Выявление ботов

Илья Козлов

6 декабря 2016 г.

Содержание

Боты и откуда они берутся

Боты появляются в социальных сетях почти сразу после их создания. Так в твитере рассылка спама была зафиксирована через неделю после его открытия[1]
Боты регистрируются и управляются (полу)автоматически. Как правило основная функция ботов — рассылка спама (реклама), но не обязательно. Есть у ботов и другие функции.

Чем боты занимаются

Рассылка спама

Отправка личных сообщений, публикация постов рекламного характера. Это одна из наиболее раздражающих деятельностей ботов, часто методы выявления ботов расчитаны как раз на выявление ботов–спамеров.

Астротурфинг

Явление астротурфинга названо в по имени компании, производящей искусственный газон.
Суть явления — создания видимости общественной поддержки (или наоборот). Эта поддержка может заключаться в оставление комментариев, вступление в группы (которые необходимо поддержать), выражение поддержки другими методами (лайки, репосты)

Другие функции ботов

Кратко про другие применения

- Поддержка других ботов — боты “дружат” друг с другом, для того, чтобы лучше мимикрировать под настоящих пользователей.
- Мошеннические действия — боты могут выдавать себя за знакомых пользователя и просить перевести деньги на их счёт.
- Сбор информации — как правило социальные сети предоставляют больше информации “друзьям” пользователя (или друзьям друзей). Боты могут помочь в сборе информации, добавляясь в друзья к пользователям.
- ...

Обнаружение и борьба с ботами. Ручные методы

Боты рассылают спам. Как правило социальная сеть позволяет пожаловаться на рассылку спама и администрация сети может блокировать такого пользователя.
Подобный подход не лишён недостатков:

- Боты блокируются *после* совершения нежелательных действий (спам уже разослан)
- Таким образом можно выделить не всех ботов, например астротурфинг подобными методами не выделяется.
- Блокировка пользователей в ручном режиме может быть трудозатратной (необходимо проанализировать большое количество жалоб)

Автоматические методы

Боты управляются автоматически или полуавтоматически.
Люди управляют своими аккаунтами “в ручную”.
Таким образом в поведении людей и ботов существуют отличия, отличия можно выявлять, в том числе и методами машинного обучения.
Для построения обучающей выборки можно использовать информацию о блокировках пользователей.

Обнаружение и борьба с ботами. Машинное обучение

Попытки обнаружить ботов методами машинного обучения предпринимались давно. Как правило для обнаружения ботов использовалось большое количество эвристических и полуэвристических признаков:

1. Использование специфических слов
2. Использование специальных слов (хештеги, упоминания пользователей)
3. Наличие закономерностей во времени публикации сообщений, не характерных для людей
4. Соотношение репостов и обычных сообщений и т.п.

Эволюционная гонка вооружений

Новые методы борьбы вынуждают владельцев ботов адаптироваться к новым условиям и лучше мимикрировать под обычных пользователей.
Например добавлять случайные временные паузы в отправку сообщений, . . .
Признаки, характерные для одних ботов (спамеров) может быть не характерно для другого типа ботов(астротурфинг).
Важно не только эффективность определения существующих ботов, но и то, насколько легко боты смогут адаптироваться

Особенности графа

В работе [2] было показано, что ботов имеет особенности:

- Боты пытаются образовать социальные связи со случайными пользователями. Это позволяет им лучше распространять информацию. Как правило боты имеют больший процент отказов.
- Боты могут образовывать социальные связи друг с другом, образовывать плотные сообщества[2].

Владельцы ботов не могут в полной мере контролировать структуру социального графа.

Использование социального графа.

Мотивация

- Есть основания предполагать, что структура графа обычных пользователей и ботов отличаются. Если это так, то можно использовать граф выявления ботов.
- Владельцы ботов не могут менять структуру графа произвольным образом. К такому методу труднее адаптироваться.
- Социальный граф занимает относительно небольшой объём (относительно других данных в социальной сети, например сообщений)

Сложность в анализе графов

Большинство классических алгоритмов классификации (SVM, Xgboost, . . .) не могут работать с графом непосредственно. Для применения алгоритмов классификации необходимо выделить признаки т.е. сопоставить каждой вершине вектор конечной размерности.

Embedding

Представление вершин графа в виде вектора производятся с помощью алгоритмов Graph Embedding.
В нашей работе использовалась модель BLM

Модель 1 Bilinear Link Model
Общая идея

Каждая вершина имеет скрытое состояние, отвечающее за вероятность образования рёбер.
На основе скрытых состояний описывается вероятность наблюдать граф, который мы наблюдаем.
С помощью принципа максимума правдоподобия мы находим скрытые состояния вершин, а затем используем из в качестве признаков для классификатора.

Формальное описание

У каждой вершины графа u есть скрытое состояние, представленное двумя векторами $\mathbf{In}_u$ и $\mathbf{Out}_u$. Зададим вероятностную модель:

Вероятность ребра из u в v

$$p(v|u) = \frac{\exp\{\mathbf{In}_u \cdot \mathbf{Out}_v\}}{\sum_{w \in V} \exp\{\mathbf{In}_u \cdot \mathbf{Out}_w\}}$$

Вероятность наблюдать ребро $u \to v$:

$$p(u, v) = p(u)p(v|u)$$

Вероятность получить граф, который мы получили

$$\ln P(G) = \sum_{u,v \in E} \ln(p(v|u)) + \ln p(u) \to \max$$

Классификация вершин

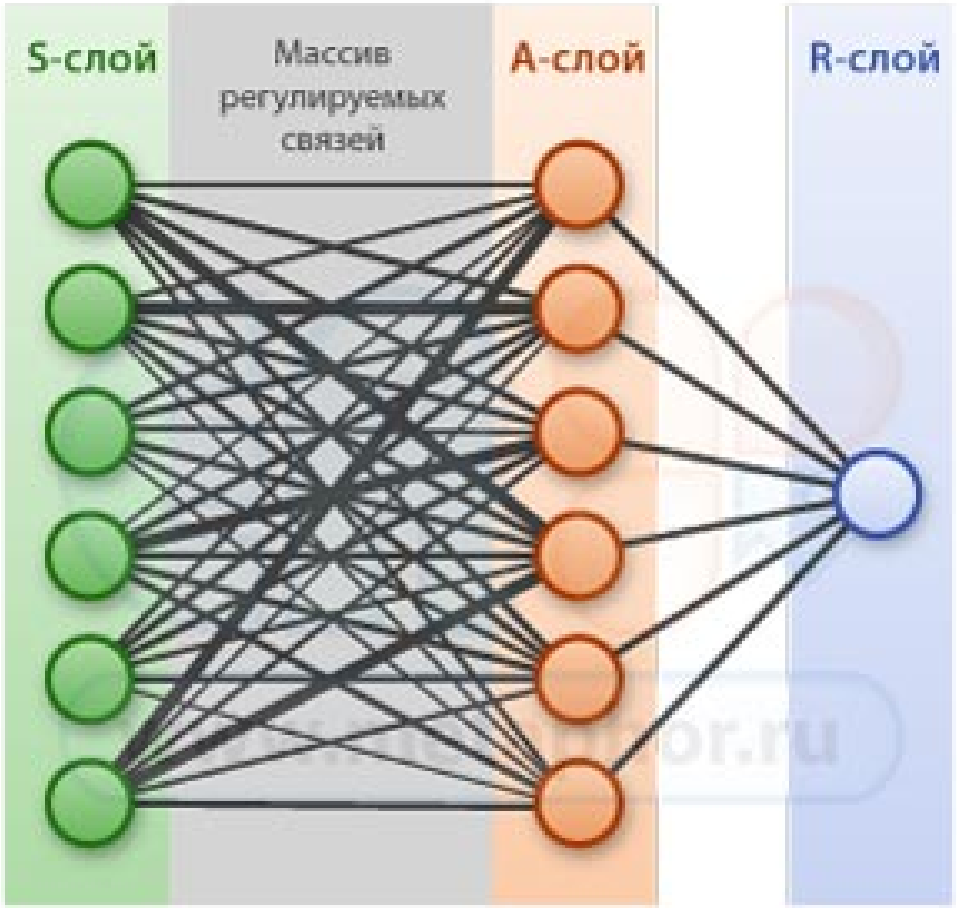
После того, как мы получили векторное представление вершин можно решать задачу классификации т.е. сопоставлять векторному представлению вершины класс

$$\vec{u} \rightarrow \{\text{Бот, Не\_бот}\}$$

Так как мы не можем достоверно определит ботов (в ручную), использовалось предположение, что боты это те аккаунты, которые подвергаются блокировке. Таким образом система обучалась предсказывать блокировки.

Нейронные сети

После того, как мы получили векторное представление вершин можно использовать любой алгоритм классификации, мы использовали многослойный перцептрон.



Качество работы

Мы моделировали работу систем с помощью сокрытия меток

$$AUC = 0.76$$

(т.е. с $p = 0.76$ если взять бота и не бота, то наша система поставит бота “выше”)

Ссылки

BLM

Исходный код — https://github.com/tigvarts/BLM
Статья http://link.springer.com/chapter/10.1007/978-3-319-26123-2_19 Статья для ленивых http://sci-hub.cc/10.1007/978-3-319-26123-2_19

DeepWalk

Исходный код https://github.com/phanein/deepwalk
Статья https://arxiv.org/abs/1403.6652
Применение в dde
http://perozzi.net/publications/15_www_age.pdf

Библиография

Pranam Kolari, Tim Finin, and A Joshi.
Spam in blogs and social media.
Tutorial at ICWSM, pages 307–361, 2007.

Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu.
Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter.
In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80. ACM, 2012.