 おわりに

本研究では，ブログ記事を Wikipedia の項目にマッピングするというアプローチで，ブロガーの得意分野を推定した．

今後は，時系列での各項目に対する言及頻度などをもとに，ブログ投稿行動のばらつきの情報を利用する予定である．話題性のある項目に関しては，その項目を得意分野としないブロガーがその項目に関するブログ記事を投稿し，一時的なバーストが発生する可能性がある．流行に関わらずコンスタントに特定項目に関するブログ記事を投稿しているブロガーは，よりエキスパート性の高いブロガーであると考えることもでき，今後はその点についての情報も含めた得意分野推定を行う予定である．

また，スプログへの対応も重要な課題である．bot による大量投稿は，本研究の推定手法において大きなノイズとなる．これらを排除することで，より正確な

ランキングが可能となる．

参考文献

[1] 総務省,”ブログの実態に関する調査研究の結果”, 2008.7

[2] Wikipedia,http://ja.wikipedia.org/

[3] Chikara Hashimoto, Sadao Kurohashi, ”Construction of Domain Dictionary for Fundamental Vocabulary”, ACL HLT 2007

[4] Chikara Hashimoto, Sadao Kurohashi, ”Blog Categorization Exploiting Domain Dictionary and Dynamically Estimated Domains of Unknown Words”, ACL HLT 2008

[5] Timothy Weale, ”Utilizing Wikipedia Categories for Document Classification”, Computer Science and Engineering Department. OSU-CISRC-4/08-TR14, http://www.cse.ohio-state.edu/ weale/

[6] Shinsuke Nakajima,Junichi Tatemura,Yoichiro Hino,Katsumi Tanaka, ”Discovering Important Bloggers based on Analyzing Blog Threads”, WWW 2005

[7] 松永拓, 平手勇宇, 山名早人, ”キーワードの出現に基づくブログコミュニティ抽出とオピニオンリーダーの発見”, DEWS 2007

[8] MeCab, http://mecab.sourceforge.net/

[9] HyperEstraier, http://hyperestraier.sourceforge.net/

[10] Wikipedia ダウンロードページ, http://download.wikimedia.org/jawiki/

[11] Wik-IE,http://sourceforge.jp/projects/wik-ie/

[12] 清田 陽司, 阿辺川 武, 吉田 稔, 田村 悟之, 坂井 哲, 増田 英孝, ”汎用シソーラス探索ライブラリの開発”, 言語処理学会 第 14 回年次大会 発表論文集 (PD1-3),pp.257-260,2008.