博士論文（要約）

Activity-aware Topic Models to Find User Preferences of Activities from Twitter Posts

(Twitter からのユーザ行動傾向を推定するための行動理解型トピックモデル)

by

朱丹丹

A dissertation submitted to the Department of Precision Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Supervisor
Professor Jun Ota

The University of Tokyo, Japan
August 2014

In this thesis, we focus on users' daily life activities, and propose the method to find users' preferences of activities from twitter posts.

Finding users' preferences of activities is crucial for recommender systems (RSs) to deliver appropriate information to different people. Generally, it may be infeasible to directly obtain user preferences before delivering recommendations. User-generated content (UGC) such as tweets, forum posts and blogs provides an indirect but practical way to access to user preferences. However, generally, it is very difficult to manually handle UGC due to the huge data and irregular format, and thus, automatic methods by mathematical models are needed for explore the hidden knowledge of UGC. We focus on the human activities, because human is keeping conducting certain activities either actively or passively, and most of these activities are driven by their concerns. By the assist of the knowledge of users' preferences of activities, recommender system can improve their information delivery to better meet the needs of customers.

To achieve the purpose, we choose the LDA model as the basic model for extension to develop new topic models. LDA model well simulates the assumed psychological process of writing tweets, and in addition, it is widely used in the field of clustering for the features of feasibility and easy extensibility.

Before starting to search the solution, two problems need to be clarified: one is the appropriate expression for activities, and the other is the accurate way to identify activities. For the first problem, we propose the verb-nonverb pair as the activity collocation which needs to be extracted from tweets; for the second problem, the information about the general topic of activities is supposed to play a complementary role in the delivery of activity information.

Therefore, the solution is the tri-layer cluster which contains a topic layer, an activity layer and a word layer. By applying this kind of cluster, each word in a given tweet can be analyzed to figure out the according activity and topic, and thus, the most mentioned activity and topic can be find out as the preference of this user.

The proposed system to generate the expected tri-layer cluster is composed of two main functional modules, that is, the data mining part and the clustering part. The part of data mining is very important in the selection of reasonable activity collocations and the filtering out noise words. In the clustering part, two newly-developed topic models are proposed: the wpLDA model is for generating the topic layer and the activity layer for the clusters; the TLCG model uses the output of wpLDA as an external input to generate the expected tri-layer clusters. The experimental results indicate that the proposed system works well in the clustering and the estimation of user preferences of activities.