Извлечение описаний событий предопределённых типов из потока сообщений пользователей микроблогов

Ефим Пышнограев

Филиал МГУ в городе Севастополе

22 апреля 2014

Постановка задачи

Данные:

- множество документов (сообщений) $\Omega = \left\{ D_i \mid i \in \overline{1, n} \right\}$,
- документ — множество слов и временная метка t_i $D_i = \left\{ w_j \mid j \in \overline{1, l_i} \right\}$.

События:

- резкое увеличение частоты ключевых слов, затем спад до нормального уровня,
- должно быть вызвано реальным событием,
- не носит периодический характер.

Особенности сети Твиттер:

- короткие сообщения (до 140 символов),
- наличие шума и ошибок,
- большая плотность сообщений,
- “взрывной” характер событий.

Схема алгоритма

Составные части:

- нахождение максимума в частотной функции, который будет соответствовать событию,
- извлечение ключевых слов, характерных этому максимуму,
- применение модели HDP с частичным обучением для того чтобы выделить все сообщения этой темы,
- проверка насколько полученный результат соответствует новому событию.

Результаты

Заключение