同行者および自宅からの距離を考慮したトピックモデル

深澤佑介　太田順

東京大学　人工物工学研究センター

yusuke.fukazawa@gmail.com

概要　本稿では、同行者および自宅からの距離を考慮したトピックモデルの提案と評価を行う。結果として、自宅から近い距離の場合、親側の視点で子供と楽しむようなトピックが増え、自宅からの距離が中距離～遠距離になると、子供視点で父親や母親と一緒に行動するときのトピック（ショッピング、週末の遠出など）が増えることが分かった。

キーワード　同行者、コンテキストアウェア、位置情報、トピックモデル

1 はじめに

モバイルユーザの屋外における行動支援の一環として、ユーザが誰（with whom）と何をし（why）にどこに行くか(where)を予測することは非常に重要である。Tweet 上には、これらの三つの情報を同時に含む Tweet が多くなされている。例えば以下のとおりである。

Tweet	同行者	場所	トピック
Off to Miami with my dad	父親	マイアミ	旅行
Fried cooking with mommy. Haha	母親	自宅	料理
It's such a joy to run around and play soccer in the rain with my son	息子	サッカー場	スポーツ
Had a great day ice skating at @intuBraehead with my daughter abd her 5 friends	娘	スケート場	遊び
Loved sitting down relaxing with kids watching Man of Steel. Henry Cavill is super hot ;)	子供	家	映画

そこで、本稿では、上記の 3 つの情報を含む Tweet を多数集め、同行者、場所、トピック間の相関関係を計算する方法を提案する。相関関係が分かれば、同行者、場所、トピックのうち、2 つが判明すれば、3 つ目の情報を予測可能となる。具体的には、同行者、場所、トピックの生成に関する因果関係について仮説を設け、その仮説に基づきトピックモデルを提案する。3 つの情報を含むデータを収集し、トピックモデルのパラメータを学習することで、同行者、場所、トピック間の相関関係を算出する。

なお、場所については、緯度経度をそのまま扱う方法、意味的な場所として扱う方法がある。意味的な場所については、同行者、場所、トピックを同時に一つの文章に含む必要があるが、Tweet の短い文章で 3 つの情報を同時に含む可能性は非常に低い。このことから、今回は場所を緯度経度情報として扱う。ただし、緯度経度情報をそのまま利用した場合、ユーザの生活圏に依存してしまい、他のユーザに適用できなくなるため、相対的な位置情報に変換する。ここでは、同行者を家族関係に限定することで、家族が一緒に住む自宅からの距離に着目する。

2 関連研究

場所や時間などのコンテキスト情報とトピック間の関係を考慮したトピックモデルとして「時間」[1][2]「場所」[3][4]に応じたトピックモデルが数多く提案されている。また、深澤らは、「同行者」をトピックを変化させるパラメータとして考慮したトピックモデルを提案している[5]。しかしながら、「同行者」「自宅からの距離」「トピック」の三者を同時に考慮したトピックモデルは提案されていない。

3 提案手法

3.1 データ収集

以下の流れでデータ生成を行った。

1. 同行者情報および位置情報を含む Tweet を抽出する。Tweet 集合 1 とする。
2. Tweet 集合 1 の各ユーザの全 Tweet から自宅（緯度経度）を推定する。
3. Tweet 集合 1 の各 Tweet に付与された位置情報と自宅との距離を算出し、同行者、自宅からの距離、Tweet の三つ組みを生成する。

まず、1 番目の Tweet 抽出について述べる。同行者情報を含む投稿文を評価用データの対象とするため、語彙統語パターン「with 同行者」と一致する文字列を含む投稿を抽出した。同行者として利用した名詞については下表に記載する。次に 2 番目の自宅推定について述べる、まず、Tweet 集合 1 で抽出した Tweet のユーザリストを作り、当該ユーザの

全 Tweet（緯度経度情報つき）を抽出する。その中で、自宅にいると思われる時間帯（午後 10 時～午前 6 時）の緯度経度情報を抽出する。その中で 100m 以内誤差をもつ Tweet の中でもっとも多いものを自宅として推定する。最後に 3 番目は、ヒュベニ (Hubeny) の公式を用いた。2014 年 4 月の Tweet を対象にデータ収集を行った。結果として、以下の表に示す文書を収集した。

同行者	語彙統語パターン（with+同行者）で使用した名詞	抽出された文書数
父親	father, dad, papa, pappy, my father, my dad, my papa, my pappy,	12536
母親	mother, mom, mama, my mother, my mom, my mama, my mother, my mom	15079
息子	son, junior, my son, my junior	3688
娘	daughter, my daughter	1257
子供	kid, children, my kid, my children	6335

3.2 グラフィカルモデル

同行者、場所、トピック間の因果関係について、いずれかの 2 つが既知で、そこから残り 1 つを決めるとすると、以下の 3 つの仮説が考えられる。

仮説 1）同行者、場所⇒トピック（誰とどこに行くかは決まっていて、そこから何をするかを決める）

仮説 2）同行者、トピック⇒場所（誰と何をするかは決まっていて、そこからどこに行くかを決める）

仮説 3）場所、トピック⇒同行者（どこで何をするかは決まっていて、そこから誰と行くかを決める）

ここで、Tweet では、同行者、トピックについては、ユーザが明示的に記載しているものの、緯度経度情報はユーザが明示的に入力したものではない。そのため、今回は、因果関係 2 に基づくトピックモデルを提案する。図 1 に提案するトピックモデルを記載する。提案モデルの推論は、LDA の推論で利用される Collapsed Gibbs Sampling（CGS）を利用する。

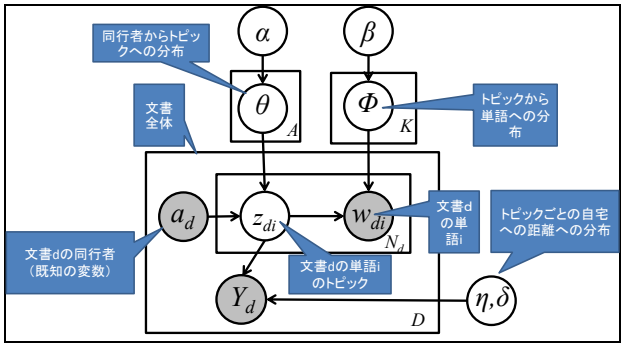

図 1　提案モデル

4 質的評価

図 2 にモデルの学習結果を示す。学習結果から、自宅から近い距離の場合、親側の視点で子供（kid,son,daughter）と楽しむようなトピック（class15,17）が多いことが分かる。一方、中距離～遠距離になると、子供が自立し、子供視点での父親や母親と一緒に行動するときのトピックが増えてくる。例えば、Class12、Class11 はショッピングに関するトピック、Class6 は旅行先での就寝に関するトピック、Class5,7 は週末の遠出に関するトピックである。このことから、ユーザから見て子供といるときと、親といるときで自宅からの距離およびトピックには違いが生じることが分かった。

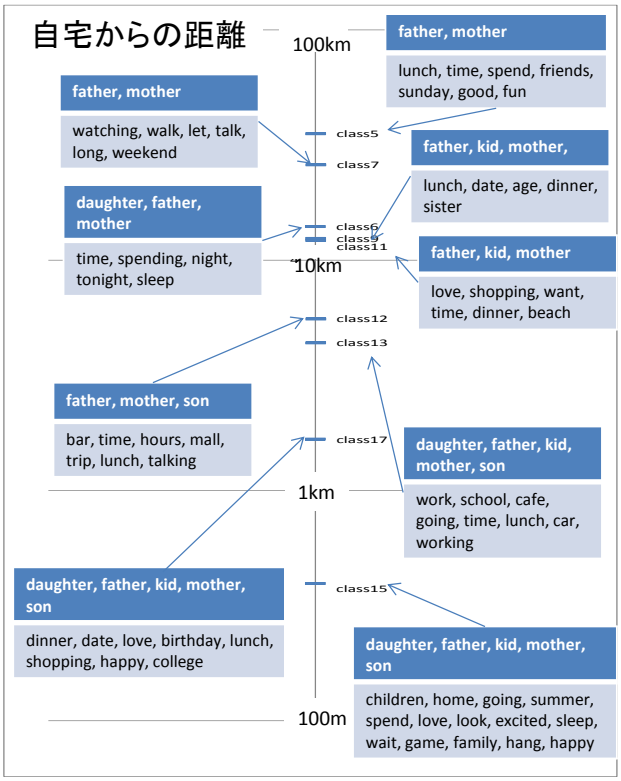

図 2　学習結果（縦軸は自宅からの距離を表す。各トピックはトピックごとの自宅への距離への分布 η に従った距離に配置している。トピックごとの表で表上では同行者からトピックへの分布 θ から導出した同行者を、表下はトピックから単語の分布 Φ から導出した単語を表示している。）

5 結論

本稿では同行者と自宅からの距離を考慮したトピックモデルを提案および評価した。

参考文献

[1] X. Wang and A. McCallum, “Topics over time: a non-markov continuous-time model of topical trends,” *Proc. of KDD*, pages 424-433, 2006.

[2] N. Kawamae, “Trend analysis model: trend consists of temporal words, topics, and timestamps,” *Proc. of WSDM*, 317-326, 2011.

[3] J. Eisenstein, B. O’Connor, N. A. Smith, and E. P. Xing, “A latent variable model for geographic lexical variation,” *Proc. of EMNLP*, pp. 1277–1287, 2010.

[4] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, K. Tsioutsiouliklis, “Discovering geographical topics in the twitter stream,” *Proc. of WWW*, 769-778, 2012.

[5] 深澤 佑介, 太田 順, 同行者に応じたトピックモデル, 情報処理学会論文誌, Vol.55, No. 1, pp.413-424, 2014.