# KDD Cup 2012 Track 1 解题报告

张作柏
17300240035

2019 年 6 月 1 日

## 目录

# 1 任务简介

本次 PJ 我选择了 KDD Cup 2012 Track 1 题目[1]，其中模型主要参考了 [] 一文。

## 1.1 任务描述

近年来，随着像 Facebook、Twitter、腾讯微博等社交平台的发展，在线社交网络引起了广泛的关注。全中国最大的微博系统之一，腾讯微博，已经成为网络社交的重要平台。目前，腾讯微博拥有超过 2 亿的注册用户，每天产生四千万条信息。海量的数据引起了数据挖掘爱好者的注意，如何利用数据信息改善用户的使用体验，成为了一个十分有趣并值得研究的问题。

本任务中，我们需要根据用户的兴趣，预测他是否会关注某个对象 (item)。对象可以是某个组织、个人、群体等等。最终我们要在所有备选推荐中，选择至多三个对象推荐给用户。

## 1.2 数据信息

### 1.2.1 名词定义

**对象 (item)：**对象是腾讯微博中的一个用户，他可以代表组织、个人或群体。数据集中大约有六千个不同的对象。

**发微博 (tweet):** 发微博是指用户可以在微博系统中发表一条信息，他的关注者会看到这条信息的提醒。

**转发 (retweet)：**用户可以转发其他用户发表的信息，并在其下添加评论。

**评论 (comment)：**用户可以在别人的微博下发表评论。

**关注者 (follower)：**用户可以关注其他用户，若用户 A 关注了 B，则称 A 是 B 的关注者。

### 1.2.2 数据文件

1. **训练数据集 rec_log_train.txt：**记录了用户与对象之间的历史推荐结果。
   文件格式：(UserId) (ItemId) (Result) (Unix-timestamp)
   在 Unix-timestamp 的时间，系统向用户 UserId 推荐了物品 ItemId，得到的结果为 Result。Result 为 1，表示接受；Result 为-1，表示拒绝。

2. **测试数据集 rec_log_test.txt：**记录了测试集中用户与对象之间的可能推荐。
   文件格式同训练数据集 rec_log_train.txt
   区别在于其中 Result 域为 0，需要我们来预测。

[1] https://www.kaggle.com/c/kddcup2012-track1

3. **用户信息 user_profile.txt：**
4. **对象数据 item.txt：**
5. **用户行为 user_action.txt：**
6. **用户关注行为 user_sns.txt：**
7. **用户关键词描述 user_key_word.txt：**

## 1.3 提交格式

## 1.4 评价方式

# 2 数据预处理

## 2.1 数据清洗

## 2.2 成对训练

# 3 用户兴趣模型

## 3.1 基本模型

## 3.2 年龄、性别因素

## 3.3 间接反馈信息

## 3.4 关键词、标签信息

## 3.5 用户、物品信息

# 4 用户行为模型

## 4.1 时间度量

## 4.2 概率预测

# 5 集成学习

## 5.1 模型原理

## 5.2 训练方法

# 6 结果评估

## 6.1 测试模块

## 6.2 测试结果

# 参考文献