Выявление сообществ в Stackoverflow

Поляков Семен

руководитель Д.А. Калашников

СПб АУ НОЦНТ РАН

12 июня 2017 г.

Введение

- Stackoverflow - популярная система вопросов-ответов для IT задач
- Более 7 миллионов пользователей
- Более 35 миллионов постов
- В постах проставлено не менее 48 тыс. тегов

Введение

Определение

Тег Stackoverflow - метка, для обозначения информационной технологии, либо термина, относящегося к той или иной информационной технологии.

Определение

Пост - набор из текста, автора, списка тегов, и нескольких метрик популярности поста

Кластеризация тегов

Из коллекции таких постов хотим получить кластеры вида:

№	Кластер тегов
1	android, android-wear, andengine, apk, ..
2	java, java-ee, junit, aop, guice, ..
3	python, pickle, virtualenv, pip, ..
4	...

,

то есть каждому тегу сопоставить номер его кластера.

А также для каждого кластера узнать количество пользователей, его использующих.

Такую задачу можно отнести к сфере исследований под названием Software repository mining.

Кластеризация тегов

Особенности текущих исследований:

- Stackoverflow - Andrej Gajduk et al. Intelligent tag groupping by using an aglomerative clustering algorithm, 2013 (дамп Stackoverflow за 2011, большая временная сложность)
- Twitter - Cristina Ioana Muntean et al. Exploring the Meaning behind Twitter Hashtags through Clustering, 2016 (обработка естественного языка)
- Grigory Begelman et al. Automated Tag Clustering: Improving search and exploration in the tag space, 2006 (не самый производительный алгоритм кластеризации графа)

Цель и задачи

Цель: выявить основные группы тегов на Stackoverflow и оценить количество пользователей в них

Задачи:

1. Подготовить данные Stackoverflow для лучших результатов кластеризации
2. Сравнить качество различных методов кластеризации на данных Stackoverflow
3. Оценить количество пользователей в каждом из получившихся кластеров

Подготовка данных

- Тегов много, их количество влияет на размерность задачи.
  - Заменим непопулярные (< 1000 использований) на более популярные, которые часто идут с ними в паре
- Пользователей много, не все явно относятся к тому или иному сообществу
  - Отсекаем по рейтингу
  - Отсекаем тех, кто слишком редко пишет
  - Выбираем активных пользователей (с положительными метриками, присущих качественным ответам)

Входные данные

Для кластеризации тегов будем использовать набор постов (35 млн.), каждый из которых описывается следующими признаками:

- Список тегов (как правило не более 10)
- Количество просмотров
- Количество ответов
- Рейтинг поста
- Средний рейтинг ответов
- Идентификатор автора

Итоговый объем обрабатываемых данных ~2 Гб.

Подходы к кластеризации тегов

Подходы:

- Выявление кластеров в графах (много алгоритмов)
- Тематическое моделирование - BigARTM (обобщение существующих LDA и PLSA)
- Комбинированный подход

Подход №1: Построение графа

- Построим граф, где теги - вершины, а ребра взвешены некоторой мерой схожести тегов
- Были придуманы несколько мер на основе меры Жаккара и отобраны те, которые дают наибольший вклад в качество кластеризации
- Линейная агрегация мер в вес на ребрах - подобор коэффициентов также через оптимизацию метрики качества

Выявление кластеров в графах

Пусть n - число вершин графа, m - число ребер.
Постановки задачи кластеризации и алгоритмы, которые нам **не подходят**:

- Разбиение графа (не известны размеры кластеров)
- Иерархическая кластеризация (Время работы: $O(n^2 log(n))$)
- Разделяющие алгоритмы
  - Girvan and Newman (Время работы: $O(m^2 n)$)
  - Radicchi et al - (Время работы: $O(\frac{m^4}{n^2})$)
- Оптимизация модулярности (точная - NP):
  - Симуляция отжига (Время работы $\sim O(n^2)$)
  - Экстремальная оптимизация (Время работы: $O(n^2 log(n))$)
- Случайные блуждания:
  - Алгоритм Zhou (Время работы $O(n^3)$)
  - NetWalk (Время работы $O(n^3)$)
- И другие

Выявление кластеров в графах (продолжение)

Постановки задачи кластеризации, которые нам **подходят**:

- Спектральная кластеризация
  (Время работы: $o(n^3)$, на практике - быстрее)
- Оптимизация модулярности
  - Жадный алгоритм (Fast Greedy,
    Время работы: $O(m)$)
- Случайные блуждания:
  - Алгоритм марковской кластеризации

  (Время работы: $O(nk^2)$, где k $\ll n$)

Оценка качества кластеризации

Метрика качества для кластеризации графов:

- Модулярность

Общие метрики качества:

- Индекс Девиса-Боулдина (DBI)
- Для сравнения с заданной, ”истинной” кластеризацией:
  - Нормализованная взаимная информация (NMI)
  - Исправленный индекс Рэнда (ARI)

Подход №2: тематическая модель

- Тематическая модель - модель коллекции документов, которая определяет, к каким скрытым темам относятся документы из коллекции. Определяется матрицами Φ и Θ, задающие вероятности $P(w|t)$ и $P(t|d)$, где w - слово, t - тема, d - документ.
- В случае кластеризации тегов примем w за тег, t - кластер тегов, d - пост.
- Обобщение наиболее известных тематических моделей - APTM (Аддитивная регуляризация тематических моделей). Позволяет находить матрицы с учетом регуляризаторов через EM-алгоритм.
- BigARTM - фреймворк для использования APTM.

Подход №3: комбинированный

- Поскольку была построена мера схожести тегов $S(w_i, w_j)$, используем ее в качестве дополнительного регуляризатора
- А именно возьмем для основных пар тегов распределения $P(t|w)$ и посчитав между ними статистическое расстояние получим дополнительное слагаемое регуляризации:

$$\sum_{i,j} JSD(p(t|w_i) \,||\, p(t|w_j))S(w_i, w_j) \to min_{\Phi}, p(t|w_i) = \phi_{w_i t},$$

$$JSD(P \,||\, Q) = \frac{1}{2}KL(P \,||\, M) + \frac{1}{2}KL(Q \,||\, M),\ M = \frac{1}{2}(P + Q),$$

где JSD - симметричное расстояние Дженсона-Шеннона

- Обозначим такую регуляризацию как RegSim

Результаты кластеризации тегов

Алгоритм	Модулярность	DBI	NMI	ARI
Fast Greedy	0.05	1.97	0.28	0.29
Spectral clustering	0.17	2.4	0.35	0.31
Markov Clustering	0.27	2.46	0.73	0.54
BigARTM			0.32	0.18
BigARTM + RegSim			0.86	0.72

Где у метрик качества:

- Модулярность: больше - лучше (от 0 до 1)
- DBI: меньше - лучше (от 0 до ∞)
- NMI: больше - лучше (от 0 до 1)
- ARI: больше - лучше (от 0 до 1)

Кластеризация пользователей

Для каждого пользователя:

1. получим набор кластеров тегов в которые он пишет
2. выберем самый часто используемый им кластер в качестве основного
3. определять кластер пользователя можно во временном окне по его постам

Таким образом получим набор временных рядов для каждого кластера тегов

Результаты

1. Были рассмотрены существующие подходы и методы кластеризации тегов
2. Оценена эффективность выбранных методов, в результате которой лучшим алгоритмом кластеризации графов оказался Марковский алгоритм.
3. Предложен и опробован комбинированный подход (BigARTM + RegSim), который дал лучшие результаты при оценке NMI для заранее заданной базовой кластеризации.
4. Выявлены сообщества для каждого из получившихся кластеров

Сообщества Stackoverflow

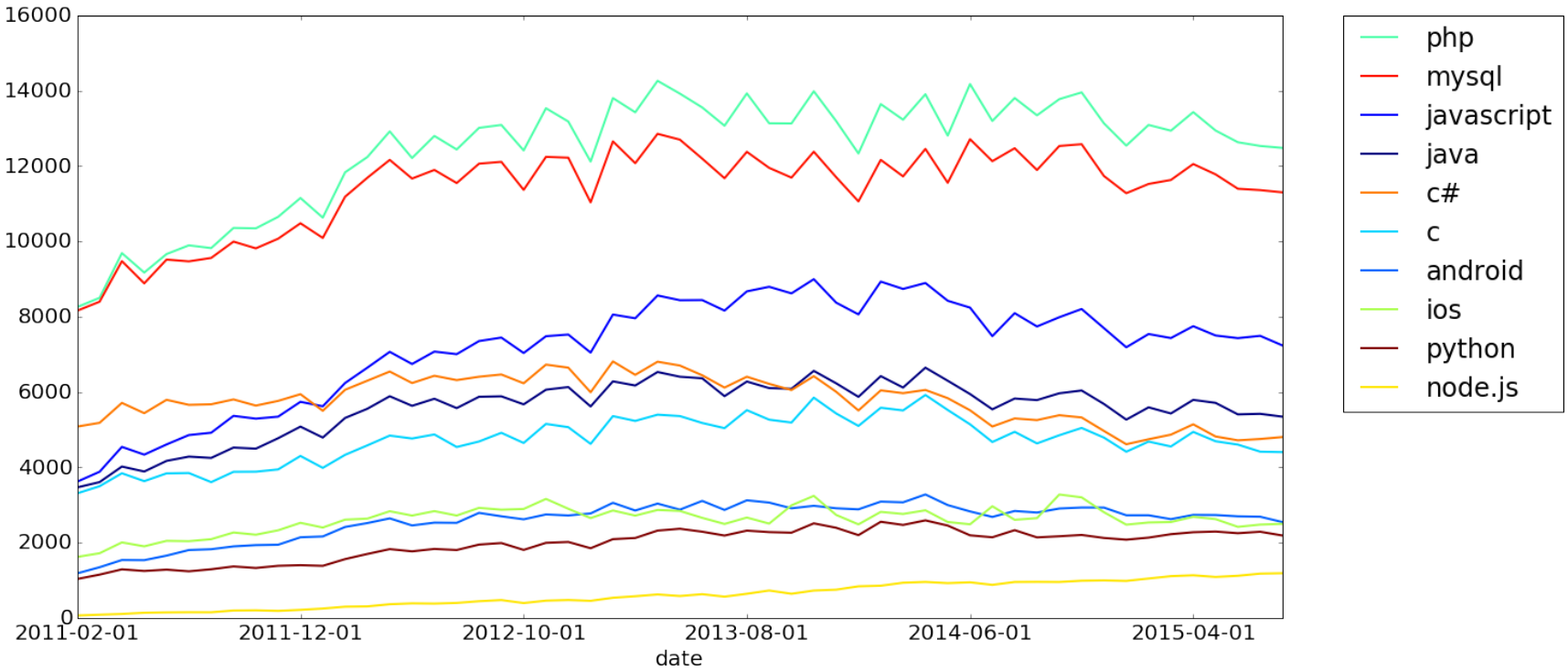

Рис.: Тренды сообществ Stackoverflow

Конец

- Ссылка на проект: `https://bitbucket.org/autumninternsjb/stackoverflowcommunityanalysis`
- Вопросы?