

Web からの効率的な行動プロセス抽出手法の検討

高橋 公海[†] 佐藤 進也[†] 松尾 真人[†]

[†] 日本電信電話(株) NTT 未来ねっと研究所 〒180-8585 東京都武蔵野市緑町 3-9-11

E-mail: [†] {t.masami, sato.shinya, matsuo.masato}@lab.ntt.co.jp

あらまし 本研究は、あるキーワードに関連する一連の行動プロセスを Web から抽出する際に、機械学習を用いて解析対象の blog 記事集合からノイズとなる記事を除くことにより、行動プロセス抽出の効率化を図るものである。また、多様なキーワードに対応するため、これまで人手で選択していたキーワードに関連する行動プロセスに出現する語を、自動で獲得する手法についても検討し、実験を通じてその有効性を検証した。抽出した行動プロセスとユーザの検索クエリ等を対応付けることにより、ユーザが今後とろうとしている行動に関する有用な情報の推薦が可能になる。

キーワード テキストマイニング, 行動プロセス, 機械学習, 知識発見, blog

1. はじめに

インターネットの普及やセンサ技術の発達により、人間の日常生活の情報を観測・蓄積することが容易になりつつある。特に Web 上には、blog や SNS, Twitter などのサービスを通じて個人の経験や行動情報が蓄積されている。これらを整理・再構成し、日常生活の行動モデルを構築することで、センサによる日常行動認識技術への応用や、ユーザが今後とろうとしている行動に関する有用な情報の推薦、行動に応じた機器の制御といった日々の生活を支援するサービスにも活用できる。

本研究では、日常生活に潜在する行動のパターンに着目し、個人が Web 上に残した情報から典型的な行動のパターンを一連のプロセスとして効率よく抽出することを目指す。本研究で対象とする行動のプロセスとは、ある状況において人々がとる一連の行動の流れを指す。例えば、「友人の結婚式に招待された」という状況であれば、「招待状に返信する→服を準備する→ご祝儀を用意する→受付をする→披露宴に参加する」といった一連の行動プロセスが想定される。他にも、「インフルエンザに罹った」という状況では「熱が出る→病院に行く→薬をもらう→家で休む」といったプロセスが考えられる。このようなプロセスを、今回は blog に記述された経験情報を解析することで抽出する。

既に、キーワードで表現された話題や対象に対してユーザがとる行動の連鎖を blog から抽出する手法を提案し、情報検索に活用しようと試みた研究^[1]は存在する。しかし、既存の手法を用いて行動連鎖抽出を行ったところ、次の 2 つの問題があることが分かった。

まず、解析対象の blog 記事に、アフィリエイト専門の blog など、個人の経験や行動が書かれていない記事が含まれると、正しい行動プロセスの頻度が下がり埋もれてしまうという問題がある。これに対し本研究では、機械学習を用いて個人の経験や行動が書かれてい

ない記事を判別し、あらかじめ除去するアプローチをとる。次に、既存研究ではキーワード毎に人手で関連語の選択を行っているが、対象とするキーワードが増えると、人手で語を選択するコストも増加し、多様なキーワードに対応することが出来ない。そこで、各キーワードに関連する語を自動で獲得する手法の検討を行った。

本稿の構成は次の通りである。2 章では関連研究、3 章では blog 記事からの行動プロセス抽出手法について説明する。4 章では実験について述べる。5 章はまとめと今後の課題である。

2. 関連研究

Web 上のデータを整理し、個人の日常生活における行動や状況を把握しようという試みの 1 つとして、経験マイニング^{[2][3][4]}が挙げられる。倉島ら^[2]は経験情報を、状況(時間・空間)、行動(動作、対象)、主観(評価、感情) から構成されると定義し、状況・行動・主観の間に存在するルールを blog から抽出する手法を提案している。例えば、「ディズニeland」という状況は「パレードを見る」という行動や、「喜び」という主観と相関が高いといった知識が得られる。経験マイニングは Web から個人の経験をマイニングし、そのパターンを抽出するという点では本研究と同じだが、行動と行動との関連やプロセスは考慮されていない。

Web から行動プロセスの抽出を試みた研究としては、旭ら^[1]の行動連鎖抽出手法が挙げられる。本研究では旭らの手法をベースに、機械学習を用いてノイズとなる記事を予め除去し、人手で選択していたキーワードの関連語を自動で獲得することにより、効率化を図る。旭らは論文中で行動連鎖がもつ性質として、プロセス性・原因結果性・時期依存性・順序可換性・並列性・繰り返し性・分岐・時間間隔という 8 つの性質を挙げている。提案されている手法と本研究の手法は、8 つ

の性質のうちプロセス性・原因結果性・時期依存性・繰り返し性・時間間隔の5つの性質を満たすことが可能であると考えられる。残りの順序可換性・並列性・分岐という3つの性質への対応については、今後の課題としたい。

3. Blog からの行動プロセス抽出手法

本章では、blog 記事からの行動プロセス抽出手法について説明する。基本的な抽出手順は次の通りである。

1. キーワードの選択
2. キーワードに関連する blog 記事の収集
3. ノイズとなる記事を機械学習で判別し除去
4. キーワードに関連する語(関連語)の集合を自動で獲得
5. blog 記事中で関連語を含む文を解析し、動作とその対象の組を得る
6. 得られた動作とその対象の組を連結し、頻出するパターンを行動プロセスとして抽出

手順2・5・6については既存研究^[1]の手法を利用し、手順3の機械学習を用いたノイズ記事除去については3.1節、手順4の関連語自動獲得については3.2節で述べる。

3.1. ノイズとなる記事の除去

既存研究^[1]の手法と同様に blog 記事を収集し、行動プロセス抽出を行ったところ、個人の経験や行動が書かれていない記事が含まれると、正しい行動プロセスの頻度が下がり埋もれてしまうことが分かった。そこで、既存研究の実験で対象としていた「インフルエンザ」「バジル」「出産」という3つのキーワードを blog 検索^[5]にクエリとして入力し、結果上位100件の記事を対象に、個人の経験や行動が書かれているか否かを人手で判定し、集計を行った。その結果、表1のような割合でノイズとなる記事が存在することが分かった。なお、記事に個人の経験情報や行動が書かれていないと判断する項目の例を表2に示す。

表1から分かるように、個人の経験や行動が書かれているか否かの割合はキーワードにより差がある。表1に挙げたキーワード以外の語についても調査を行ったが、具体的な商品名や「チャンネル」といったブランド名、「ヒアルロン酸」のような美容関係のキーワードを含む記事集合は、「結婚式」「お正月」「お歳暮」といったイベントを示す語に比べてノイズとなる記事が多い傾向があった。そのため、具体的な商品名などをキーワードとして、関連する行動プロセス抽出を行う際には、ノイズとなる記事の除去が特に重要となる。

関連研究として、川口らは blog 記事が商品のレビュー記事か否かを判別するために、機械学習を用いる手法^[6]を提案している。判別のための属性としては、特

表1：ノイズとなる記事の割合調査結果

keyword	経験・行動情報あり	経験・行動情報なし
インフルエンザ	68%	32%
バジル	28%	72%
出産	40%	60%

表2：ノイズとなる記事の例

カテゴリ	説明
商品等の紹介	アフィリエイトを目的とした記事
リンク集	リンクが列挙されている記事
引用	ニュース記事等をそのまま引用した記事
感想・批評・解説	映画・小説・テレビなどの感想や批評、社会現象・用語などの解説のみが書かれている記事
広報	個人や企業が、事業や商品等の広報活動の一環で書いている記事
メモ	行動情報は書かれておらず、単なるメモ書きと思われる記事

徴語辞書・アフィリエイトリンクの有無・品詞の割合の3つを利用している。本研究では、個人の経験や行動が書かれているか否かの判別を目的としており、川口らとはタスクが異なるため、同じ属性をそのまま利用して判別を行うことは難しい。まず、個人の経験や行動に出現する表現の特徴語辞書を作成することは容易ではない。また、個人の経験や行動が書かれている記事であっても、自動で本文中の語にアフィリエイトリンクが付いている記事も見られる。そのため、今回は記事中の語の品詞割合を属性として、ある程度判別出来ないか検討と検証を行った。

川口らはレビュー記事を対象としているため、評価表現に多い形容詞・形容動詞・動詞を手掛かりとしていたが、個人の経験や行動が書かれているか否かを判別するためにどの品詞を属性にすべきか自明ではない。そこで、属性とする品詞の検討を行うため、複数のキーワードに対して blog 記事を収集し、人手で各記事に対してラベル付けを行い、記事中の各品詞の割合を調査した。結果を図1に示す。今回、対象の blog 記事はキーワードを blog 検索^[5]にクエリとして入力した際の結果を利用し、形態素解析は Yahoo!テキスト解析 API^[7]を利用した。

図1は「結婚式」というキーワードについて個人の経験が書かれている記事50件、書かれていない記事50件を対象として調査した結果である。縦軸は割合(%), 横軸は品詞の種類を示す。図1から名詞・助詞・助動詞・動詞・記号という5つの品詞は、個人の経験や行動が書かれている記事(marriage_yes)と書かれていない記事(marriage_no)とで割合に差があることが分かる。イベントや商品名、ブランド名など他の複数の

キーワードについても調査を行ったが、ほぼ同様の割合分布となった。

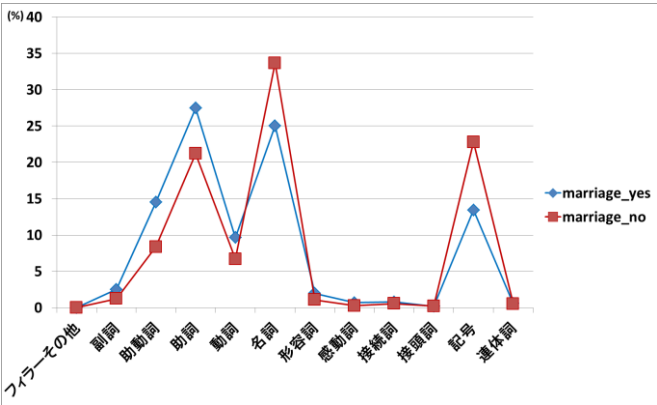


図 1：記事中の各品詞の割合(結婚式)

そこで、次の式で blog 記事に含まれる語の品詞の割合を算出し、属性として利用する。

$$P(x) = \frac{C(x)}{N} \quad (1)$$

N はキーワードに関連する全 blog 記事数、 x は名詞・助詞・助動詞・動詞・記号のいずれかである。 $C(x)$ は各 blog 記事における x の出現回数であり、 $P(x)$ は x の記事中における割合を示す。このように名詞・助詞・助動詞・動詞・記号の割合をそれぞれ算出し、この 5 つの値を属性として与え、記事の判別に用いた。

3.2. 関連語の獲得

既存研究^[1]では、blog 記事から動作とその対象の組を行動として得る際に、キーワードに関連する語を手で選択し、それらの語が含まれる文のみを解析対象としていた。例えば、「バジル栽培」というキーワードに対して「苗」や「種」といった語を選択する必要がある。しかし、対象とするキーワードが増えると人手で語を選択するコストも増加するため、多様なキーワードに対応することは困難である。

また、各キーワードに対して適切な関連語をある程度網羅的に選択することも難しい。例えば、「インフルエンザ」というキーワードについて考えた時、関連語として「熱」「病院」「薬」「タミフル」「症状」といった語が挙げられるが、これは「インフルエンザに感染した」状況における関連語である。別の状況として、「インフルエンザの予防接種」を考えた時には、「注射」「ワクチン」「予防」などの関連語の方が適切であると考えられる。このように、各キーワードには複数の状況が想定されるため、あらかじめキーワードに関連する状況を洗い出し、それぞれの状況における関連語を手で選択していくことは現実的ではない。

一般に、関連語を得る既存の手法は多数存在するが、

あるキーワードに関連する行動のプロセスに出現する語を得る方法は自明では無い。そこで本稿では、次の 2 つの手法を検討し、4 章で各手法を比較・検証した。

- 1. キーワードとよく組合せて検索される語の集合を関連語とする。
- 2. 対象とするキーワードを含む blog 記事を収集し、潜在ディリクレ配分法 (Latent Dirichlet Allocation, 以下 LDA)^[8]を用いて関連語を得る。

まず手法 1 だが、一般に、検索を行う際は検索自体が目的ではなく、検索を通じて何か別のタスクを達成することが目的となる場合が多い^[9]。そのため、何らかのタスクを達成するために検索エンジンに入力するクエリ列には、行動プロセスに出現する語が多く含まれる可能性がある。そこで、キーワードとよく組合せて検索される語の集合を関連語として利用できないかと考えた。

次に手法 2 だが、LDA^[8]は代表的なトピックモデルで、文書が潜在意味(トピック)に基づいて生成される過程を確率的に表現したモデルである。データを用いてモデルを学習することで、各文書が含むトピックや、各トピックの特徴的な単語を発見することができる。この各トピックの特徴語を、関連語として利用できないか次章で検証を行った。

4. 実験

4.1. 機械学習を用いたノイズ除去

本節では、ノイズとなる記事の除去方法として、3.1 節で述べた手法が有効であるか検証を行った。

今回、高い分類性能と汎化能力をもつ SVM を利用し、カーネル関数は線形カーネルを用いた。訓練データとして、「インフルエンザ」というキーワードに関する blog 記事を収集した。キーワードを blog 検索^[5]にクエリとして入力し、結果として返された上位 1000 件の記事から、個人の行動が書かれている記事 50 件、書かれていない記事 50 件を手でラベル付けし、実験に利用した。SVM に属性として与えたのは、記事に含まれる語の品詞のうち名詞・助詞・助動詞・動詞・記号の割合をそれぞれ算出した値である。評価は 10 分割交差検定を用いて行った。

結果を表 3 に示す。yes は個人の行動が書かれている記事、no は個人の行動が書かれていない記事である。訓練データのタグ付けにコストがかかるため、今回検証に用いたデータセットは少ないが、いずれも 8 割以上の高い精度で分類できていることが分かる。

表 3：ノイズ除去実験結果(インフルエンザ)

Class	Precision	Recall
yes	0.911	0.82
no	0.836	0.92

同様に、「結婚式」というキーワードについても個人の行動が書かれている記事 50 件, 書かれていない記事 50 件についても調査を行った。結果を表 4 に示す。「インフルエンザ」の結果ほどではないが, いずれも 7 割以上の精度で分類出来ていることが分かる。

表 4: ノイズ除去実験結果(結婚式)

Class	Precision	Recall
yes	0.77	0.94
no	0.923	0.72

今回の手法は品詞の割合のみを属性とした非常にシンプルな手法だが, なぜ 8 割前後の精度で判別出来たのか考察を行った。いわゆるスパムブログやアフィリエイト専門の blog だけでなく, 批評や感想が書かれた記事に対してもなぜ有効であったのか検証するため, 批評記事に文体が似ていると考えられる新聞記事の品詞割合の分布を調査した。結果を図 3 に示す。

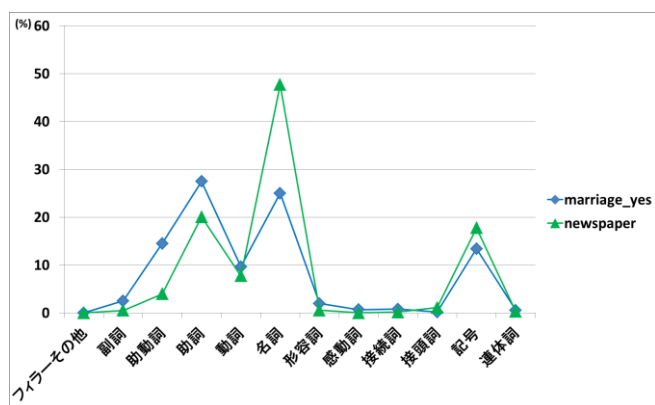


図 3: 記事中の各品詞の割合(結婚式)

図 3 は「結婚式」というキーワードについて個人の経験が書かれている blog 記事 50 件(marriage_yes), 毎日新聞 2008 年の社会面の記事(newspaper)を対象として調査した結果である。縦軸は割合(%), 横軸は品詞の種類を示す。図 3 から今回 SVM の属性として用いた名詞・助詞・助動詞・動詞・記号という 5 つの品詞の割合は, 個人の経験や行動が書かれている記事(marriage_yes)と新聞記事(newspaper)とで割合に差があることが分かる。これは, 図 1 と同様の結果である。そのため, 批評や感想が書かれた記事に対してもある程度の精度で判別を行うことが出来たと考えられる。

さらに、「インフルエンザ」というキーワードについて, 除去される行動プロセスと, 除去されずに残る行動プロセスとの比較を行った。除去される行動プロセスとしては, 「パンデミックが題材→新型インフルエンザに大騒ぎ」といった小説の概要が書かれた記事から抽出されたもの, 「鳥インフルエンザの対策→厚生省

が定める」といった政府の資料を引用した記事から抽出されたものが挙げられる。一方で, 「インフルエンザ B 型感染発症→熱がある」「インフルエンザの予防接種→注射を打つ」「熱が下がらなかった→インフルエンザにかかりました」「インフルエンザ B 型でした→熱を測ってみる→病院へ行って→治療薬の点滴」といった正しいと思われる行動プロセスは除去されずに残っていることを確認した。

4.2. 関連語の獲得

本節では, 3.2 節で述べた 2 つの手法を利用し, あるキーワードに関連する行動プロセスに出現する語を得ることが出来るか検証を行った。あらかじめ正解集合として, 確実に抽出すべき関連語を人手で 20 語程度選択した。対象のキーワードは「インフルエンザ」とし, 正解集合の中には既存の研究^[1]の実験で用いられた「熱」「咳」「病院」「治療」「感染」「予防接種」「タミフル」などの関連語も含めた。正解集合のうち, 何割の語を抽出することができたか比較した。

関連検索ワードは, Yahoo!関連検索ワード API^[10]にキーワードをクエリとして入力し, 結果として得られる上位 50 語を利用している。LDA はキーワードをクエリとして blog 検索^[5]を行い, 検索結果上位 1000 件の記事に対して適用した。記事集合から本文を抜き出し形態素解析^[7]を行い, 各記事に含まれる語から名詞と未知語のみを抽出し, bag-of-words を作成した。今回は崩壊型ギブスサンプリング^[11]を利用し, トピック数は 150 とした。得られた上位 10 件のトピックから確率の高い 5~6 個の単語を選び, 重複を除き計 50 語を利用した。正解集合のうち, 何割が 2 つの手法で得られた語(それぞれ 50 語)に含まれているか調査した結果を表 5 に示す。

表 5: 関連語抽出結果(インフルエンザ)

手法	正解率
関連検索ワード	55%
LDA	65%

正解集合は確実に抽出出来ることが望ましいが, そのような結果は得られなかった。しかし, LDA で抽出した潜在クラスには, 状況に対応した行動プロセスに出現する可能性のある語を多く含むトピックが存在していた。今回の実験で得られたトピックの例と, 各トピックの生成確率上位 10 語程度を表 6 に示す。

表 6 中の 2・3・5 という 3 つのトピックだけに着目すると, 特徴語の中に正解集合の 80%が含まれていた。トピック 2 は予防接種に行く状況, トピック 3 はインフルエンザに感染したという状況, トピック 5 は日々の予防に関連すると考えられる。関連語自動獲得のた

めには、状況に対応した関連語を多く含むトピックを選択する手法の検討が必要となるが、この点については今後の課題としたい。

表 6：抽出されるトピックの例

topic	生成確率上位の特徴語
0	自分, 人, インフルエンザ, 今回, 前, 今, 場合, 時, 仕事
1	話, 問題, 必要, 中, 日本, 意味, 可能性, 現在, 人間, 実際
2	予防接種, インフルエンザ, 注射, 小児科, 予約, 今年, 病院, 予防, 季節, 医院
3	熱, インフルエンザ, 病院, 薬, 風邪, 症状, 咳, 検査, ウイルス, 感染, 体調, 水
4	時, 今, 前, 今日, 東京, 本当, 気, 顔, 夜, 大変
5	インフルエンザ, 冬, 風邪, 手洗い, うがい, ダウン, 今年, 菌, 体調, 乾燥, 気温
...	...

5. まとめと今後の課題

本稿では、多様なキーワードに関連する一連の行動プロセスを効率的に Web から抽出するため、あらかじめ個人の行動や経験が書かれていない blog 記事を除去する手法、関連語を得る手法について検討し、検証を行った。

具体的には、品詞の割合に着目し SVM を用いてノイズとなる記事を除去する手法を検討し、実験を通じてその手法が有効であることを示した。属性として品詞の割合のみを利用しているため、キーワードに依存せず汎用的に利用できる可能性がある。今後は動詞の時制を考慮することで精度がどの程度向上するか検証を行う。また、データセットを増やし、より多くのキーワードに対しても同様の効果があるか検証を行う。

関連語の獲得については、関連検索ワードを用いる手法と LDA を用いる手法とを比較し、LDA については一定の有効性を確認した。人手での関連語選択を行う場合と比較して効率的ではあるが、状況に対応した関連語を多く含むトピックを選択する手法の検討が必要である。今後は多様なキーワードに対して精度よく関連語を抽出できるよう改良を行う。

さらに、実際に多様なキーワードに関連する行動プロセスを抽出し、既存研究と比べて抽出精度がどの程度向上したか、また、本手法で対応できるキーワードはどの程度あるかなど、評価を行う。そして、順序が可換なプロセス、分岐や並列したプロセスの抽出手法についても、検討を行う必要がある。

参 考 文 献

- [1] 旭 直人, 山本 岳洋, 中村 聡史, 田中 克己, “行動連鎖を用いた情報検索支援と Web からの行動連鎖の抽出”, DEIM2009, 2009.
- [2] 倉島健, 藤村考, 奥田英範, “大規模テキストからの経験マイニング”, DEWS2008, 2008.
- [3] 乾健太郎, 原一夫, “経験マイニング: Web テキストからの個人の経験の抽出と分類”, 言語処理学会第 14 回年次大会, 2008.
- [4] 阿部修也, 江口萌, 隅田飛鳥, 大崎梓, 乾健太郎, “みんなの経験: ブログから抽出したイベントおよびセンチメントの DB 化”, 言語処理学会第 15 回年次大会, 2009.
- [5] Yahoo! blog 検索 “<http://search.yahoo.co.jp/blog>”.
- [6] 川口 敏広, 松井 藤五郎, 大和田 勇人, “SVM と新聞記事を用いた Weblog からの意見文抽出”, 第 20 回人工知能学会全国大会, 2006 年 6 月.
- [7] Yahoo! テキスト解析 Web API: 日本語形態素解析, “<http://developer.yahoo.co.jp/webapi/jlp/ma/v1/parse.html>”.
- [8] Blei, D. M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp.993-1022, 2003.
- [9] C.Lucchese, et al. “Identifying Task-based Sessions in Search Engine Query Logs”, WSDM’11, China, 2011.
- [10] Yahoo! 関 連 検 索 ワ ー ド A P I , “<http://developer.yahoo.co.jp/webapi/search/assistsearch/v1/webunitsearch.html>”.
- [11] T. L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 101 Suppl 1:5228–5235, 2004.