論文の内容の要旨

Thesis Summary

論文題目　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）

氏　　名　朱　丹丹

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

Finding user 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 contents (UGCs) such as tweets, forum posts and blogs provide an indirect but practical way to access to user preferences, because they are created by user themselves and much affection has been poured into them. However, generally, it is very difficult to manually handle UGCs due to the huge data and irregular format, and thus, automatic methods by mathematical models are needed for exploring certain hidden knowledge of UGCs. We focus on the general concept of activity and study all the activities that human can be involved, because human is keeping conducting certain activities either actively or passively, and most of these activities are driven by their interests. By the assist of the knowledge of user preferences of activities, service providers like recommender systems can improve their information delivery to better meet the needs of customers.

To achieve the purpose, we proposed a system to estimate user preferences of activities, which is built based on the two newly-developed topic models. The Latent Dirichlet Allocation (LDA) model was chosen as the basic model for the creation of new topic models. The LDA

model well simulates the assumed psychological process of writing tweets, and we consider that it is suitable for the knowledge discovery from twitter posts. 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. To filter out noisy activity collocations, we introduce the concept of dependency to only retain the verb-nonverb pairs with dependency relation; 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, and thus, our goal is to exploit the preferred activity and the according topic of this activity from users' twitter posts.

Therefore, the solution is to build a strong database which is composed of the tri-layer cluster containing a topic layer, an activity layer and a word layer. By referring to these tri-layer clusters, each word in a given tweet can be analyzed to figure out the according activity and topic, and then, the most mentioned activity and topic can be find out as the activity 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 word-pairs. We use the concept of dependency to find reasonable activity collocations and the concept of confidence to measure the link strength of two words in the activity collocations. The experiments show that the dependency can greatly decrease the number of collocations for calculation and effectively filter out noisy, while the confidence can optimize the clustering result by participating in the clustering process.

The clustering part is mainly composed of two series topic models: the word-pair generative LDA (wpLDA) model is for generating the topic layer and the activity layer for the clusters; the tri-layer cluster generative (TLCG) model uses the output of wpLDA model as an external input to generate the expected tri-layer clusters.

The wpLDA model assumes that the corpus is generated by word pairs instead of individual words. The advantage of constraining the clustering process in the unit of word pairs is that the topic discovery is based on the specific collocations which represent specific objects. We identify a particular collocation, that is, a verb plus a non-verb, as the expression of activity. Therefore, by extracting the activity collocations with this pre-set format, the wpLDA can calculate the link strength of the word pairs and then assign them to the according topics.

The TLCG model is the follow-up study of the wpLDA model, which can cluster corpus in the form of topic-activity clusters. We prefer to utilize the results of wpLDA to estimate user preferences of activities. The training data are text documents tagged with activity labels, but without topic labels, namely partially labeled corpus. And then, by inserting the activity node between the topic and the word nodes, the TLCG is capable of generating the expected tri-layer clusters.

To verify the proposed models, we use the perplexity and the KL-divergence to conduct the quantitative evaluation. Besides, human judgments in the form of questionnaire and cluster presentation are also utilized for the qualitative evaluations. Considering the purpose of this research, estimation accuracy of user preferences of activity is definitely an important metrics. The experimental results indicate that the proposed system works well in the clustering and the estimation of user preferences of activities.

This research can be applied into many services for personalization, such as recommendation and searching. A huge number of original tweets is needed to build the tri-layer clusters as the activity database. And then, by capturing users' real-time SNS post, we can refer to the established activity database to estimate the user preference of activities. So we can deliver the according recommendations or searching results for this user.

In the future, we plan to put the proposed system into practical application, that is, the personalized recommender system. In a real world application, large-scale data is needed to fully explore the possible activities that users may concern about. Chapter 5 provides a first glimpse of data mining technology, but for the large-scale data, we must have better-develop strategy for filtering and NLP. In the other hand, how to utilize this research result to find appropriate recommendations is significant issue. One possible method is to provide comprehensive information about the preferred topics while focusing on the key recommendations according to the preferred activities. In addition, the integration of the two topic models may be feasible to provide more effective function of clustering.