社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93（11）
2009－ＮＬ－189（11）
2009／1／23

単語の出現ホストを利用した未知語の分野推定

下山 剛司† 秋岡 明香†† 村岡 洋一‡
† 早稲田大学大学院 基幹理工学研究科 †† 電気通信大学 ‡ 早稲田大学理工学術院

概要

近年，Web からオピニオンリーダと呼ばれる影響力のある人物を抽出する研究が Web マーケティングの分野や知識抽出の分野で注目されている．オピニオンリーダの抽出には，キーワードを使用する手法があるが，キーワードが未知語であったとき，何の分野におけるオピニオンリーダか分別するのが難しいという問題があった．本研究では，本文の要約が書かれていることが期待されるタイトルタグを解析し，各単語の出現ホストにおける類似性によって，未知語の分野を推定する手法を提案する．

Guessing Category of Unknown Words by using Web Hosts of Word Appearance

Takeshi SHIMOYAMA† Sayaka AKIOKA†† Yoichi MURAOKA‡
† WASEDA University Graduate School of Fundamental Science and Engineering
†† The University of Electro-Communications
‡ WASEDA University Faculty of Science and Engineering

Abstract

Recently, the research that extracts powerful people that is called an opinion leader from web have attracted attention. There is a method to extract opinion leader by using keyword. But, this method have a problem that it is difficult to sort opinion leader by category in case of unknown keyword.

In this paper, we propose that guessing category of unknown word by similarity of web hosts of word appearance.

1. はじめに

世の中では，"ビリーズブートキャンプ"といった商品名や"のだめカンタービレ"といったドラマ名，空気が読めないを意味する"KY"と言った略語など，話題性のあるキーワードが絶えることなく出現している．このようなキーワードの発信者は世間への影響力が強い，もしくは，情報収集能力が高いことが予想されるため，オピニオンリーダと呼ばれており，近年 Web マーケティングの分野や知識抽出の分野でオピニオンリーダの抽出の研究が盛んである．例えば，松永ら[1]は，ある期間までは出現頻度の低く，イベント性と重要性を兼ね備えるキーワードを"ニュース性のあるキーワード"と定義し，オピニオンリーダの抽出を行っている．しかし，キーワードが未知語であったとき，オピニオンリーダを分類できないという問題点があった．本研究では，この問題を解決するために，未知語の分野を推定する手法を提案する．

2. 関連研究

未知語の分野を推定する研究は，既に行われている．例えば，Hashimoto ら[2]は，基本語ドメイン情報を構築することによって，未知語の分野推定を試みている．基本語ドメイン情報とは，形態素解析システム JUMAN[3]に登録されている既知の名詞，動詞，計 26658 語を表 1 に示す 12 の分野に分類した情報である．Hashimoto らはこの情報と検索エンジンを用いて，未知語の分野を推定するという手法を提案している．Construction of Domain Dictionary for Fundamental Vocabulary[2]では，未知語をクエリーとして Yahoo 検索 API に投げたとき，上位 30 位以内にヒットしたページを解析し，各ドメインへの分類を行うことによって，未知語の分野を推定するという手法を用いている．この手法により，67.5%の精度で未知語を推定することができたとこの論文内では報告されている．

表1:ドメイン体系表

文化・芸術	家庭・暮らし
レクリエーション	料理・食事
スポーツ	交通
健康・医学	教育・学習
科学・技術	メディア
ビジネス	政治

3. 問題点

関連研究では，未知語を1つのカテゴリに分類することを目的としていた．しかし，未知語の中には，様々な見方によって，複数のカテゴリに分類できるものもある．例えば，“Suica”という電子マネーは，機能としては”交通”というジャンルにカテゴライズできるが，”科学・技術”とも解釈できる．つまり，未知語に対して複数のカテゴリを付与する必要性がある．

また，多義の未知語に関しても考慮する必要がある．例えば，”ヌンチャク”という未知語に対して，これがゲーム機である Wii のコントローラを示すのか，武器としてのヌンチャクを示すのかにより，カテゴリが大きく異なってくる．特に，キーワードに基づいてオピニオンリーダを抽出する際には，この問題は致命的である．

本稿では，これらの問題点を解決するために，未知語が出現するホストの類似関係から，多義の未知語が何を意味しているのか特定した後，分野推定を行い，複数のカテゴリを付与する手法を提案する．

4. 提案手法

本稿では，単語の出現ホストに注目し，出現ホストの類似性から，単語のクラスタリングを行い，既知の言葉を利用して，未知の言葉の分野を推定する．提案手法は，単語における出現ホストのベクトル化，単語のクラスタリング，Wikipedia を用いたクラスタへのカテゴリ付与，検索エンジンを用いたカテゴリ絞込みの 4 ステップから成る．これらについて以下で詳細に説明する．

4.1 単語における出現ホストのベクトル化

このステップでは，各単語における出現ホストを表 2 のように抽出し，各単語ごとに，出現ホストのバイナリベクトルを形成する．

表 2:出現ホスト分布表

	a.com	b.com	c.com	d.com	…
Wii	1	1	0	1	…
PS3	1	1	0	1	…
PHP	0	1	1	0	…
Ruby	0	1	1	0	…

4.2 単語のクラスタリング

このステップは，単語間距離の算出，MDS を用いた 2 次元マッピング，k-means 法を用いたクラスタリングの計 3 つのサブステップから成る．これらについて以下で詳細に説明する．

4.2.1 単語間距離の算出

このサブステップでは，単語間の距離を算出する．出現ホスト数を n，単語間距離を l，閾値を k としたとき，式(1)より，求められる距離を，単語間距離として求める．例えば，表 2 の単語間距離を算出すると，表 3 のようになる．

$$l = \begin{cases} 1 & (k > n) \\ 0 & (k \leq n) \end{cases} \quad (1)$$

表 3:単語間距離

	Wii	PS3	PHP	Ruby
Wii	-			
PS3	0	-		
PHP	1	1	-	
Ruby	1	1	0	-

4.2.2 MDSを用いた2次元マッピング

このサブステップでは，多次元尺度構成法(MDS)と呼ばれる手法を用いて，各単語の 2 次元マッピングを行う．本稿では，ニュージーランドの Auckland 大学の Ross Ihaka と Robert Gentleman により作られた統計解析向け言語，R 言語[4]で用意されている，計量多次元尺度法を用いて，2 次元マッピングを試みた．計量多次元尺度法とは，主座標分析[5]とも呼ばれ，距離マトリクスを $D_{m\times m}$ としたとき，式(2)より得られる，$Z_{m\times m}$ を固有ベクトルの点として，2次元マッピングする手法である．

$$Z_{ij} = \frac{1}{2}\left(\sum_{i=1}^{m}\frac{d_{ij}^{\,2}}{m} + \sum_{j=1}^{m}\frac{d_{ij}^{\,2}}{m} - \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{d_{ij}^{\,2}}{m^2} - d_{ij}^{\,2}\right) \quad (2)$$

4.2.3 k-means 法を用いたクラスタリング

このサブステップでは，二次元にマッピングされた各単語を k-means 法[6]を用いて k 個のクラスタに分ける．k-means 法について以下で詳しく述べる．

k-means 法は，データ数を n としたとき，以下の4つのステップから構成される．(i = 1,2,…n)

1. 各データ d_i の所属クラスタをランダムに決定する．

2. 各クラスタの中心の座標群 Z_k を求める

3. 各データ d_i は，クラスタの中心座標群 Z_k に最も近いクラスタに再割り振りを行う．

4. 1～3 のステップを繰り返し，各データ d_i の所属クラスタに変化がなかったとき，終了する

4.3 Wikipedia を用いたクラスタへのカテゴリ情報付与

このステップでは，既知の単語のカテゴリ情報を Wikipedia を用いて取得することで，各クラスタにカテゴリ情報を付与する．具体的には，図 1 のように，"ピラティス"という既知の単語と，"ビリーズブートキャンプ"という未知の単語が同クラスタ内に存在しているとき，Wikipedia によって"ピラティス"という単語に付与されているカテゴリタグを，クラスタのカテゴリ情報として付与する．

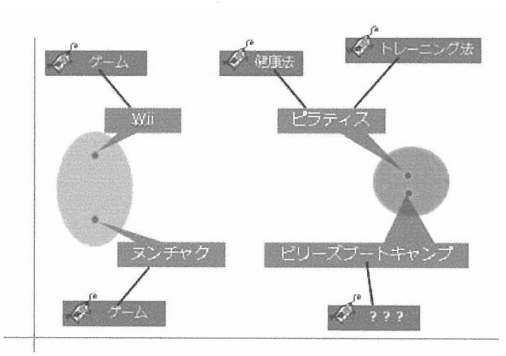

図1:Wikipedia を用いたクラスタへのカテゴリ付与例

4.4 検索エンジンを用いたカテゴリの絞込み

このステップでは，検索エンジン Google を用いて，未知語と所属クラスタのカテゴリ情報の共起数を調べることで，カテゴリ情報の絞込みを行い，未知語のカテゴリを決定する．

ここでは，未知語を X ,所属クラスタのカテゴリ情報を C_i とし，検索エンジンへのクエリーとヒット数の関係を，表 4 のように定義する．また，式(3)を共起ポイントとして定義し，閾値 θ を越えたとき，カテゴリ情報 C_i を未知語 X のカテゴリ情報として付与する．

表 4:検索エンジンへのクエリーとヒット数の関係

クエリー	ヒット数
X	s 件
C	t 件
X AND C	u 件

$$\frac{u}{t} + \frac{u}{s} > \theta \quad (\theta\text{は閾値}) \quad (3)$$

5. 実験

提案手法による，未知語の分野推定実験を行った．これらについて以下で詳細に説明する．

5.1 実験方法

時系列データを用いたタイトルタグからの新語抽出法の提案[7]内で，使用した 5825 万 2033 件のデータを用いて実験を行った．本実験では，本文の要約が書いてあると期待されるタイトルタグより，表 5 で示すような未知語を約 1000 個抽出し，その未知語の出現ホストを利用することで，30 のクラスタに分別することに成功した．これらを利用し，未知語の分野推定を行った．

表 5:未知語抽出例

ウェブリブログ
アニメイト
アンチエイジング
Wii
ドラゴンクエストモンスターズジョーカー
サイボウズ
ニート
CanCam
チャンドンゴン
クォンサンウ
コミュファ
アビシニアン
QUALIA
トップバリュ
モナー
スイーツ
リクナビ
ピラティス
ゴスロリ
ポッドキャスティング
デュエルマスターズ
レーシング
テニプリ
DODA
アコーディア
スパイダーマン
LOMO
ドクターシーラボ
アイマス
カレ
…

5.2 実験結果

未知語のクラスタリング結果，分野推定結果を以下の述べる．

5.2.1 クラスタリング結果

出現ホストを利用した単語のクラスタリング結果を図 2 に示す．

また，単語のクラスタリング結果の一部を表 6，表 7 にまとめる．

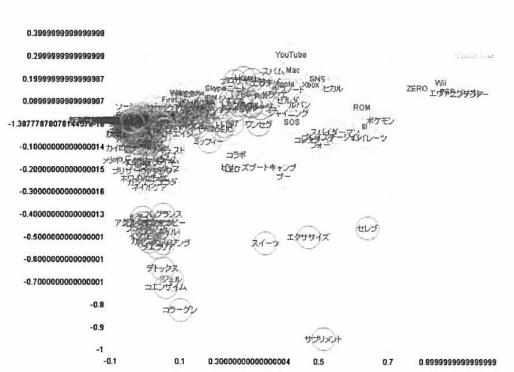

図 2: 出現ホストを利用した単語のクラスタリング

表 6,クラスタ要素（クラスタ No.1）

ジェル
マカ
プラセンタ
デトックス
コラーゲン
フレグランス
ルテイン
グルコサミン
コエンザイム
ネイル
イソフラボン
アロマテラピー
プエラリア
フコイダン
アスタキサンチン
カルニチン

表 7,クラスタ要素(クラスタ No.2)

GBA
ウィー
ローゼンメイデン
エクセル
シャア
レゴ
NDS
ジブリ
ヤフオク

5.2.2 未知語の分野推定結果

表 6,表 7 のクラスタにおける分野推定結果と 2008 年 11 月時点で wikipedia に掲載されていた各単語のカテゴリ情報を表 8,表 9 にまとめる.

表 8:未知語の分野推定結果(クラスタ No.1)

未知語	提案手法で推定された分野	wikipedia のカテゴリ
ジェル	化粧品 健康食品 ハーブ	
マカ	健康食品 栄養素 環境技術	質問があるページ 出典を必要とする記事 アブラナ科 ハーブ
プラセンタ	化粧品 健康食品 栄養素	妊娠 女性生殖器
デトックス	栄養素 化粧品 健康食品	医学関連のスタブ項目 デトックス
コラーゲン	化粧品 健康食品 栄養素	出典を必要とする記事 生物学関連のスタブ項目 タンパク質 健康食品 化粧品
フレグランス	化粧品 健康食品 ハーブ	
ルテイン	健康食品 栄養素 タンパク質	有機化合物 色素 栄養素
グルコサミン	栄養素 健康食品 アミノ酸	健康食品 アミノ糖 単糖
コエンザイム	化粧品 栄養素 アミノ酸	分子生物学 補酵素 補因子
ネイル	化粧品 健康食品 質問があるページ	
イソフラボン	栄養素 健康食品 デトックス	化学関連のスタブ項目 複素環式化合物 ケトン 生体物質 健康食品
アロマテラピー	化粧品 アミノ酸 健康食品	代替医療 環境技術 香料
プエラリア	化粧品 健康食品 栄養素	マメ科 健康食品
フコイダン	健康食品 質問があるページ 栄養素	告知事項があるページ 化学関連のスタブ項目 多糖類
アスタキサンチン	化粧品 健康食品 栄養素	生体物質
カルニチン	栄養素 健康食品 アミノ酸	アミノ酸

6. 評価

ランダムに抽出した複数のクラスタに対して, 人手で分野推定精度を測り評価した. ここでは, 未知語の付与された分野が妥当であると判断した数を m,妥当でないと判断した数を n としたとき, 式(4)によって精度を算出した. その結果, 約 63%の精度で分野を推定できていることがわかった. また, クラスタによっては, 81%と高い精度を保つものもいれば, 41%という低い水準のものもあった.

$$精度 = \frac{m}{m+n} \times 100 \qquad (4)$$

表 9:未知語の分野推定結果(クラスタ No.2)

未知語	提案手法で推定された分野	wikipedia のカテゴリ
GBA	ゲームボーイ レゴ 玩具	出典を必要とする記事 2001 年のコンピュータゲーム ゲームボーイ
ウィー	Yahoo! JAPAN ゲームボーイ プレイステーション 2 用ソフト	
ローゼンメイデン	ローゼンメイデン ドラマ CD ロボット関連企業	継続中の作品 出典を必要とする記事 ローゼンメイデン ドラマ CD プレイステーション 2 用ソフト 2006 年のコンピュータゲーム
エクセル	ゲームボーイ レゴ Yahoo! JAPAN	
シャア	玩具 ドラマ CD スタジオジブリ	出典を必要とする記事 出典を必要とする記事/2008 年 9 月 宇宙世紀の人物
レゴ	レゴ ゲームボーイ 玩具	レゴ 玩具メーカー デンマークの企業 玩具 デンマークの文化 子供の遊び ロボット関連企業 エレクトロニック・アーツ
NDS	ゲームボーイ レゴ 玩具	
ジブリ	スタジオジブリ 玩具 ドラマ CD	出典を必要とする記事 東京都の企業 スタジオジブリ
ヤフオク	Yahoo! JAPAN ゲームボーイ レゴ	出典を必要とする記事 インターネットオークション Yahoo! JAPAN

7. 考察

今回の実験では，クラスタ数を 30 として実験したが，クラスタによって精度の差が激しいことから，最適なクラスタ数を設定することにより，精度の向上が望めるのでないかと考察する.

8. まとめ

本稿では，単語の出現ホストに注目し，出現ホストの類似性から，単語のクラスタリングを行い，既知の言葉を利用して，未知の言葉の分野を推定する手法の提案を行った. また，本実験下(クラスタ数:30)においては，約 63%の精度で未知語の分野推定を行えることがわかった. 今後の課題として，精度向上のため，最適なクラスタ数の設定を行う予定である.

参考文献

[1] 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," DEWS2007, C3-7 (2007.3)

[2] Chikara Hashimoto and Sadao Kurohashi. Construction of Domain Dictionary for Fundamental Vocabulary. ACL 2007 pp.137-140. 2007. 6.

[3] 黒橋禎夫，河原大輔. 日本語形態素解析システム JUMAN version 5.1 . 京都大学大学院情報学研究科, 2005.

[4] W. N. Venables, D. M. Smithand the R Development Core Team R 入門 Version 1.7.0 (2003-04-16)

[5] Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika ,Vol.53, 325-328.

[6] Hartigan, JA, Wong, MA: A k-means clustering algorithm. JR Stat. Soc. Ser. C-Appl. Stat. 28, 100-108 (1979)

[7] 下山剛司,秋岡明香,村岡洋一,時系列データを用いたタイトルタグからの新語抽出法の提案 情報処理学会第 70 回全国大会, 2008.