1F3-2

行動変容のための習慣行動情報の抽出

An Extracting Method of Habitual Behavior Information for Behavioral Modification

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

Many researches are being conducted about the analysis of human behavior using sensor devices in the real world or various information all over the Internet. The objective mainly is to improve social behaviors and habits such as the prohibition of smoking and prohibition of the use mobile phones while driving. These unhealthy social behaviors and habits tend to cause health problems and antisocial behaviors. Behavioral modification understands the habitual behavior is one of the most important behaviors to solve these issues. This paper proposes a new technique to extract habitual behaviors for discovering objectives of behavioral modification. Specifically, Latent Dirichlet Allocation or LDA is used for getting appropriate topics of frequent behavior literary expressions and Point-wise Mutual Information or PMI is applied to select suitable words for habitual behaviors. The experiment of this technique by using text data from question answering Web sites of telecommunication industry area was evaluated and showed good performance results.

1. はじめに

近年，センサやインターネット上の大量のデータを使った人間の行動把握に関する研究が多く行われている．また，禁煙や運転中の携帯電話の利用などに代表されるような，利用者の行動を変化させて，より良い社会を目指そうとする行動変容の研究も継続して行われている[Bianchi 2005]．著者らは，コンピュータによる行動変容手法の構築を目的として，インターネット上のテキスト情報から行動変容に関する知識を抽出する手法を検討している．行動変容の目的には様々なものが検討されているが，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など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，本研究ではインターネット上のテキスト情報から習慣的な行動を抽出する手法を提案する．具体的には，高精度のテキストマイニング手法として最近注目されている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) によってテキスト情報の潜在的なトピックを推定し，トピックに含まれる複数の単語の候補の中から相互情報量 PMI (Point-wise Mutual Information) を使って習慣行動に適すると考えられる単語を抽出する．提案手法に対して，通信事業者による質問回答サイトのテキストデータを使った実験を行い，本手法の評価を実施したところ，良好な結果を得ることができた．

2. 行動情報の抽出技術

インターネット上のテキスト情報から人間の行動情報を抽出する技術には多くの研究事例がある．例えば，ブログ記事から行動情報を抽出する代表的な技術に，乾らと倉島らの「経験マイニング」がある．乾らは，人手で収集した辞書を使ってテキスト情報から出来事と行為を同定し，評価・感情・時間・態度を合わせることで経験情報として抽出している[Inui 2008]．倉島らは，行動情報を動作と対象とを合わせた情報として定義し，動詞「する」で終わる文を対象として辞書による照合を使って抽出している[Kurashima 2008]．また，高橋らは，行動を item と action に細分化しトピックモデルと特定の事象のキーワードとの共起頻度を使って行動情報を抽出しており，さらに，Twitter のデータを使って行動の時間的な前後関係も付与している[Takahashi 2012]．これらの研究は，人手による辞書をベースとした手法であるために作業コストが高く，著者らが対象としている習慣行動についても考慮されていない．次に，ブログや Twitter のテキスト情報を使って習慣行動を解析したものとしては田中らの研究がある[Tanaka 2013]．この研究では，記事の投稿履歴から時間的な行動の発生確率を算出することで習慣行動のモデル化を行っている．行動の情報それ自体の抽出には行動に関する用語の辞書をあらかじめ用意することで照合している．また，投稿時刻と投稿数から tfidf を使って時間毎の行動スコアを算出し，このスコアから，睡眠中・出勤中・勤務中・食事中・帰宅中・その他というあらかじめ定義された 6 つの習慣行動を推定している．これに対して，著者らはテキスト情報からあらかじめ定義されていない習慣行動そのものを抽出することを目的としている．

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

3. LDA と PMI を用いた習慣行動情報の抽出

3.1 習慣行動

本研究で扱う習慣行動とは，行動変容を行なう対象の候補抽出が目的であるため，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を指す．そのため，これまでの関連研究にて提案されている行動の要素である動作と対象に対して周期的な頻度情報を加えたものを習慣行動と定義する．すなわち，習慣行動 HB とは式(1)の組み合わせと定義する．

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

3.2 LDA による習慣行動の候補抽出

近年，文書や各種履歴などの離散データを解析する手法として bag-of-words 表現された文書の生成過程を確率的にモデル化したトピックモデルが注目されている．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]. 本研究では, このような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することを試みる. 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して LDA の入力情報とする. 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. 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単語が含まれている.

3.3 PMI による候補選定

LDA により抽出されたトピック中の複数の単語は, 必ずしも習慣行動を表わす単語のみが含まれているわけではない. そのため, これらの候補単語から頻度, 動作, 対象の各属性を示す単語を抽出する必要がある. まず, 「頻度」についてはキーワードの単語を抽出してあてはめる. 「動作」については, 候補単語中の動詞-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. 上位 2 つの単語を使うのは, 例えば, 「メールする」のような場合に 1 つだけでは「メール(名詞-サ変接続)」しか抽出されず「する(動詞-自立)」という具体的な動作まで網羅できないためである. 「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. この結果, 表 1 のようなスロットの条件を持つ情報を得ることができる. ここで上位 3 つの単語を使うのは, 多くの文章の事例から 3 つの単語で「対象」を表わすことが可能と判断したためである.

表 1 習慣行動情報の取得内容

	頻度	動作	対象
品詞/キーワードのスロット	「よく」 「毎」 「いつも」	動詞-自立 名詞-サ変接続 名詞-副詞可能	名詞-サ変接続と名詞-非自立を除いた名詞
選択	常に選択	上位 2 つの単語	上位 3 つの単語

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各動作と対象を示す候補単語との関連性の強さを示している. x は周期表現のキーワード, y は動作および対象の単語を示す. また, $f(x)$は周期表現を持つ文の中での単語の頻度, $f(x,y)$は周期表現のキーワードと単語の共起頻度, N は周期表現を持つ文の中の全単語数である.

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

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

これまで述べた習慣行動を抽出する手順を図 1 に示す.

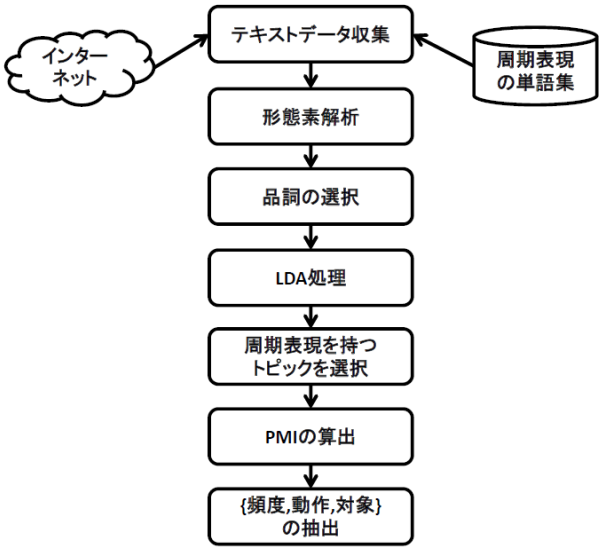

図 1 習慣行動の抽出手順

4. 評価

これまで述べた手法を使って習慣行動情報を抽出する実験を行った. 実験の対象としては, 質問応答サイトのテキストデータを使った. 特に, スマートフォンなどを使った通信における行動変容への適用を想定するために, 通信事業者の質問応答サイトを利用した. まず, 質問応答サイトから 8,953 文の発言データを収集し, その中から先に示した周期表現キーワードである単語を含む文を抽出した. これによって抽出された文は 226 文となり, これらに対して形態素解析を行った. ここで, 形態素解析には茶筌を使用した. 次に, 形容詞, 動詞, 名詞, 副詞の各品詞の単語を抽出して発言毎に LDA の処理を施した. LDA のツールとしては Blei らの LDA-C を用いた[Blei 2003]. LDA の処理の結果として得られたトピックは 49 個であった. ここで得られたトピックの例を表 2 に示す. 得られた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 20 個のトピックが得られた.

表 2 質問応答サイトのトピックの例

トピック	単語(品詞)
Topic 006	いつも(副詞-一般), し(動詞-自立), 電話(名詞-サ変接続), 変更(名詞-サ変接続), 変換(名詞-サ変接続), 請求(名詞-サ変接続), 料金(名詞-一般), の(名詞-非自立), 円(名詞-接尾)
Topic 007	毎月(名詞-副詞可能), し(動詞-自立), 思い(動詞-自立), 安心(名詞-サ変接続), 携帯(名詞-サ変接続), メール(名詞-サ変接続), 無料(名詞-一般), 料金(名詞-一般), て(動詞-非自立), れ(動詞-接尾)

これらの各トピックにおいて 3.3 節で示した PMI による候補選択の方法を使って習慣行動に適した単語を選択した. 結果の例を表 3 に示す.

表 3 トピックから選択した習慣行動の例

トピック	頻度	動作	対象	習慣行動の解釈
Topic 006	いつも	し 電話	変更 請求 翌月	いつも翌月に請求の変更を電話する.
Topic 007	毎月	安心 メール	携帯 無料 料金	料金が無料なので, 毎月, 安心して携帯でメールしている.

つづいて, 全てのトピックについて本方式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, 20 トピック中 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.

この評価実験では, 2 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. それらの結果を表 4 に示す. ここで, Topic021 については, 形態素解析の段階でサービス名である「Win」が「W」と「in」に分割されてしまい意味の無い単語となっている. これは, 形態素解析ツールの辞書のメンテナンスを継続的に行うことで解決可能と考える. Topic023 については, 動作として抽出された動詞の「やり」と「あり」だけでは習慣行動の動作を表現することができていない. しかし, PMI 値の 3 位の動詞である「し(する)」を使えば, 「よく電話する」のような行動表現が可能であるので, 特定の動詞が選択された場合には次候補を選択するような処理が必要と考える.

表 4 不正解の例

トピック	頻度	動作	対象	不正解の原因
Topic 021	よく	あり 対応	携帯 in ドコモ	サービス名のWinの一部である「in」のみが単語として抽出された.
Topic 023	よく	やり あり	電話 回答 人	動作として抽出された単語が習慣行動に不適格.

5. おわりに

本稿では, インターネット上のテキスト情報から行動変容の対象となる習慣行動を抽出するための手法を提案した. まず, 習慣行動を頻度・動作・対象の組であると定義した. 次に, LDA を使って周期表現を含むトピックを求め, さらに求められたトピック中の複数の候補単語に対して PMI の上位単語を選ぶことで習慣行動に適した単語を抽出する. 通信事業者の質問応答サイトのテキストデータを使った評価実験の結果, 90%の高い正解率を得ることができた. しかし, 評価に利用したテキストデータの数量が少ないことや, 対象とする情報のテーマが通信に限られていることから行動変容の対象とする習慣行動が広範囲に収集できているとは言えない. 今後は, より多くの量のテキストデータを収集し, 通信以外のテーマも評価の対象とすることで, さらなる課題の抽出と解決策の検討を行う. また, 今回提案の方式では, インターネットから最初に収集したテキストデータを LDA の入力とする際に多くの種類の品詞を使用している. しかし, その後の処理フェーズでは主に動詞と名詞のみを判断基準に用いている. そのため, LDA へ入力するデータの選定時に抽出する品詞の範囲を狭めることが可能であると考えられるので, これについても評価を実施していく. さらに, 本方式の有効性を比較評価するために, 係り受け解析などによる習慣行動の抽出手法についても検討し比較実験を行う予定である.

参考文献

[Bianchi 2005] Adriana Bianchi and James G. Phillips: Psychological Predictors of Problem Mobile Phone Use, Cyberpsychology & Behavior, Vol.8, No.1, pp.39-51, 2005.

[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.

[Inui 2008] 乾, 原: 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 回年次大会論文集, pp.1077-1080, 2008.

[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.

[Kurashima 2009] 倉島, 藤村, 奥田: 大規模テキストからの経験マイニング, 電子情報通信学会論文誌, Vol.J92-D, No.3, pp.301-310, 2009.

[Takahashi 2012] 高橋, 佐藤, 松尾: Web からの行動プロセス抽出手法の提案, 電子情報通信学会 人工知能と知識処理研究会, Vol.112, No.319, AI2012-20, pp.31-35, 2012.

[Tanaka 2013] 田中, 中村, 寺口, 中本, 加藤: マイクロブログを対象としたユーザの習慣的な行動の解析に関する研究, 情報処理学会第 75 回全国大会 5N-4, 2013.