人物の属性に着目した時期・時間帯に関連のある事象の獲得

山元 航平　　嶋田 和孝
九州工業大学大学院 情報工学府
{k_yamamoto, shimada}@pluto.ai.kyutech.ac.jp

1 はじめに

昨今，知識獲得についての研究が盛んに行われている．獲得の対象となっている知識は，文脈上の数値に対する大小感覚知識 [1] や，一般的に物理的に近い場所にあるオブジェクトペアの知識 [2] など，多くの種類が存在する．大規模な知識ベースも複数作成されており [3, 4]，そのような知識が質問応答 [5] や非タスク指向型対話 [6]，また複数の言語処理ベンチマークタスク [7] において有効であることが報告されている．

我々は以前の研究 [8] において特定の時期や時間帯に関連がある事象（人間の行動，自然現象など）についての知識の収集を行った．この知識は { 事象，時期・時間帯 } のように，事象と時期・時間帯の二つ組として表記される．知識の具体例としては { 雪が降る，冬 }，{ 湯船に浸かる，夜 } などが挙げられる．時期と時間帯は粒度の異なる 5 種類を定義し，対話システムなどでの使用を想定して，それぞれの時期・時間帯に対応する事象の知識を獲得した．

以前の研究 [8] の問題点の 1 つとして，獲得した知識が世間一般的なものに汎化されており，人間の行動における人物の属性が考慮されていないという点がある．社会人や学生などの社会的な人物の属性は，その人物の生活傾向などに大きな影響を与えるであろう．この問題点の具体例を 1 に示す．例えば，{ 出勤する，7 時 } や { バイトに行く，16 時 } という知識は一般的にはもっともらしい知識であると思われる．しかし { 出勤する，7 時 } は大学生に，{ バイトに行く，16 時 } は社会人に当てはまるとは考えにくい．このように，社会の人々を区別なく平均化した知識は特定の人々にとっては正しくない場合がある．人物の属性を踏まえた知識獲得を行えば，より厳密に実社会を反映した知識が得られ，様々な言語処理タスクに寄与することが期待できる．

本研究では様々な人物属性の中でも特に人物の生活傾向に影響を与えていると予想できる社会的な人物属性（社会人，学生，専業主婦など）に着目し，時期・時間帯に関連のある事象の獲得を試みる．具体的には，幅広く利用されている SNS である Twitter[1] のデータからの知識の獲得を行う．

[1] https://twitter.com

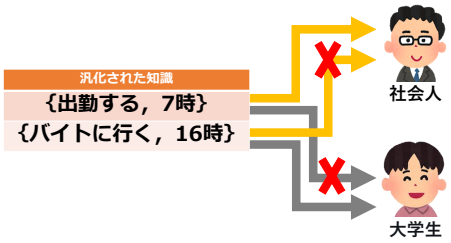

図 1: 汎化された知識の問題点．

2 関連研究

本研究では人間の行動を始めとする様々な事象の中でも，特定の時期や時間帯に関連を持つ事象について，人物の属性ごとに獲得することを目的としている．Ge ら [9] は Wikipedia データを利用して事象の知識の収集を行った．しかし，ここで Ge ら [9] が獲得の対象としている事象は，事件，事故，災害などの突発的な出来事やオリンピックなどの大規模な行事などであった．社会的な影響の大きな事象についての知識は重要なものであると思われるが，特定の時期や時間帯に関連するものではないことが多いであろう．さらに，社会的影響の大きな事象と同様に人々の身の回りの事象知識にも価値があると考え，本研究では人間の行動に関する事象を中心に獲得を行う．人間の行動と時間情報に着目した知識獲得を行っている論文としては，Tandon ら [10] の研究と Yao ら [11] の研究が挙げられる．Tandon ら [10] はドラマ，映画，小説などの物語作品から人物の行動知識の獲得を行い，その際，行動時の時間などの情報も抽出をした．しかし，抽出の対象とした時間情報は昼，夜などで種類が少なく，それぞれの定義も曖昧であった．本研究では，ひと月から 1 時間までの幅広い粒度の時間区分を設定する．Yao ら [11] は Web テキストから動詞およびその動作主を抽出することで，近しい行動のペアなど，行動同士の関係知識の獲得を行った．時間の情報も抽出しているが行動間の相対的な時間情報であり，事象どうしの関係知識に近いものであるといえる．さらに，Tandon ら [10] の研究と Yao ら [11] の研究のいずれにおいても行動を行った人物の属性については考慮されていない．本研究では，人物の属性の沿った知識の獲得を目指す．

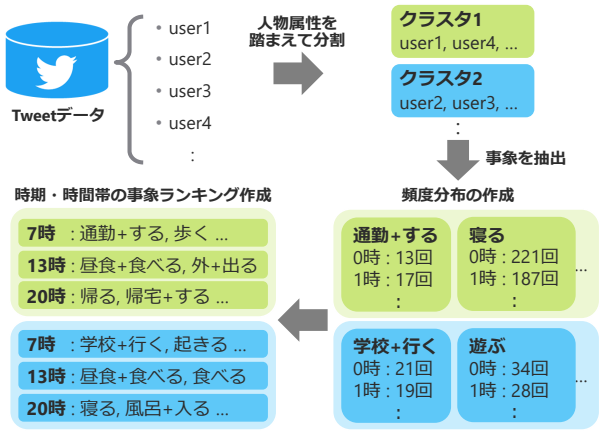

図 2: 提案手法の概略図.

3 提案手法

本研究では，人物の属性に着目した時期・時間帯に関連のある事象の獲得を目指し，Tweet データからの知識の獲得を行う．ある事象について言及する Tweet が，特定の人物属性のユーザらによって特定の時期・時間帯において集中的に投稿されているとする．加えて他の人物属性のユーザの投稿に同様の傾向が見られないのであれば，言及されている事象はその時期・時間帯に関連を持ち，その人物属性に特徴的な事象であると考えられる．これが提案手法の基本的なアイデアである．提案手法の概要を図 2 に示す．まず，人物属性を考慮したユーザの分割を行う (3.1 節)．次に, それぞれの属性のユーザの Tweet から事象の抽出と，事象の頻度分布の作成をする (3.2 節)．最後に，時期・時間帯ごとにランキングを作成することで，特定のクラスタにおいてある時期・時間帯に集中して発生する事象を獲得する (3.3 節)．以下，各手順について説明を行う．

3.1 人物属性を考慮したユーザの分割

1 節でも述べたとおり，社会的属性の同じ人物は似たような生活パターンを有すると予想できる．例えば，社会人と学生の生活パターンよりも，社会人同士の生活パターンの方がより似ているであろう．また，生活パターンは Tweet の投稿時間の傾向にも影響を与えると考えられる．このことから，同じ属性を持つ人物の Tweet の投稿時間パターンには同様の傾向があるのではないかと予想できる．各ユーザの Tweet の内容もそのユーザの人物属性を反映しているとは考えられるものの，本研究では各ユーザの Tweet の投稿時間を用いて，人物属性を考慮したユーザの分割を行う．具体的には 24 時間の各時間における Tweet 投稿の割合を値に持つ 24 次元のベクトルを作成し，それを各ユーザのベクトルとしてクラスタリングを行う．クラスタリング手法には k-means を使用する．

3.2 事象の抽出と頻度分布の作成

Tweet からの事象の抽出及び事象の頻度分布の作成は，我々の以前の研究 [8] における手法をクラスタごとに適用して行う．以下が手法の詳細である．Tweet 中には，投稿したユーザ自身の行動をはじめとして，自然現象や社会の出来事などの様々な事象が含まれている．このような事象は動詞や，名詞と動詞のペアで表現されることが多い．そこで，まず Tweet 中から動詞を抽出する．さらに，動詞の直前に名詞が存在する場合には，動詞単体とは別に名詞+動詞のペアを抽出する．本研究で抽出する動詞および名詞と動詞のペアを「事象語」と呼ぶ．事象語の抽出の後，抽出元の Tweet の投稿時間を基に各事象語の頻度分布を作成する．頻度分布の作成においては月単位，曜日単位，平日週末，朝昼晩，1 時間単位の 5 つの時間区分を設定する．

3.3 各時期・時間帯に集中して発生する事象の獲得

特定の時期・時間帯に強く関連する事象は，その時期・時間帯に投稿された Tweet において言及されることが多いと考えられる．そこで，Tweet 中で各時期・時間帯に集中して使用される事象語を抽出する．具体的には，その事象語の総出現数に対するその時期・時間帯での出現数を，各事象語の各時期・時間帯におけるスコアとし，時期・時間帯ごとに関連の強い事象語のランキングを作成する．ランキングの作成はそれぞれのクラスタにおいて，個別に行う．しかし，こうして作成したランキングの上位に存在する事象語であっても，以下のいずれかの条件を満たす事象語は今回の目的である人物属性を考慮した事象の獲得と合致しない．

1. 長期間の頻度分布での出現数のばらつきが大きい．
2. いずれの属性のクラスタにおいても同様に，ランキング上位に存在する．

条件 1 を満たす語は，災害や大規模な社会的イベントなどの影響による急激な Tweet 投稿の増加などを指す，バーストと呼ばれる現象に関連した事象語である．バースト中の事象は特定の時期や時間帯に紐づくものではないが，短時間に多くの Tweet が投稿されることで，バーストの発生した時期・時刻のランキングのノイズとなる傾向がある．本研究では，年間の頻度分布での頻度の分散を利用してバースト中の事象の除去を行う．また，条件 2 を満たす事象語は特定の時期・時間帯に関連のある事象語であるとは言えるものの，特定の人物属性において特徴的な事象語であるとは考えづらい．本研究では属性を考慮した知識の獲得を行うため，条件 2 を満たす事象語は今回獲得を行う対象となる事象語ではない．

以上の理由により，条件 1 と条件 2 のいずれかを満た

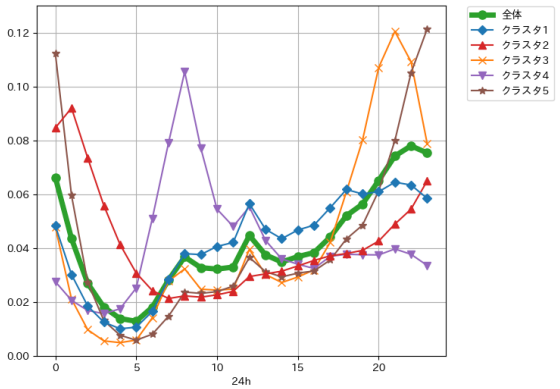

図 3: 全体およびクラスタごとの投稿頻度分布.

す事象語はランキング中から除外する．このようにして作成したランキング上位の事象語を，その属性クラスタにおける時期・時間帯に関連のある事象であるとする．

4 実験・考察

本節では実験設定と実験結果，考察について述べる．

4.1 実験設定

本研究では，Twitter 社の提供する TwitterAPI を使用して収集した Tweet データを使用した．収集はランダムなアカウントのタイムライン（投稿した Tweet の履歴）を取得するという方法で行った．プログラムによる複数の定型文の自動投稿を行っているユーザや，スパム的な投稿を行ってユーザなどの Tweet はノイズになると考え，フィルタリングを行っている．この手法で 2 年間以上のタイムラインを取得できたユーザの，任意の 1 年間の Tweet を解析対象とした．収集の結果，12,000 ユーザの約 270 万 Tweet が解析対象となった．なお，3 年以上の期間のタイムラインはごく少数しか取得できなかった．月のランキングにおいてバースト除去を行うためには，複数年期間での頻度の分散を求める必要があるため，3.3 節のバースト除去手法は月のランキングには適応することができなかった．また，ユーザの分割の際のクラスタリングにおけるクラスタ数 k の値は経験的に 5 に設定した．

4.2 実験結果

本節ではクラスタリングの結果とランキングの作成結果について述べる．

4.2.1 クラスタリング

クラスタリングによって 5 つのクラスタを作成した．各クラスタに属するユーザの Tweet 投稿の分布が大きく異なっていれば，意図したとおりの適切なクラスタリングが行えているといえる．データセット全体と作成したクラスタそれぞれの 24 時間での Tweet 投稿時間分布を図 3 に示す．緑の太線がデータセット全体，その他の色がそれぞれのクラスタの分布である．図 3 より，クラスタ 1 は 12 時と 21 時に投稿時間の山を持ち，早朝にかけて減少する分布となっており，比較的規則正しい生活パターンを送っていると予想できるユーザの集合となっている．また，クラスタ 2 は 1 時に大きなピークを持つ分布となっている．日中に山を持たないことから，決まった生活パターンを持たないユーザの集合であると予想できる．このように，クラスタリングによって Twitter の投稿傾向の異なるユーザのクラスタを獲得した．

表 1: クラスタ 1　時間ごとのランキング　一部抜粋

時間	事象語
7 時	朝+なる, 会社+行く, 弁当+作る
8 時	遅延+する, 昨日+寝る, 乗り遅れる
9 時	遅延+する, 2 度+寝る, 昨日+寝る
:	:
12 時	完売+する, 名乗る, 診る
13 時	お昼+食べる, もの+買う, 弄る
14 時	休憩+する, 事+できる, 働ける
15 時	昼寝+する, 引っかかる, いえる
16 時	一人+いる, 昼寝+する, お知らせ+する
17 時	ご飯+作る, 仕事+おわる, 残業+する
:	:
21 時	輝く, 整う, 寝かしつける
22 時	寝かしつける, 宣言+する, 浮上+する

4.2.2 ランキング作成

作成したクラスタごとに，月，平日休日，曜日，朝昼晩，時間に強く関連する事象語のランキングを提案手法により作成した．作成したランキングのうち，図 3 で示したクラスタ 1 のランキングの一部を表 1，表 2，表 3 に，クラスタ 2 のランキングの一部を，表 4，表 5 に示す．クラスタ 1 のランキング（表 1，表 2，表 3）には,「会社＋行く」（7 時，平日）,「仕事＋終わる」（17 時）,「残業＋する」（17 時）などの勤務に関する事象語や,「弁当＋作る」（7 時）,「ご飯＋作る」（17 時）,「寝かしつける」（21 時，22 時）などの家事や育児に関連する事象語，また「昼寝＋する」（15 時，16 時，昼）などの事象語が含まれている．これらの事象語から，子育てをしながら働いている人や専業主婦などのユーザを含むクラスタであると予想できる．またクラスタ 2 のランキング（表 4，表 5）には，深夜の「ラーメン＋食べる」（1 時），社会人や小中学生，高校生では考えられない時間の「寝坊＋する」（12 時，13 時）などの事象語や,「バイト＋行く」（16 時，昼）などの比較的若い人々に関係するであろう事象語が含まれている．このことから，生活パターンの乱れやすい大学生などのユーザを含むクラスタであると予想できる．今回スペースの都合で結果を載せていないクラスタ 3，4，5 にも，クラスタ 1，2 と同様に特徴的な事象語がそれぞれのランキング中に存在する．しかし，クラスタに含まれるユーザの属性の推測までは難しい結果であった．

表 2: クラスタ 1　平日週末のランキング

時間	事象語
平日	会社+行く, 日+限る, ブロック+する
週末	完売+する, 無事+終わる, すける

表 3: クラスタ 1　朝昼晩のランキング

時間	事象語
朝	2 度+寝る, 度+寝る, 昨日+寝る
昼	昼寝+する, 休憩+する, 販売+する
晩	明日+する, 酔う, 最高+すぎる

4.3　考察

クラスタ 1 やクラスタ 2 の結果から，人々の属性に沿った事象語の獲得には一定の成功を収めていると考えられる．一方で前述のように評価の難しい結果となったクラスタ（3，4，5）も複数存在する．原因の 1 つとして，属性でのユーザ分割が適切なものでないことが考えられる．本研究では，各ユーザの 24 時間での Tweet 投稿分布を使ってクラスタリングを行うことで，属性でのユーザの分割を試みた．同じ属性の人物は Tweet 投稿時間にも似た傾向があるのではないかという予想に基づくものであった．クラスタ 1，2 のように分割できた属性も存在するため，人物の属性の Tweet の投稿傾向への影響は存在するといえるものの，個人的な生活習慣や Twitter の使用傾向の違いの Tweet 投稿傾向への影響も同様に存在すると考えられる．人物属性によるより明確な分割の実現には，投稿内容などの投稿時間分布以外の要素を考慮する必要があると考えられる．クラスタリング手法の改良とともに，ユーザの人物属性での分割を，事前に設定したいくつかの属性へのユーザの分類問題として解くことも視野に入れ，より適切な手法の検討に取り組む必要がある．

また，本研究の実験結果への評価はあくまで定性的なものにとどまっており，定量評価はおこなっていない．定量的な評価手法の検討も必要であると考える．

5　おわりに

本研究ではユーザの人物属性に着目した時期・時間帯に関連のある事象の Tweet データからの獲得を試みた．Twitter ユーザを属性で分割するために各ユーザの投稿の分布を用いてクラスタリングを行い，特定の時期や時間帯に集中して発生する事象の抽出を各クラスタに対して行った．結果として複数の知識の獲得に成功しが，問題も複数存在する．ユーザの属性でのより良い分割手法の検討及び獲得した事象の評価手法の検討が今後の課題である．

表 4: クラスタ 2　時間ごとのランキング　一部抜粋

時間	事象語
1 時	ラーメン+食べる, 何+できる, キス+する
:	:
12 時	寝坊+する, 嘘+つく, 受かる
13 時	寝坊+する, 回収+する, 傷つける
14 時	休み+する, コラボ+する, 焼く
15 時	おなか+すく, 曇る, 遭遇+する
16 時	成る, バイト+行く, 曲がる
17 時	フォロー+する, 募る, お待ち+する

表 5: クラスタ 2　朝昼晩のランキング

時間	事象語
朝	今+寝る, 座れる, 目覚める
昼	今日+帰る, バイト+行く, 入院+する
晩	ギター+弾く, 歌える, キス+する

参考文献

[1] Katsuma Narisawa, Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. Is a 204 cm man tall or small? acquisition of numerical common sense from the web. In *Proceedings of ACL*, pp. 382–391, 2013.

[2] Frank F Xu, Bill Yuchen Lin, and Kenny Zhu. Automatic extraction of commonsense locatednear knowledge. In *Proceedings of ACL*, pp. 96–101, 2018.

[3] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, pp. 4444–4451, 2017.

[4] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. In *Semantic Web Journal*, pp. 167–195, 2015.

[5] Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Improving question answering by commonsense-based pre-training. In *Proceedings of NLPCC*, pp. 16–28. Springer, 2019.

[6] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of ICLR*, 2019.

[7] Zhang Zhengyan, Han Xu, Liu Zhiyuan, Jiang Xin, Sun Maosong, and Liu Qun. Ernie: Enhanced language representation with informative entities. In *Proceedings of ACL*, pp. 1441–1451, 2019.

[8] Kohei Yamamoto and Kazutaka Shimada. Acquisition of knowledge with time information from twitter. In *Proceedings of IALP*, 2019.

[9] Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, Furu Wei, and Ming Zhou. Eventwiki: a knowledge base of major events. In *Proceedings of LREC*, pp. 499–503.

[10] Niket Tandon, Gerard De Melo, Abir De, and Gerhard Weikum. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of CIKM*, pp. 223–232, 2015.

[11] Wenlin Yao and Ruihong Huang. Temporal event knowledge acquisition via identifying narratives. In *Proceedings of ACL*, pp. 537–547, 2018.