Классификация записей из микроблогов с помощью Википедии

Абишев Т. М.
545 группа, математико-механический факультет, СПбГУ

Научный руководитель: Барашев Д. В.
Доцент, математико-механический факультет, СПбГУ

Конференция СПИСОК
СПбГУ, Санкт-Петербург, апрель 2012

Введение

- Данных всё больше
- Микроблоги как пример таких данных
- ...Которые необходимо классифицировать
  - зачем?
- И которые имеют особенности
  - какие?

Идеи

- Википедия, как источник помощи при классификации
  - категории
  - ссылочная структура
- Контекст
  - о чем еще пишет автор данного сообщения?

Проблемы

- Размер википедии
- Структура категорий
  - совсем не дерево
  - и даже не ацикличный граф
  - *Programming languages*?

Текущие результаты (1)

- Попытка классифицировать сообщения из микроблогов на имеющие отношения к программированию и не имеющие
- Попытка классифицировать пользователей на основе полученной информации

Текущие результаты (2)

- Программисты
  - sstephenson: 16.153254370938036
  - joehewitt: 2.04975624651863
  - joestump: 4.142122742967949
- Не программисты
  - katyperry: 0.3060912490764008
  - ladygaga: 0.9178572404672108
  - BarackObama: 0.2046168083345868 4

Дальнейшее развитие

- Решение проблемы со структурой категорий
  - не все родительские категории одинаково важны
- Классификация сообщений на всё дерево категорий
- Использование полученных результатов в классификации на произвольные категории