



МЕДИАЛОГИЯ

ПРОБЛЕМЫ КЛАСТЕРИЗАЦИИ НОВОСТНОГО ПОТОКА

Петр Воляк

pvolyak@mlg.ru

NLPseminar, Санкт-Петербург

8 октября 2011 года

О КОМПАНИИ И ТЕХНОЛОГИЯХ



О компании «Медиалогия»

www.mlg.ru
780-9040

- специализируется на технологиях лингвистического анализа текстовой информации
- российский лидер в области разработки онлайн-решений для мониторинга и анализа СМИ
- на рынке с 2003 года



Что такое система «Медиалогия»

www.mlg.ru
780-9040



БАЗА СМИ

+



**ТЕХНОЛОГИИ
ОБРАБОТКИ И
ИЗВЛЕЧЕНИЯ ДАННЫХ**

=



**СИСТЕМА ОНЛАЙН
АНАЛИЗА СМИ**



ТВ-программы (252)



Журналы (583)



Информгентства (280)



Блоги (952)



Радиопередачи (80)



Газеты (1 607)



Интернет (2 515)

6 269 СМИ



Данные на 09.09.11



Федеральные	364
Региональные	2 409
ЦФО	656
ПФО	507
СФО	320
СЗФО	275
ЮФО	194
УФО	240
ДФО	153
СКФО	64
СНГ и Балтия, в т.ч.	393
Украина	197
Казахстан	67
Зарубежные	242
Отраслевые	1 754
Блоги	952
Глянec	155



Данные на 26.08.11



- Агрегация онлайн- и оффлайн-СМИ, а также соцмедиа (блоги, форумы, соц.сети) в режиме реального времени
- Классификация и кластеризация потоков информации
- Выделение именованных понятий
- Мониторинг и анализ
- Визуализация результатов мониторинга и анализа



Named Entity Recognition

www.mlg.ru
780-9040

- Выделение позиций
- Соотнесение с базой объектов (персоны, организации, бренды, геопонятия)
- Работа правил
- Ранжирование объектов на позиции (в том числе с неизвестным)
- Подсветка



На том же этапе

- Выделение прямой и косвенной речи
- Жанровая классификация
- Рубрикация
- Выделение фактов и связей
- Далее - кластеризация



- Мониторинг упоминаний объектов в СМИ
- Генерация периодических отчетов
- Различные продукты с новостной картиной дня
- Мониторинг блогосферы и соцмедиа

КЛАСТЕРИЗАЦИЯ



- Нормализуем лексику в документе, выкидываем стоп-слова
- В каждом документе выделяем топ по TF-IDF
- Подсвечиваем документы
- По векторам слов и объектов строим расстояния между документами
- Если расстояние меньше заданного радиуса, документы попадают в один кластер
- Также по расстоянию можно выделить плагиаты и дубликаты



- **Непрерывная кластеризация:** анализ вновь поступивших документов и включение их в уже имеющиеся кластера, проверка схожести с независимыми документами для последующего объединения в новый кластер
- **Дискретная перекластеризация:** периодически из имеющихся кластеров выбираются те, которые были обновлены с момента последнего процесса перекластеризации, затем выбранные кластеры проверяются на возможность объединения или разбиения
- **Проверка на связанность:** количество документов, с которыми связан вновь добавляемый в кластер документ, деленное на общее количество документов в кластере является связанностью документа, которая должна быть больше/равна по величине связанности кластера - усредненной связанности документов в кластере



- Влиятельность источника
- Свежесть
- Максимальная связанность с другими документами кластера
- Заголовок выбирается из документов, непосредственно связанных с главной статьей



- Большие кластера, собирающиеся вокруг похожих событий (стихийные бедствия, происшествия, биржевые котировки)
- Плохое деление на подкластера в случае масштабных событий
- Недостаточная точность работы алгоритма выбора заголовка
- «Мусорные» документы в кластерах



- Отдельный вектор с биграммami
- Учет биграмм в лексических векторах
- Точное определение географии
- Подключение тезауруса с синонимами
- Подключение модуля коррекции опечаток



- Сбор данных о географии:
 - Объекты
 - Прилагательные
 - Онтологические связи
- Определение локации с помощью геобазы:
 - Иерархия
 - Система координат



Выбор заголовка (задача)

www.mlg.ru
780-9040

лексика – отсутствие оценочной, жаргонной, ненормативной лексики

Например:

(хорошо) *1 января для водителей московских такси вводится обязательная лицензия*
(плохо) *Зимой столичные бомбилы попадут на новые штрафы*

объекты – в заголовке должны фигурировать главные участники сюжета + фактическая информация наиболее полно

Например:

(хорошо) *ВТБ заявил о продаже «Газпрому» 70% акций «Связьбанка» за \$100 млн*
(плохо) *Крупнейший госбанк продает свою дочку*
(хорошо) *Председатель фракции «Справедливая россия» в Госдуме Николай Левичев сложил полномочия*
(плохо) *Левичев заявил об уходе*

уровень обобщенности – заголовок должен описывать сюжет в общем, а не его фазу или деталь

Например:

(хорошо) *При взрыве в «Домодедово» пострадало несколько десятков человек*
(плохо) *Два харьковчанина числятся пропавшими после теракта в Москве*

знаки препинания – заголовок не должен состоять из нескольких предложений, крайне нежелательны символы «тире», «двоеточие», восклицательный и вопросительный знаки

Например:

(хорошо) *Президент России обсудил спортивное образование в школах*
(плохо) *Медведев: Самое лучшее – детям!*



Выбор заголовка (критерии)

- Длина - в районе 50-70 символов
- Наличие ключевых слов и объектов
 - а) из других заголовков кластера
 - б) из первых абзацев статей в кластере
- Источник – с максимальным весом
- Вес статьи внутри кластера (близость к ядру)
- Считать статистику только по уникальным заголовкам
- В конце заголовка не должно быть знаков препинания
- В заголовке должен быть глагол (боремся с такими заголовками, как «Авария в центре Москвы», «Беспорядки в Лондоне» и т.п.)
- Заголовок не должен состоять только из заглавных букв
- В заголовке не должно быть менее 3 слов



Спасибо за внимание!

www.mlg.ru
780-9040

Воляк Петр

Компания «Медиалогия»

Руководитель направления лингвистических
решений

+7 (916) 534 35 61

pvolyak@mlg.ru

127018, Москва, ул. Складочная, д.3, стр.1

телефон/факс: +7 (495) 780 90 40