5K-04

A Method of Temporal Feature Characterization with Application to Tourism Recommender Systems

Fang Guanshen† Kamei Sayaka‡ Fujita Satoshi‡
Graduate School of Engineering, Hiroshima University†‡

1. Introduction

Recommender systems are widely used in our daily life. In this paper, we are particularly interested in the recommendation of objects to have *temporal features* which dynamically vary according to the change of the context, such as season, weather, reputation and the type of customers. Although there are many existing recommender systems which could reflect such temporal features to the outcome of recommendation [A. Karatzoglou, 2010] [O. Khalid, 2013], most of them force the service provider to manually input temporal features, or are designed for a restricted domain, e.g., restaurant, traffic, and museum.

The main contribution of the current paper is the proposal of a general framework which automatically generates *temporal feature vectors* of such objects. The basic idea of the proposed method is outlined as follows: 1) At first, it identifies the vocabulary concerned with each object through online documents; 2) it then identifies the trend (i.e., temporal variance of features for each season) over all objects through SNS; and finally, 3) it highlights the weight of words contained in each identified trend to obtain temporal feature vectors for each object. A typical data source for the first step is Wikipedia, and a data source for the second step is Twitter which has become extremely popular according to the rapid development of mobile communication [K. Oku, 2014].

The effectiveness of the proposed framework is evaluated by building a tourism recommender system to recommend appropriate POIs to the users for a designated season.

2. Proposed Method

2.1 Framework to Generate Temporal Feature Vectors

In this paper, we assume that the time axis is separated into several ranges so that the feature of objects is regarded to be invariant. Each range is called a *season*. The goal of the proposed framework is to generate a temporal feature vector (TFV, for short) of each object for each season. TFV is calculated by extending the basic feature vector (BFV, for short), in such a way that it reflects the trend of words in each season. More concretely, BFV is a vector of TF-IDF weights and TFV is its extension. Let O be the set of objects and d_i be a document concerned with object $o_i \in O$. In general, d_i is the union of statements on object o_i relevant to different seasons. In other words, in order to generate TFV for each season, we need to distinguish word sets relevant to each season in document d_i. Let W_i be the set of words contained in document d_i and $\mathrm{W} = \cup_i W_i$. Then, BFV $\vec{v}_{\mathrm{i}}^{\mathrm{b}}$ of object o_i is defined as

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

where TF_{i,w_j} and IDF_{w_j} are the TF (term frequency) and IDF (inverse document frequency) weights of word w_j in document d_i, respectively.

The key idea of the proposed framework is to extend the definition of the TF weight by considering the trend of words. Let t_k be a collection of tweets issued in season s_k. By considering t_k as a single document, we can define the TF weight of word w_j in season s_k as follows:

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

where n'_{k,w_j} is the number of occurrences of word w_j in t_k and W is the set of words contained in documents concerned with O; namely *we omit words in tweets which do not appear in any document*. With the above notion, TFV $\vec{v}_{\mathrm{i}}^{\mathrm{t}}$ of object o_i for season s_k is defined as

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

where $0 \le \alpha \le 1$ is an appropriate parameter.

2.2 Tourism Recommender System

To demonstrate the effectiveness of the proposed framework, we build a tourism recommender system in which objects correspond to POI (point-of-interest) and their temporal features. Such a seasonal recommendation of POIs is desired by many travelers visiting Japan since the appearance of the scenery completely different for different seasons.

Figure 1 illustrates an overview of the tourism recommender system with a flow of generating TFVs. We focus on 6057 POIs given in the category of "sightseeing spots in Japan" in Wikipedia. The set of words (nouns) W_i in each document d_i is obtained by conducting the morphological analysis using MeCab. A set of tweets relevant to tourism is acquired from Twitter using Twitter Streaming API. More concretely, we acquired 50 million Japanese tweets issued from evaluation given in the next section, we assume that one year is divided into 12 disjoint seasons. More precisely, we derive seven documents from T which

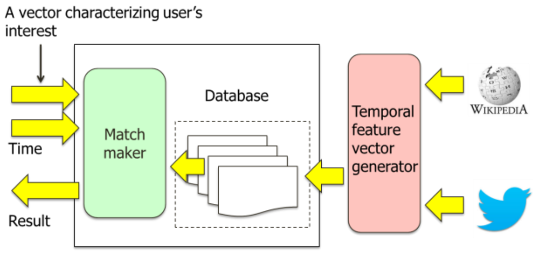

Figure 1 Tourism recommender system based on the proposed framework.

Table 1 Details of four typical clusters.

	relevant features	season	# of POIs
C_r	red leaves, waterfall	Nov.	109
C_i	illumination,cafe	Jan.	36
C_s	snow,event	Jan.	35
C_c	cherry-blossom,park	Mar.	61

represent the trend of each season, as we collected tweets for seven months. The parameter α is fixed to 0.995.

3 Evaluation

3.1 Variance of TFVs

With a normalization of TFVs by the length in the L^2 -norm, we evaluate the time transition of the similarity of POIs which is obtained by applying the K-means method [Elkan, 2003] to TFVs of all POIs, with $k = 70$ for each season. Table 1 summarizes observed four typical clusters, namely C_r, C_i, C_s and C_c. Each of them is only observed and defined for a specific season. It indicated that seasonal features are extracted in the relevant season.

3.2 Impact to Recommendation

To evaluate the performance of the proposed method, we regard the mean of TFVs contained in each aforementioned cluster as the preference of user. Since a vector space is spanned by all TFVs, in which such means can be mapped as points. Thus we have four points in the vector space characterized by red leaves, illumination, snow and cherry-blossom. The performance as a tourism recommender system is evaluated by analyzing the Top- $\boldsymbol{t}$ POIs according to the cosign similarity of the corresponding TFVs to the designated points for each season k. Such a subset of POIs is denoted as Q_t^k hereafter. For comparison, we also calculate the Top- $\boldsymbol{t}$ POIs according to the cosign similarity of the corresponding BFVs to the designated points, which is denoted as P_t hereafter.

Assume that p is the mean of cluster C. We measure the goodness of subset X concerned with designated point p by $|C \cap X|$. Thus the advantage of using TFV instead of BFV can be measured by calculating

$$\xi(t,k) \stackrel{\text{def}}{=} \left|Q_t^k \cap C\right| - |P_t \cap C|$$

Table 2 The value of $\boldsymbol{\xi(30,k)}$ with $\boldsymbol{k} \in \{\boldsymbol{Sep., Oct., Nov., Dec., Jan., Feb., Mar.}\}$

	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
C_r	5	5	**5**	5	5	4	4
C_i	-8	4	4	4	**4**	4	4
C_s	6	4	4	5	**8**	3	3
C_c	11	10	11	11	12	13	**13**

Table 3 The value of $\boldsymbol{\xi(t,k)}$ where the value of $\boldsymbol{t}$ is fixed to be equal to the corresponding cluster size.

	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
C_r	6	8	**7**	8	9	-4	2
C_i	-15	-5	0	0	**0**	-3	0
C_s	7	6	4	5	**9**	4	4
C_c	8	8	11	10	8	17	**17**

which depends on the value of parameter t and the selection of season k.

Table 2 summarizes the results for t = 30, where emphasized numbers in the table designates the season which defines the corresponding cluster. The result implies that by using TFVs instead of BFVs, we can recommend more POIs contained in the given cluster, and such an effect is maximized when the designated season is coincide with the season defining the cluster.

Recall that the value of $\xi(t,k)$ depends on the value of parameter t. Table V summarizes the results where t is fixed to the cluster size, e.g., we let $t = 109$ for cluster C_r. Comparing with Table IV, the larger gap of $\xi(t,k)$ in each row implies that there are various Q_t^k for a given cluster. In other words, if the designated season is not relevant with given C, less POIs that contained in C will be recommended.

4 Concluding Remarks

In this paper, we proposed a general framework which automatically generates TFV to characterize objects for recommender systems. The result of evaluations shows that the TFVs indicate the temporally variable features of the POIs. As a future work, we will improve the SNS trends extraction.

References

A. Karatzoglou, X. A. (2010). Multiverse recommendation: N-dimensional tensor factoriza-tion for context-aware collaborative filtering. *Proceedingsof the 4th ACM Conference on Recommender Systems*, (pp. 79–86).

Elkan, C. (2003). Using the triangle inequality to accelerate k-means. *Twentieth International Conference on Machine Learning (ICML-2003)*, (pp. 147–153).

K. Oku, K. U. (2014). Mapping geotagged tweets to tourist spots for recommender systems. *the IIAI 3rd International Conference on Advanced Applied Informatics*, (pp. 789 – 794).

O. Khalid, M. U. (2013). Om-nisuggest: A ubiquitous cloud based context aware recommen-dation system for mobile social networks. *IEEE Transactionson Services Computing*, (pp. 401–414).