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 ранжирования источников
3. Разработать программное обеспечение
4. Провести исследование для выбора параметров метода

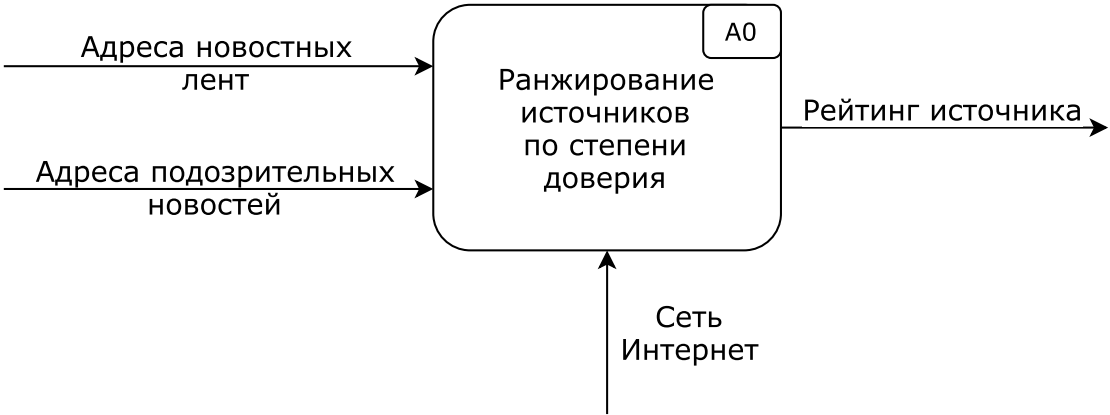
Актуальность

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

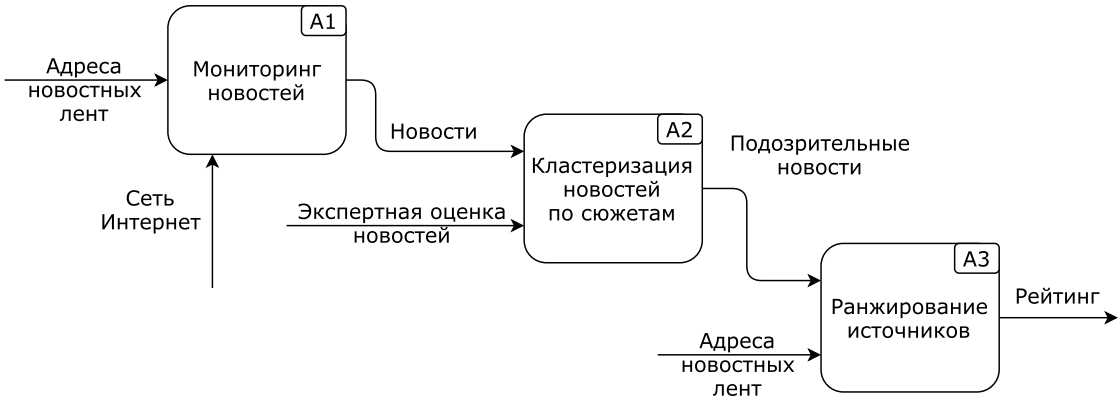
Вывод

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

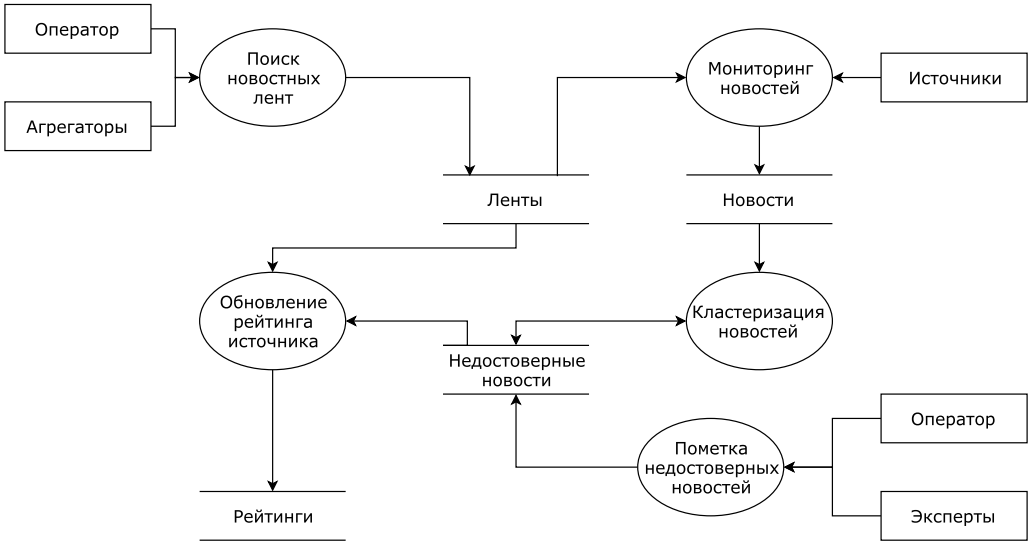
Постановка задачи



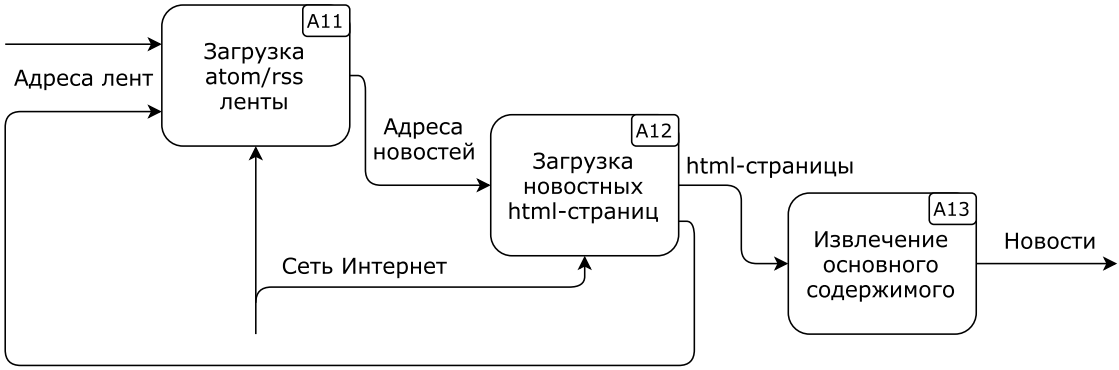
Предложенный метод



Потоки данных в системе



Мониторинг новостей

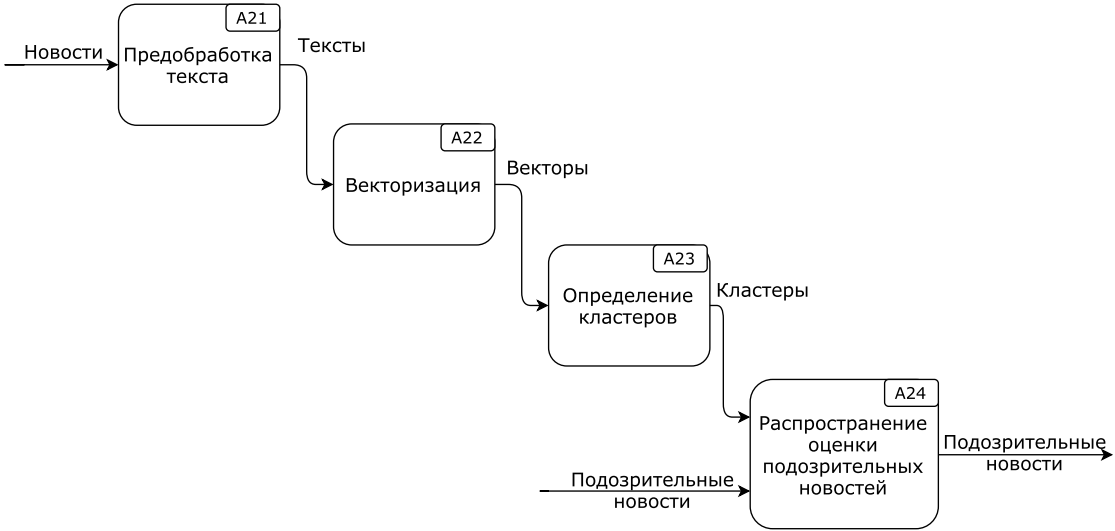


Извлечение содержимого

Readability Algorithm

1. Оценка узлов DOM-дерева по различным показателям:
   1. Длина текста и доля ссылок в тексте
   2. Количество запятых в тексте
   3. Доля изображений, списков и т.д.
   4. Классы, идентификаторы и теги
2. Ранжирование узлов по данной оценке
3. Объединение наиболее значимых узлов
4. Извлечение текстового содержимого

Кластеризация новостей



Выбор алгоритма кластеризации

	k-ср.	разд. k-ср.	HAC	DHCA	ICA
Адаптивное кол-во класт.	−	+	+	+	+
Иерархические кластеры	−	+	+	+	−
Онлайновый алгоритм	−	−	−	+	+
Возможность оптимизаций	+	−	−	−	+

Алгоритмы

k-ср — метод K-средних

разд. k-ср — разделяющий метод k-средних

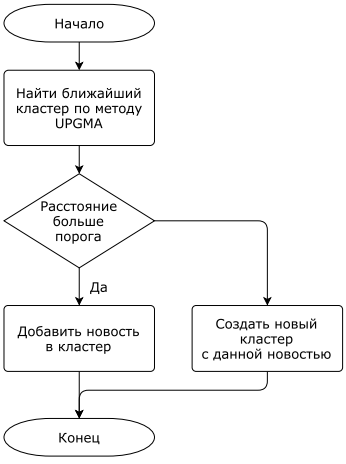
HAC — Hierarchical Agglomerative Clustering

DHC — Dynamic Hierarchical Compact Algorithm

ICA — Incremental Clustering Algorithm

Incremental Clustering Algorithm

Добавление новости:



UPGMA:

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

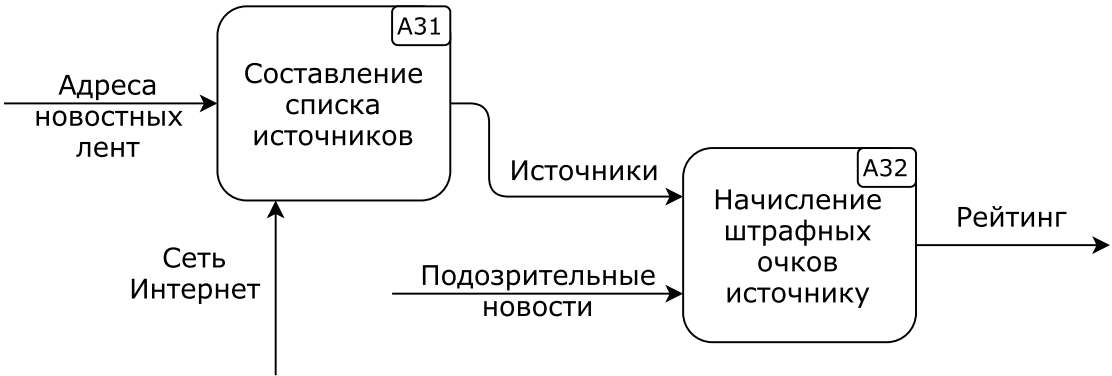
где C_i, C_j — кластеры

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

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

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

Ранжирование источников



Исследование влияния меры сходства на качество кластеризации

Цель эксперимента

Оценить влияние на качество кластеризации меры схожести

Набор данных

2005 новостей и 263 кластера от агрегатора «Яндекс.Новости»

Метрика качества

$$F = \sum_{o \in O} \frac{|o|}{n} \max_{c \in C} F_1(c, o), \text{ где } \quad F_1(c, o) = \frac{2 \cdot P(c, o) \cdot R(c, o)}{P(c, o) + R(c, o)}$$

Исследование влияния меры сходства на качество кластеризации

Мера схожести		F-метрика
Эвклидово расстояние	$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	0.89
Косинусная мера	$sim_c(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$	0.92
Мера Жаккара	$sim_j(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|^2+\|\mathbf{y}\|^2-\mathbf{x}\cdot\mathbf{y}}$	0.91
Эвклидово расстояние со штраф.	$d(\mathbf{x},\mathbf{y}) \cdot (1 + penalty(x,y))$	0.89
Косинусная мера со штрафами	$sim_c(\mathbf{x},\mathbf{y}) \cdot (1 - penalty(x,y))$	0.93
Мера Жаккара со штрафами	$sim_j(\mathbf{x},\mathbf{y}) \cdot (1 - penalty(x,y))$	0.92

Выводы из работы

1. Проанализирована предметная область
2. Разработан метод ранжирования источников
3. Разработано программное обеспечение
4. Исследовано влияние меры схожести на качество кластеризации

Дальнейшее развитие

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