時間特徴ベクトルの生成手法と観光地推薦システムへの応用

房　冠深†　　亀井　清華‡　　藤田　聡‡

†‡広島大学工学研究科 〒739-8527 東広島市鏡山１丁目４－１

E-mail:　†bou@se.hiroshima-u.ac.jp,　‡{s-kamei, fujita}@ se.hiroshima-u.ac.jp

あらまし　現在、ユーザの日常のニーズを満たし、アイテムを推薦するために、情報推薦システムが広く使われている。本文は情報推薦システムにおけるアイテムの時間的な特徴の抽出または抽象に注目し、自動的な時間特徴ベクトルの生成手法を提案する。主なアイディアは：1)　Wikipedia を用いてアイテムに関連するコーパスを定義する;　2) Twitter を用いてアイテムに関連するトレンドを定義する;3)　アイテムのトレンドに含まれた時間の特徴をハイライトする。提案手法の有効性を検証するため、旅行推薦システムを構築した。実験の結果は以下の結論に導いた：1)　収集した 6057 個の観光地の中、提案手法は約 9%に対して優れた機能を果した;2)　この 9%の観光地の時間特徴ベクトルはベクトル空間の中の分布は互いの類似度に関連している;　3)　時間特徴ベクトルのベクトル空間での変化は季節的な観光地推薦に用いられる。

キーワード　観光推薦システム，季節特徴ベクトル，Wikipedia，Twitter

1. 研究背景

現在、推薦システムは各分野で広く使用されている[1][2]。本文は推薦システムにおけるアイテムが有する時間とともに変化する時間特徴に注目する。その特徴は常に季節、天気などのコンテクスト要素と関連している。例えば季節または期間限定のメニューを提供するレストラン、天気に敏感する高速道路、季節のアトラクションを備えている観光地の特徴が分析の対象となる。複数の分析の方法が既存研究に提案されたが、人工的にコーパスの用意、ドメインに制限されることや様々の欠点が存在する。

本文は時間の特徴に対し、通用性の高い自動的な時間特徴ベクトルの生成手法を提案する。主要な考えは以下の三点にある:　1)　Wikipedia を用いてアイテムに関連するコーパスを定義する;　2) Twitter を用いてアイテムに関連するトレンドを定義する;3)　アイテムのトレンドに含まれた時間の特徴をハイライトする。Wikipedia は 480000 ページ以上の規模のソーシャルメディアとし、ステップ 1 に相応しいデータソースだと考える。また推薦システムの研究で広く使用されているデータソースとした Twitter は近年、モバイルコミュニケーションと共に迅速に発展した[3]。

提案手法を評価するため、本文では提案手法に基づいた観光地推薦システムも提案し、構築された。提案システムではユーザのプロファイルに加え、予定の旅行時期も考慮して推薦結果を決定する。収集された 6057 個の観光地情報は全日本に覆い、Wikipedia において詳しい紹介文が設けられている。観光地に関連するトレンド情報は Twitter Streaming API を用いて抽出された。

2. 提案手法

2.1. 時間特徴ベクトルの生成

先ず、時間軸を分割し、時間の特徴が変化する時間スライスを定義する。与えられたスライスにおいてアイテムの特徴は不変と仮定する。目標はスライス毎に、時間特徴ベクトル(TFV)を生成する。TFV は与えられたスライスの特徴を表し、基本特徴ベクトル(BFV)から拡張してくる。BFV はベクトルとして TF-IDF で計算され、時間の要素に関わらない視点でアイテムの特徴を表現する:

$$\vec{v}_{\mathrm{i}}^{\mathrm{b}} = \left\{ \left(w_j, TF_{i,w_j} \right) \times IDF_{w_j} \right) \middle| w_j \in W_i \} \qquad (1)$$

その中、与えられたアイテムはo_iとし、W_iはアイテムの紹介文d_iに含まれた単語の集合とする。全アイテムの紹介文の集合をD =∪ d_i と定義する。

TFV を生成するために、BFV の式(1)の TF 部分を拡張する。与えられたスライスs_kに発表された tweet の集合をt_kとし、W_iに含まれた単語w_jの Twitter における TF は以下のように:

$$TF'_{k,w_j} = \frac{n'_{k,w_j}}{\sum_{w \in W} n'_{k,w}} \qquad (2)$$

中に、n'_{k,w_j}は単語w_jの頻度とし、Wはすべてのアイテムに関連する単語の集合とする。即ち、tweet の中のアイテムの紹介文Dに含まれていない単語を除外する。変数αでTF_{i,w_j}とTF'_{k,w_j}を調整することで、BFV の$\vec{v}_{\mathrm{i}}^{\mathrm{b}}$を TFV に拡張する:

$$\vec{v}_{\mathrm{i,k}}^{\mathrm{t}} = \left\{ \left(w_j, \left((1-\alpha) TF_{i,w_j} + \alpha TF'_{k,w_j} \right) \times IDF_{w_j} \right) \middle| w_j \in W_i \right\} \quad (3)$$

2.2. 旅行推薦システム

この節は提案手法の有効性を証明するために、提案された季節のアトラクションに対応できる観光地推薦システムを説明する。提案システムは観光地の季節的な特徴(e.g.,花見、紅葉またはイベント)を考慮し、全日本の観光地を推薦対象とする。

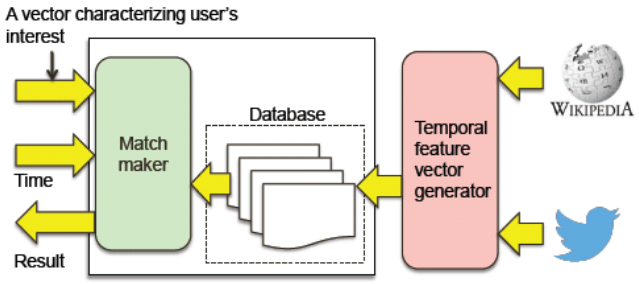


Figure 1:提案されたシステムの構造

Figure 1:提案されたシステムの構造は提案された推薦システムを構造を表す。システムは予め Wikipedia のカテゴリ:"日本の観光地"に属する観光地の紹介文を収集し、Mecab で形態素解析を行う。一方、収集された観光地の名称を用いて Twitter Streaming API で 2013.9 から 2014.3 まで約 50 万の関連 tweet を抽出した。実装の段階で、時間スライスを一ヶ月間と定義した。i.e. 一年間を季節に 12 個分割した。式(3)のTF_{i,w_j}とTF'_{k,w_j}を 1 対 1 の比率にするため、αの値を 0.995 に修正した。

3. 実験と評価

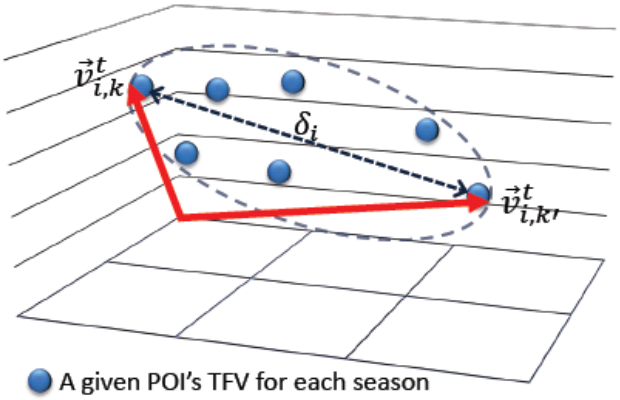
提案手法の有効性を提案された観光地推薦システムで評価する。このセクションでは先ず、各季節の TFV の多様性について評価する。次に、提案システムに対して、TFV の変化の視点からシステムの性能を測る。

3.1. TFV の多様性

全観光地の TFV が属したベクトル空間をΩと定義し、Ωの次元ごとに単語と関連する。故に観光地の季節s_kの TFV である$\vec{v}^{t}_{i,k}$を空間Ωでの点とマッピングする。従って与えられた観光地o_iの直径を定義する：

$$\delta_i = max_{k \neq k'}\{\|\vec{v}^{t}_{i,k} - \vec{v}^{t}_{i,k'}\|\}$$

$\|\cdot\|$はL^2norm とする。以下の図のように表す：



TFV の多様性の評価するために、収集された観光地

直径の大きさ	観光地の数
[0, 0.005)	4531
[0.005, 0.01)	892
[0.01, 0.015)	424
[0.015, 0.02)	86
[0.02, 0.025)	41
[0.025, 0.03)	18
[0.03, ∞)	1

Table 1 各々の観光地の直径

各々の直径を計算する。テーブル Table 1 各々の観光地の直径は観光地の直径を表す。その中、75%の観光地は小さい直径を得ており、9%の直径は 0.01 より大きい。即ち、提案手法は特にこの 9%の観光地に有効であり、生成された TFV は十分季節の特徴を表現している。故に 0.01 を閾値として収集された観光地を active と inactive に分ける。

3.2. TFV の時間的な変化

この節では提案手法がベースとして提案システムでの性能を評価する。即ち、与えられたユーザのプロファイルと旅行の予定時期に対し、ベクトル空間Ωで TFV が最もユーザプロファイルに類似度の高い観光地を観察する。現在ユーザのプロファイル構築は未実装なので、仮に或る観光地の TFV をプロファイルとしてシミュレートする。

実験の結果を表示するため、Active 観光地から仁和寺を選び、ユーザのプロファイルとし、他の観光地の TFV とのコサイン類似度を計算する。類似度が高い上位三個をテーブル Table 2 仁和寺の類似度が高い観光地に表している。

11 月の TFV	2 月の TFV
下呂温泉合掌村	姫路城
大覚寺	八戸公園
再び公園	高遠城

Table 2 仁和寺の類似度が高い観光地

4. まとめ

本文は通用性が高い自動的な時間特徴ベクトルの生成手法を提案した。その上、提案手法を評価するため、季節のアトラクションに対応する観光地推薦システムを提案した。実験の結果は、提案手法は正しく観光地の時間的な特徴を抽象したことを証明した。

文　献

[1] J. Hong, W.-S. Hwang, J.-H. Kim, and S.-W. Kim,, "Context-aware music recommendation in mobile smart devices," Proceedings of the 29th Annual ACM Symposium on Applied Computing, 2014, pp. 1463–1468..

[2] X. W. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li, "We know what you want to buy: A demographic-based system for product recommendation on microblogs,"Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014,pp.1935-1944.

[3] K. Oku, K. Ueno, and F. Hattori, "Mapping geotagged tweets to tourist spots for recommender systems," Proceedings of the IIAI 3rd International Conference on Advanced Applied Informatics, 2014, pp. 178-794. .