Метод ранжирования информационных источников по степени доверия

Выполнил: Павелко Павел Юрьевич, ИУ7-81
Руководитель: Бекасов Денис Евгеньевич, ИУ7

МГТУ им. Баумана

Москва, 2017

Цель

Разработка системы мониторинга новостей с последующим ранжированием источников по степени доверия

Задачи

1. Проанализировать предметную область
2. Разработать метод ранжирование источников
3. Разработать программное обеспечение
4. Исследовать применимость метода

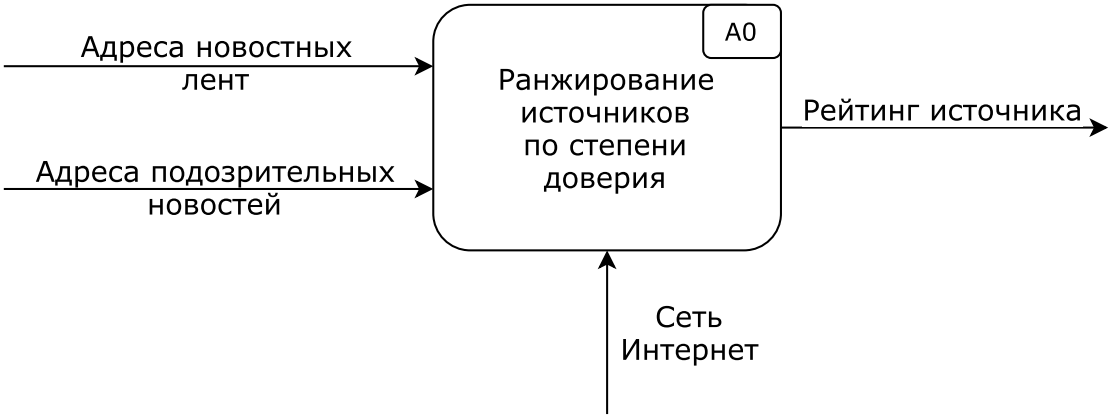
Актуальность

1. Более 100 тыс. новостных сообщений ежедневно от почти 7 тыс. значимых источников
2. Достоверность новостей вызывает сомнения
3. Эксперты способны проверить лишь новости популярных источников

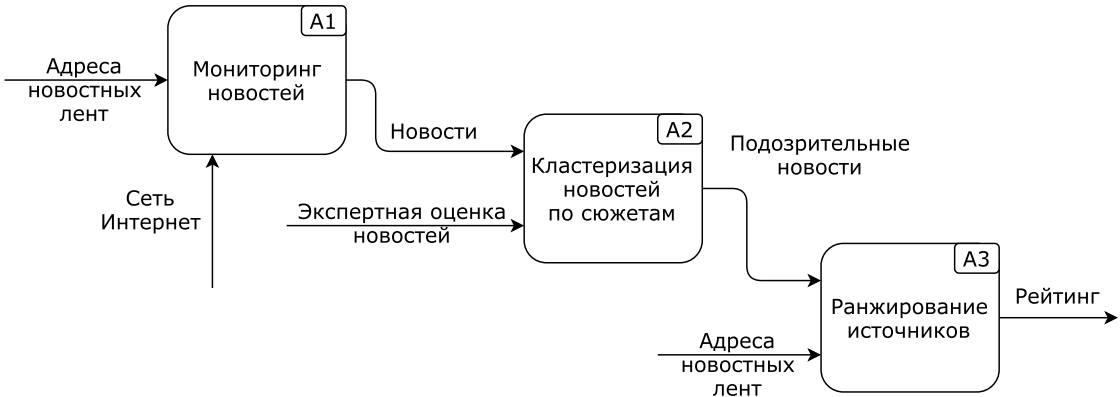
Вывод

Необходим метод распространения оценки эксперта на схожие новости менее популярных источников с последующим ранжированием источников по степени доверия

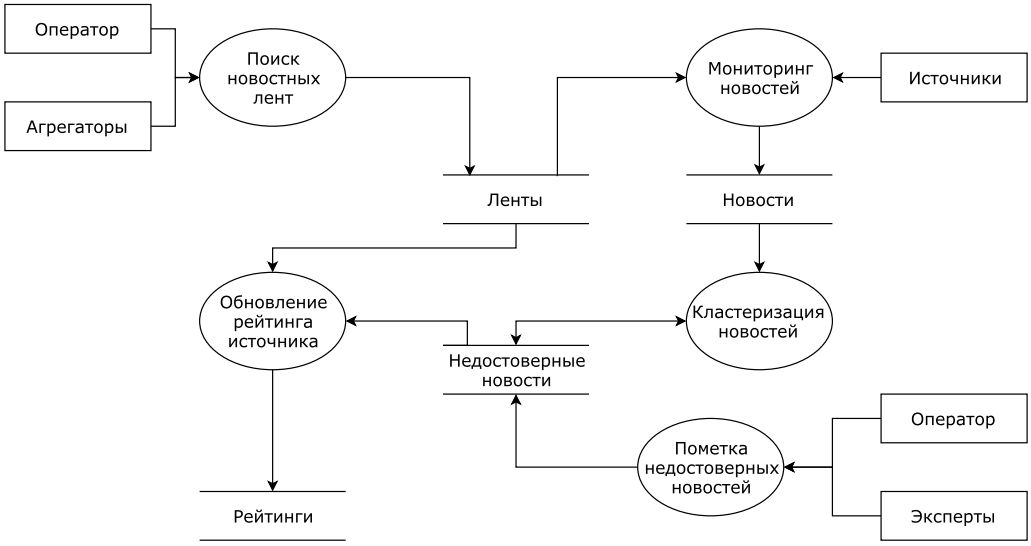
Постановка задачи



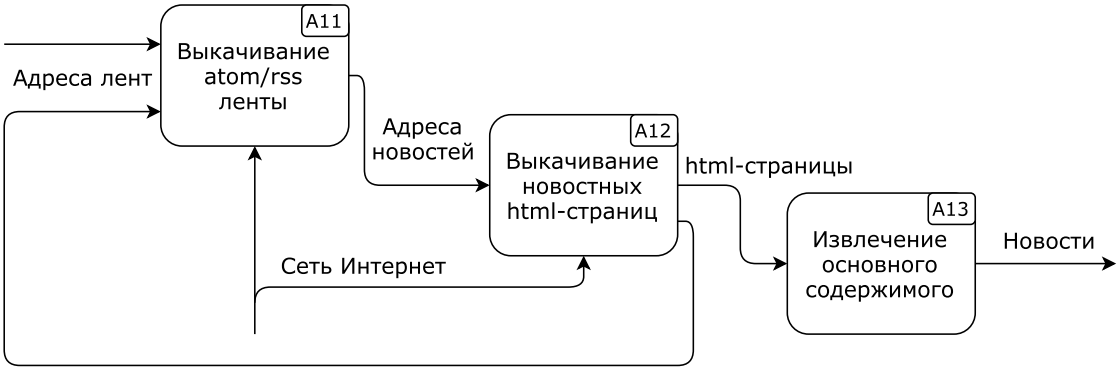
Предложенный метод



Потоки данных в системе



Мониторинг новостей



Извлечение содержимого

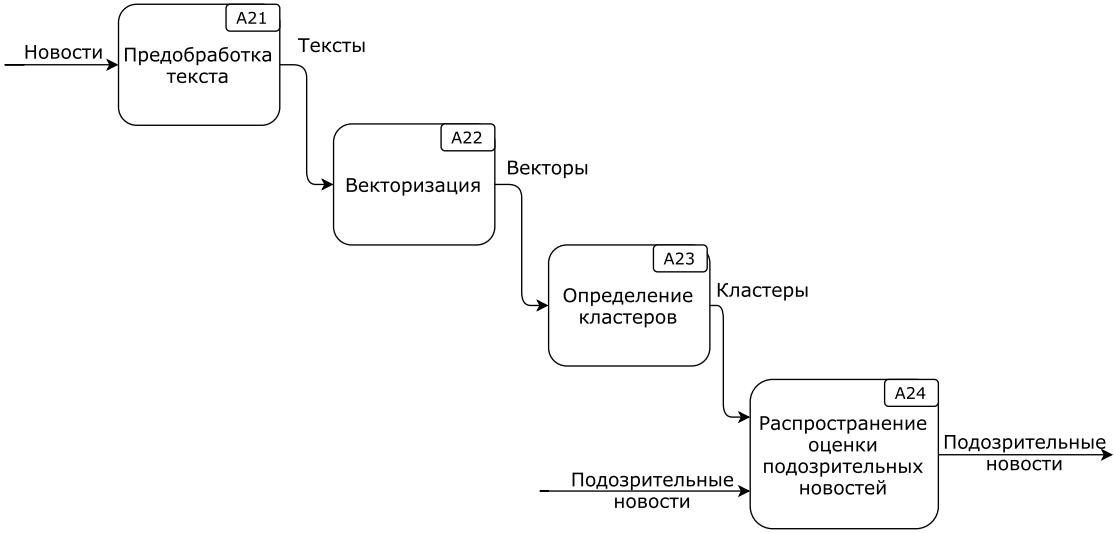
Readability Algorithm

1. Оценка узлов DOM-дерева по различным показателям
2. Ранжирование узлов по оценке
3. Извлечение текстового содержимое наиболее значимых узлов

Основные показатели

- Длина текста и доля ссылок в тексте
- Количество запятых в тексте
- Доля изображений, списков и т.д.
- Классы, идентификаторы и теги

Кластеризация новостей



Выбор алгоритма кластеризации

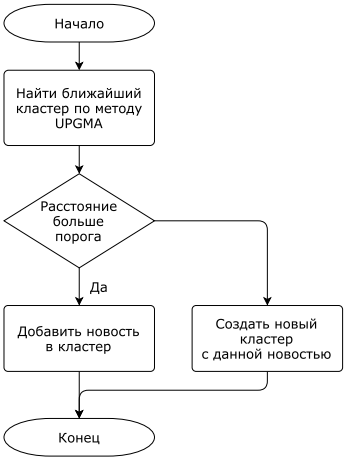
	k-m	HAC	DHCA	ICA
Адаптивное кол-во кластеров	−	+	+	+
Иерархические кластеры	−	+	+	−
Онлайновый алгоритм	−	−	+	+
Возможность оптимизаций	−	−	−	+

Алгоритмы

k-m — K-means
HAC — Hierarchical Agglomerative Clustering
DHC — Dynamic Hierarchical Compact Algorithm
ICA — Incremental Clustering Algorithm

Incremental Clustering Algorithm

Добавление новости:



UPGMA:

$$upgma(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} sim(\mathbf{x}, \mathbf{y}),$$

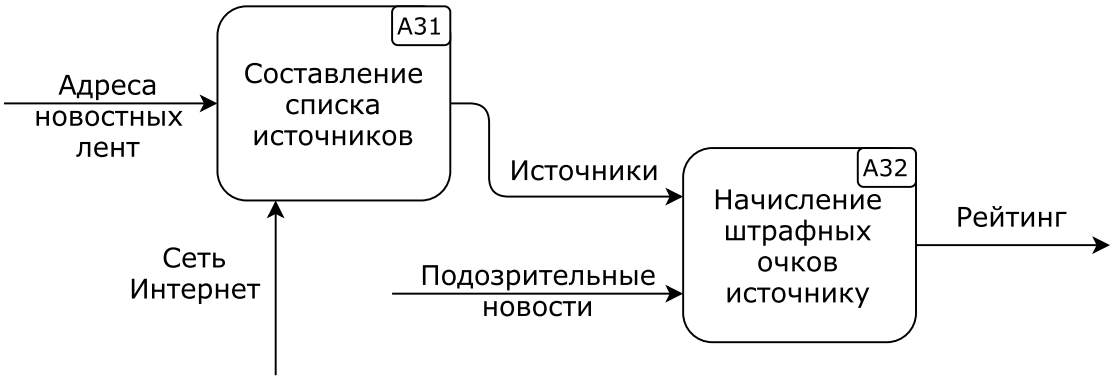
где C_i, C_j — кластеры

$\mathbf{x}, \mathbf{y}$ — новости как «мешки слов»

$sim(\cdot, \cdot)$ — мера схожести новостей:

$$sim(\mathbf{x}, \mathbf{y}) = \mathbf{xy} \cdot (1 - penalty(x, y))$$

Ранжирование источников



Исследование влияния выбора меры сходства на качество кластеризации

TODO

1. постановка
2. цель
3. размер выборки

Выводы из работы

1. Проанализирована предметная область
2. Разработан метод ранжирования источников
3. Разработано программное обеспечение
4. Исследована применимость метода

Дальнейшее развитие

- Построение тематического рейтинга источников
- Агрегированная экспертная оценка
- Ранжирование экспертов
- Нахождение дубликатов и определение первоисточника
- Связывание сюжетов по тематике