1A3-5

トピックによる習慣行動抽出手法の評価
Evaluation of Habitual Behavior Extraction Methods with Topics

鈴木信雄[*1]
Nobuo SUZUKI

津田和彦[*2]
Kazuhiko TSUDA

[*1] (株)KDDI 研究所
KDDI R&D Laboratories Inc.

[*2] 筑波大学大学院
University of Tsukuba

This study has proposed the habitual behavior information extraction method from the data on Internet to build effective behavioral change support system so far. It is well known that habitual behavior improvement is important to avoid risk behaviors for a safety driving and a health improvement. It used Latent Dirichlet Allocation approach and already evaluated by using communication behaviors in Question and answering Web sites. This paper describes another evaluation by using travel behavior information. The dependency relation is often used to extract valuable information from text data. This paper also shows the comparative evaluation between our proposed method and the dependency relation method. It is realized the proposed method is more accurate than the dependency relation method according to the result of the evaluation.

1. はじめに

インターネット上の SNS や質問応答サイトなどでは，大量のテキストが日々書き込まれている．著者らは，安全運転や健康維持に効果的な行動変容を支援するシステムの構築を目的として，これらのデータから行動情報を抽出する手法を提案した[Suzuki 2013]．禁煙などの健康改善や運転中の携帯電話の利用などの危険行動の回避を対象とした行動変容に対しては，特に習慣的な行動の改善が重要であることが知られている[Kukkonen 2010]．そのため，本研究では習慣行動に注目して情報抽出を試みた．

具体的な習慣行動の抽出手法としては，トピックモデルである潜在的ディリクレ割当法 LDA(Latent Dirichlet Allocation)を用いてテキストに含まれる潜在的なトピックを推定し，そのトピックに含まれる複数の単語候補の中から相互情報量 PMI(Point-wise Mutual Information)により習慣行動に適する単語を判定している（提案手法と呼ぶ）．本稿では，提案手法のこれまでに行った質問応答サイトにおける通信に関する行動の情報（通信行動と呼ぶ）を使った評価に加えて，交通機関を使った移動に関する情報（交通行動と呼ぶ）における評価実験を行った結果を報告する．一方，テキストデータから情報を抽出する手法としては係り受け関係を利用するものが多く使われている．今回，提案手法の評価実験にて用いたものと同じデータを使い，提案手法と係り受け関係を使った手法（係り受け手法と呼ぶ）を比較評価した．比較評価の結果，提案手法の方が係り受け手法よりも高い正解率を得られることがわかった．

2. 潜在トピックと PMI を使った行動情報抽出

2.1 抽出手法

提案手法は，次のような LDA と PMI を使った習慣行動情報の抽出手法である．まず，習慣行動を歯磨きや睡眠などの生理的な習慣に限らず，高い頻度で現れる人間の行動全般のことと定義した．そこで，行動の要素として，動作，対象，周期的な頻度情報を習慣行動とした．すなわち，習慣行動 HB を式(1)の組み合わせと定義した．

$$HB = \{頻度, 動作, 対象\} \quad \cdots (1)$$

次に，トピックモデルの一つである LDA を用いて習慣行動を抽出した．トピックモデルの特徴は，一つの文書が複数のトピックの混合として表現されることであり，高い精度で文書をモデル化できることが示されている[Canini 2009]．まず，周期表現として頻繁に使われる「よく」「毎」「いつも」などをキーワードとして準備し，これらの単語を含む文をインターネットから抽出する．抽出した文に対して形態素解析を行い，頻度，動作，対象として使われやすい形容詞，動詞，名詞，副詞を bag-of-words として選択し，LDA の処理を行う．その結果，複数の単語から構成されるトピックが抽出され，それらのトピックの中から周期表現を持つトピックを抽出する．ここで，抽出された各トピックの中には周期表現の他に習慣行動を表わすと思われる動作と対象の候補となる多くの単語が含まれている．しかし，このままでは抽出されたトピック内には習慣行動以外の単語も多く含まれているため，高精度に習慣行動を抽出することはできない．そこで，トピック内の単語に対して PMI を指標として習慣行動に関連する単語を推定した．次に，習慣行動の「頻度」に対してキーワードの単語をあてはめる．「動作」に対しては，動詞-自立，名詞-サ変接続，名詞-副詞可能の各品詞にあてはまる単語について周期表現のキーワードとの間の PMI を計算し，上位 2 つの単語を「動作」を示す単語として抽出する．「対象」は，動作で選択されなかった名詞-サ変接続を含み名詞-非自立を除いた名詞に対して周期表現のキーワードとの間の PMI を計算し上位 3 つの単語を選択する．ここで，PMI は式(2)と(3)のように表され，単語間の結びつきの強さを表わすことができる指標であり，周期表現の単語に対して「動作」と「対象」を示す単語との関連性の強さを示している．

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \quad \cdots (2)$$

$$p(x) = \frac{f(x)}{N},\ p(y) = \frac{f(y)}{N},\ p(x,y) = \frac{f(x,y)}{N} \quad \cdots (3)$$

連絡先：鈴木信雄，(株)KDDI 研究所，nu-suzuki@kddilabs.jp

2.2 交通行動における評価

先の研究において，通信事業者の質問応答サイトのテキストデータ 8,953 文を使い，通信行動における本手法の評価を行った．その結果，20 トピックが抽出され，その中で 18 トピックが正しく習慣行動を抽出したと判断でき，正解率は 90%となった．

今回は，通信行動以外に交通行動についても評価を実施した．渋滞緩和や地球温暖化への対策としてマイカーではなく公共交通機関への利用を促進するような行動変容が求められており，行動変容支援の対象となる．交通に関する質問応答の中には，人の移動に関する習慣行動が多く含まれている．交通に関する質問応答のテキストデータから習慣行動を抽出し，その情報を行動変容に活用できると考えらえる．今回，交通に関する質問応答サイトから 6,627 文の発言を収集し，その中から通信行動と同じ周期表現キーワードを含む文を抽出した．これによって抽出された文は 521 文となり，これらに対して形態素解析を行った．次に，形容詞，動詞，名詞，副詞の各品詞の単語を抽出し発言毎に LDA ツールである LDA-C を使ってトピックを抽出した[Blei 2003]．結果として 50 個のトピックを得た．得られたトピックの例を表 1 に示す．これらのトピックから周期表現のキーワードを持つものを選択することで 16 個のトピックが得られた．

表 1 質問応答サイトのトピック例（交通行動）

トピック	単語(品詞)
Topic 000	いつも(副詞-一般),あり(動詞-自立),バス(名詞-一般),い(動詞-非自立),し(動詞-自立),円(名詞-接尾-助数詞),料金(名詞-一般),日(名詞-接尾-助数詞),利用(名詞-サ変接続),長原(名詞-固有名詞-人名-姓)
Topic 026	よく(副詞-一般),の(名詞-非自立-一般),バス(名詞-一般), アドベンチャーワールド(名詞-一般),とれ(動詞-自立),日(名詞-接尾-助数詞),目(名詞-接尾-一般),し(動詞-自立),時間(名詞-副詞可能),朝(名詞-副詞可能)

これらの各トピックにおいて PMI による候補選択を行い，習慣行動に適した単語を抽出した．結果の一部を表 2 に示す．

表 2 トピックから選択した習慣行動の例（交通行動）

トピック	頻度	動作	対象	習慣行動の解釈
Topic 000	いつも	あり し	バス 利用	いつもバスを利用する．
Topic 026	よく	し 時間	バス アドベンチャーワールド 朝	よく，朝の時間にバスでアドベンチャーワールドへ行く．

つづいて，全てのトピックについて提案手法により求めた単語が習慣行動を表現しているかどうかを手動で確認した．その結果，16 トピック中 12 トピックが正しく習慣行動を抽出したと判断でき，正解率は 75%となった．

この評価実験では，4 つのトピックにおいて抽出された単語から習慣行動を得ることができなかった．それらの例を表 3 に示す．不正解のトピックについては，全てが「動作」や「対象」として正しい単語を抽出できていなかった．姓や「名義」などの交通とは関連のないと思われる単語も出現しており，データを調査したところ，交通とは関連の無い話題が含まれていた．日本の姓の大半は地名と同じ単語を使用していることがトピックにこれらの単語が含まれている要因と考えられる．例えば，「小淵沢」という単語は地名でもあり姓でもある．そのため，「名義」という単語との関連では姓が適用されると考えられるが，移動の視点からは地名であることが期待される．これらを文脈から地名なのか姓なのかを判別する仕組みを取り入れることが必要である．これらについては，データの前処理にて，単語頻度情報などにより交通との関連性が低い発言を削除するような対策が考えられる．また，移動においては時間と場所の情報が重要である．対象の単語で時間や場所に該当する単語が含まれている場合には、それらを優先的に選択する処理も必要である．

表 3 不正解の例（交通行動）

トピック	頻度	動作	対象	不正解の原因
Topic 011	よく	し	京都 青島 北海道	動作の抽出が不正解．トピック内の他の単語にも正解なし．
Topic 035	よく	割引 行く	(なし)	対象の抽出が不正解．トピック内の他の単語にも正解なし．
Topic 049	よく	あり し	小淵沢 名義	対象の抽出が不正解．移動と「名義」は関連していない．

3. 係り受け関係を用いた習慣行動情報抽出

3.1 抽出手法

テキストデータから特定の情報を抽出する手法の一つに係り受け関係を用いたものがある．伊藤らは，ブログを対象に行動と興味の時系列推移を抽出しており[Itoh 2011]，遠藤らは，感情表現の抽出を行っている[Endo 2006]．さらに，池田らは，係り受け関係を用いて有害情報の検出も提案している[Ikeda 2010]．このように情報抽出には有効な手法であるので，係り受け関係を用いた習慣行動情報抽出を実現し，提案手法との比較評価を行った．

今回実現した係り受け関係を用いた抽出手法を図 1 に示す．まず，テキストデータの中から 2 項にて示したものと同じ周期表現を含む文を選択する．この周期表現のキーワードを「頻度」とする．次に構文解析処理を行い，各文節の係り受け関係を求める．求めた係り受け関係から，周期表現が係る先の文節（係り先文節と呼ぶ）中に現れる動詞を「動作」とする．周期表現が最初に出現する文節と係り先文節の間で，同じく係り先文節にかかる文節中の名詞を習慣行動の「対象」とする．ここで，動作や対象に該当する文節が複数ある場合には，最初の文節を抽出対象とする．

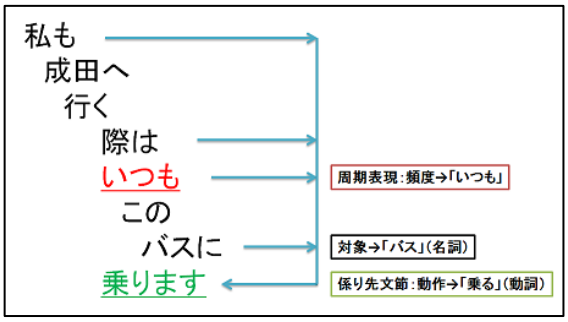

図 1 係り受け関係を使った習慣行動の抽出

3.2 評価

係り受け手法の評価では，提案手法の評価の時と同じデータを使い，通信行動に関連するデータと交通行動に関するデータの 2 種類について評価を行った．係り受け関係の解析ツールとしては CaboCha を利用した[Kudo 2002].

まず，通信事業者の質問応答サイトのテキストデータ 8,953 文を使い，通信行動における係り受け手法の評価を行った．その結果，153 個の習慣行動が抽出され，その中で 63 個が正しく抽出したと判断でき，正解率は 41.2%となった．表 4 に正解の例を表 5 に不正解の例を示す．

表 4 係り受け関係を用いた正解例(通信行動)

頻度	動作	対象	習慣行動の解釈
よく	かける	相手	よく相手にかける
いつも	使う	定額	いつも定額を使う
よく	間違える	BCC	よく BCC を間違える

表 5 係り受け関係を用いた不正解例(通信行動)

頻度	動作	対象	不正解の原因
よく	する	画素	動作の情報が不十分
毎	する	様々	動作の情報が不十分
いつも	助かる	*	対象が不正

次に，提案手法の評価時と同じく交通に関する質問応答サイトから 6,627 文を使い，通信行動と同じ周期表現キーワードを含む 521 文を抽出した．交通行動における係り受け手法の評価を行った結果，86 個の習慣行動が抽出された．その中で 29 個が正しく抽出したと判断でき，正解率は 33.7%となった，表 6 に正解の例を表 7 に不正解の例を示す．

表 6 係り受け関係を用いた正解例(交通行動)

頻度	動作	対象	習慣行動の解釈
いつも	乗る	バス	いつもバスに乗る
よく	止める	近隣	よく近隣に止める
よく	調べる	電車	よく電車を調べる

表 7 係り受け関係を用いた不正解例(交通行動)

頻度	動作	対象	不正解の原因
いつも	する	百	対象の抽出が不正
よく	わかる	方	対象の抽出が不正
よく	いく	かた	対象の抽出が不正

不正解の原因としては「動作」と「対象」において意味がとおらない単語を抽出してしまっており，単語辞書や固有名詞抽出の併用などが必要となっている．一方で，係り受け手法と提案手法との間で正解率を比較すると，通信行動と交通行動の両方において提案手法の方が抽出精度は向上していることがわかる．このことより，係り受け手法に比べ，提案手法の方が有効な手法であることがわかる．ただし，両方の手法に共通した課題として，情報内容の充実が不足していることがある．例えば，「いつもバスに乗る」という習慣行動が両手法にて抽出されているが，いつ，どこで，誰がという情報が不足している．さらに，行動変容につなげるためは行動の条件も必要である．これらについては，習慣行動の定義を拡張し，フレームによる情報抽出を使うことなどにより対応することが考えられる．

4. おわりに

本稿では，すでに提案している潜在トピックモデルと相互情報量を用いた習慣行動の抽出手法に関する評価について報告した．これまでは通信行動のみの評価であったが，交通行動に関するデータも加えた評価を行い，高い正解率を確認した．さらに，テキストデータからの情報抽出に多く使われている係り受け関係を用いた手法との比較評価を行った．その結果，通信行動と交通行動の両方において，提案手法の方が高い正解率が得られることがわかった．

今後は，抽出情報内容の充実を図ると共に，健康改善行動へも領域を広げ，本手法による習慣行動情報の収集を進める予定である．

参考文献

[Blei 2003] David M. Blei, Andrew. Y. Ng and Michael I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp.993-1022, 2003.

[Canini 2009] Kevin R. Canini, Lei Shi and Thomas L. Griths: Online Inference of Topics with Latent Dirichlet Allocation, Proceeding of the 12th International Conference on Artificial Intelligence and Statistics, 2009.

[Endo 2006] 遠藤, 齊藤, 山本: 係り受け関係を利用した感情生起表現の抽出, 言語処理学会第 12 回年次大会, pp.947-950, 2006

[Ikeda 2010] 池田, 柳原, 松本, 滝嶋: 係り受け関係に基づく違法・有害情報の高精度検出方式の提案, DEIM Forum 2010, C9-5, 2010

[Itoh 2011] 伊藤, 吉永, 豊田, 喜連川: ブログユーザの行動・興味に関する時系列推移 3 次元可視化システム, DEIM Forum 2011, E7-5, 2011

[Kudo 2002] Taku Kudo and Yuji Matsumoto: Japanese Dependency Analysis using Cascaded Chunking, CONLL 2002

[Kukkonen 2010] Harri Oinas Kukkonen: Behavior Change Support Systems: The Next Frontier for Web Science, Proceedings of the Second International Web Science Conference (WebSci10), 2010.

[Suzuki 2013] Nobuo Suzuki and Kazuhiko Tsuda: An Effective Method for Habitual Behavior Extraction from the Internet, Procedia Computer Science, Vol.22, pp.599-605, 2013