PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/161874

Please be advised that this information was generated on 2019-08-04 and may be subject to change.
ABSTRACT

The 2016 ACM Recommender Systems Challenge focused on the problem of job recommendations. Given a large dataset from XING that consisted of anonymized user profiles, job postings, and interactions between them, the participating teams had to predict postings that a user will interact with.

The challenge ran for four months with 366 registered teams. 119 of those teams actively participated and submitted together 4,232 solutions yielding in an impressive neck-and-neck race that was decided within the last days of the challenge.

CCS Concepts

• Information systems → Recommender systems; Data mining; Test collections;

Keywords

Recommender Systems; Data Mining Challenge; XING

1. INTRODUCTION

The ACM RecSys Challenge allows researchers and engineers around the world to jointly work on real-world recommender system problems in various domains such as social media [3] or product recommendations [2]. In 2016 the challenge was dedicated to the problem of job recommendation\(^1\) [1]. We released a dataset from XING\(^2\)—a business-oriented social network with around 18 Million users worldwide and more than 10 Million users in Germany.

XING supports people in discovering career opportunities. Job recommendations are therefore an essential part of the XING platform and its mobile apps. Those recommendations aim to satisfy the demands of both the job seekers who have certain preferences concerning their next career step and the recruiters who aim to hire the most appropriate candidate for a given job. In this challenge, we focused on the demands of the job seekers by defining the following task: Given a XING user, the recommender had to predict those job postings that a user will positively interact with by clicking on it or bookmarking it.

2. DATASET

We provided a training dataset featuring user profiles, job postings, and interactions that users performed on job posts. The dataset also incorporated user-item impressions, i.e., information about job postings that were shown to users. The training data\(^3\) contained 1,367,057 users and 1,358,098 items. Users and items were described by several similar attributes such as job categories, career level, industry, lo-

\(^1\)http://2016.recsyschallenge.com
\(^2\)http://xing.com

Figure 1: Number of interactions per user and item. 328,618 items (24.2%) and 582,370 users (42.9%) remained without interactions during the training period. More than 80% of the interactions are clicks, followed by deletes, replies and bookmarks. The distributions for training data and test data (ground truth) follow similar characteristics.
Anonymization

Anonymization of the training dataset was carried out with an iterative protector/attacker procedure. We took a simple and straightforward approach to threat modeling: The attacker profile was implicit the choice of a highly experienced Data Science team that attempted to de-anonymize the data. At each step the dataset was further bleached, and additional synthetic users were added. Then, tests were carried out to check if the dataset could be de-anonymized by the attacker and also if the dataset supported training a recommender system algorithm effective on a plain-text test set. This procedure ensured that enough information was left in the data to make this to be a useful data set for solving the problem on the actual plain-text data, while at the same time protecting the privacy of XING users.

The bleaching procedure involved replacing named entities with IDs, removing a selection of user attribute values, and removing a selection of interactions. The relative ordering of the interactions was maintained. Synthetic users were created by clustering real users. Further protections included the obvious measure of including only a fraction of XING’s users and job postings in the dataset, and also protecting the dataset legally with a user agreement that protected the dataset legally with a user agreement that

3 https://github.com/recsyschallenge/2016/#evaluation

4. REFERENCES

