DATA 643: Recommender Systems

Discussion 4: Attacks on Recommender Systems

Walt Wells, Summer 2017

Overview

Recommender Systems that utlize Collaborative Filtering are particularly vulnerable to troll attacks or sybil attacks. It's causing issues across many social or collaborative systems, not just recommender systems. A good example includes the 'bot-itisation' of Twitter.

Essentially, business or political interests have created large armies of Twitter bots to attempt to sway opinion or user behavior.

Some good overview links include:

- Robert Mercer: the big data billionaire waging war on mainstream media
- First! How '#WhoIsNeil' Gamed Trump's Twitter Bots

In an interesting paper recently published studying injection of trolls into social or collaborative networks, noted that the size of network (or in the case of recsys, magnitude of the ratings utility matrix) is a primary factor.

> "Our results showed that small amounts of trolls have a higher impact when connecting to users in the networks' periphery, as those users receive and exercise less peer influence and cannot compensate for the negative influence of trolls as well as highly connected and highly active users can. However, larger amounts of trolls influence activity levels more when performing informed selection of high-degree users. While these users—building the core of the networks—are able to compensate for the trolls' influence longer, overall activity is drastically reduced once high-degree users are infected and start spreading unproductive activity themselves."

Potential RecSys Defense: Anomaly Detection

The 2010 paper from UC London, Temporal Defences for Robust Recommendations, outlines some interesting options for handling attacks from sophisticated sybils attacking a deployed RecSys.

The authors ultimately recommended implementing and training a predictive model for anomaly detection at the user, item and global levels. The authors also note the difficulty of managing a patient or slow attack against a system using this defense. For example, if a new user profile is created, and they create a bunch of 'false' ratings at once, it can be easy to flag this activity as fraudulent. If the profile(s) inject 'false' ratings over time it can be more difficult to detect.

ASIDE: My inner sci-fi nerd wants this type of study to be named after the Holtzmann effect from Frank Herbert's Dune.

> "The shield turns the fast blow, admits the slow kindjal!"

One iterative option for implementing recsys defenses would be to setup a kind of temporal firewall. Incoming user signals could be pooled in a kind of 'holding pen' before being injected back into the full system dataset. Then a kind of pooled anomaly detection could be run over that data to test for attacks. Any user data that passes could be assimilated back into the system dataset. Any rejected could be flagged for further review.

Of course, for a RecSys to implement defenses of any kind there has to be buy-in from the decision leadership and an acknowledgement that this type of attack fundamentally threatens the well being of the system. As of

this posting, there is still no indication that Twitter (while not fundamentally a RecSys) cares that it is a breeding ground for bots and trolls.

References

- Wisdom of the crowd? IMDb users gang up on Christian Bale's new movie before it even opens.
- Google's anti-trolling AI can be defeated by typos, researchers find
- Temporal Defences for Robust Recommendations
- Exploring the Impact of Trolls on Activity Dynamics in Real-World Collaboration Networks
- Robert Mercer: the big data billionaire waging war on mainstream media
- First! How '#WhoIsNeil' Gamed Trump's Twitter Bots