Автоматическое составление обзорных (сводных) рефератов новостных сюжетов

С.Д.Тарасов

Балтийский Государственный Технический Университет им. Д.Ф. Устинова «ВОЕНМЕХ»

Автоматизация реферирования текстовой информации

- SDS (Однодокументное реферирование)
- MDS (Многодокументное реферирование) – как минимум с 2001 года

Конференции: **TREC**, **DUC**, **TSC**.

Практические реализации:
http://news.google.ru/
http://news.yandex.ru/
http://www.newsblaster.com/

Метод Луна

[Luhn, 1958] G. Luhn. The Automatic Creation of Literature Abstracts (context). http://citeseer.ist.psu.edu/context/74679/0

$$V_s = \frac{N_{important}^2}{N_{all}}$$

V_s -значимость предложения;

$N_{important}$ - число значимых слов в предложении (длина последовательности значимых слов);

N_{all} - полное число слов в предложении.

Manifold Ranking Algorithm

- Может быть использован для ранжирования любых информационных примитивов: текстов, предложений, изображений, звуков. В этом случае любой вид информации должен быть представлен в векторном пространстве.

- В задачах ранжирования результатов информационного поиска «отправной точкой» алгоритма является запрос, и ранг предложений определяется как мера их «информационной близости» запросу.

- В задаче автоматического реферирования «отправной точкой» алгоритма можно считать «тему» кластера.

Manifold Ranking Algorithm

- Позволяет описать связную структуру текста
- Для описания связной структуры текста используется математический аппарат векторов и матриц.

Manifold Ranking Algorithm

1. Вычисление ранга каждого предложения (***Информационная значимость***)

2. Применение алгоритма отсечения предложений, наиболее похожих на те, что уже попали в обзорный реферат (***Информационная новизна***)

Алгоритм

1. Задается набор структур:

$$\overline{\mathrm{X}} = \left\{ \overline{x_i} \,|\, 0 \le i \le n \right\} \subset R^m$$

x_0 – предложение, которое формулирует тему кластера

Алгоритм

2. Вводится отображение:

$$f : \overline{X} \to R,$$

которое ставит в соответствие каждому x_i некоторый ранг f_i

$$f = [f_0, f_1, .. f_n]^T$$

Алгоритм

3. Задается вектор:

$$y = [y_0, y_1, .. y_n]^T$$

Согласно алгоритму $y_0=1$, т.к. x_0– тема кластера (в задачах информационного поиска соответствует фразе поискового запроса), и $y_i=0$ для всех остальных предложений ($1<i<n$).

Алгоритм

4. Каждое предложение (объект) представляется в векторном пространстве следующим образом:

$$x_i = \left[tf_0, tf_1, ..tf_n\right]^T,$$

где tf_k - стандартная TF_ISF мера относительной важности терма t_k

Алгоритм

5. Набор предложений представляет собой взвешенный граф с матрицей весов W. Для каждой пары x_i и x_j предложений вычисляется вес их «лексической близости» при помощи стандартной евклидовы меры:

$$W_{i,j} \equiv Sim(\overline{x_i}, \overline{x_j}) \qquad W_{i,i} = 0$$

$$Sim(\overline{x_i}, \overline{x_j}) = \frac{\overline{x_i} \cdot \overline{x_j}}{\|x_i\| \cdot \|x_j\|}$$

«Мама мыла раму»

№	Обозн.	Текст
0	***X0***	***Мама мыла раму***
1	X1	Первого октября мама мыла раму
2	X2	Мама мыла раму
3	X3	Мама мыла раму тряпкой

«Мама мыла раму»

№	МАМА	МЫТЬ	ОКТЯБРЬ	ПЕРВОЕ	РАМА	ТРЯПКА
0	0,33	0,33	0,00	0,00	0,33	0,00
1	0,20	0,20	0,48	0,48	0,20	0,00
2	0,33	0,33	0,00	0,00	0,33	0,00
3	0,25	0,25	0,00	0,00	0,25	0,60

$$x_0 = (0.33;0.33;0.00;0.00;0.33;0.00)$$
$$x_1 = (0.20;0.20;0.48;0.48;0.20;0.00)$$
$$x_2 = (0.33;0.33;0.00;0.00;0.33;0.00)$$
$$x_3 = (0.25;0.25;0.00;0.00;0.25;0.60)$$

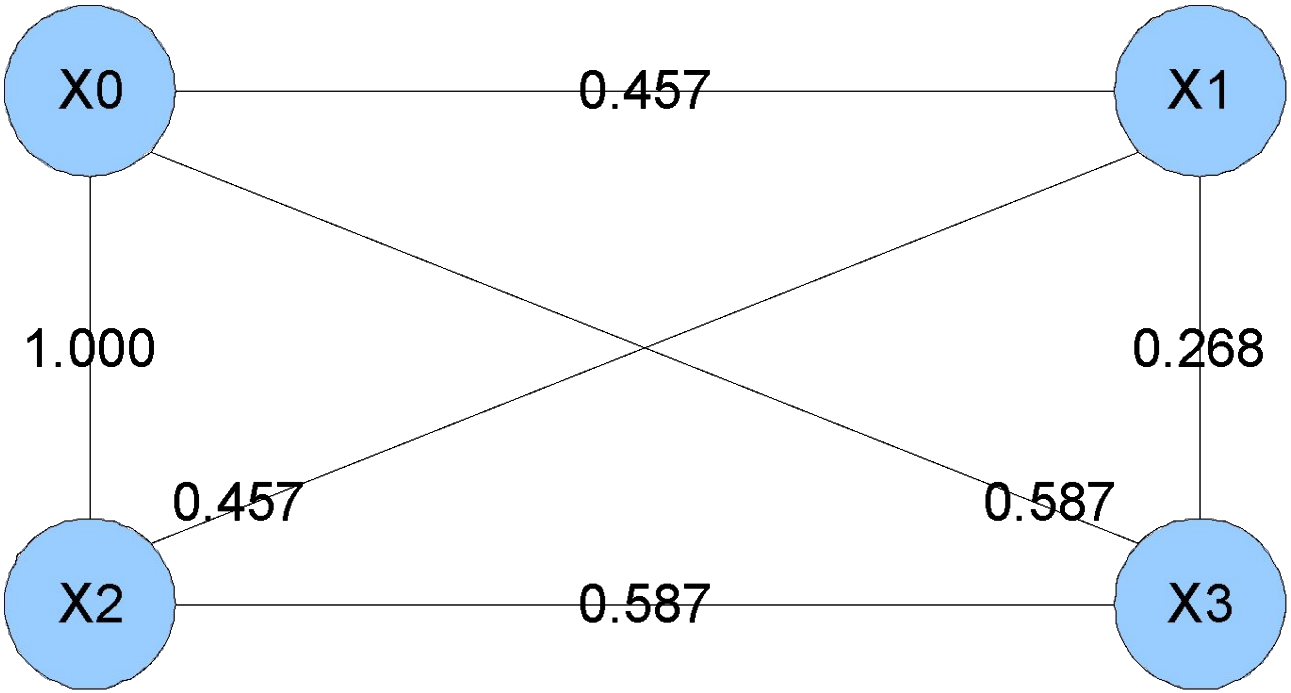
«Мама мыла раму»

Матрица весов в этом случае будет выглядеть:

$$W = \begin{matrix} 0,000 & 0.457 & 1.000 & 0.587 \\ 0,457 & 0,000 & 0,457 & 0,268 \\ 1,000 & 0,457 & 0,000 & 0,587 \\ 0,587 & 0,268 & 0,587 & 0,000 \end{matrix}$$

«Мама мыла раму»

Граф связности текста:



Алгоритм

6. Матрица весов подвергается симметричной нормализации:

$$S = D^{-1/2} \cdot W \cdot D^{-1/2}$$

$$D_{i,i} = \sum_{j=1}^{N} W_{i,j}$$

Алгоритм

6. F вычисляется как результат итеративного процесса: :

$$\overline{f}(t+1) = \alpha \cdot S \cdot \overline{f}(t) + (1-\alpha) \cdot \overline{y}$$

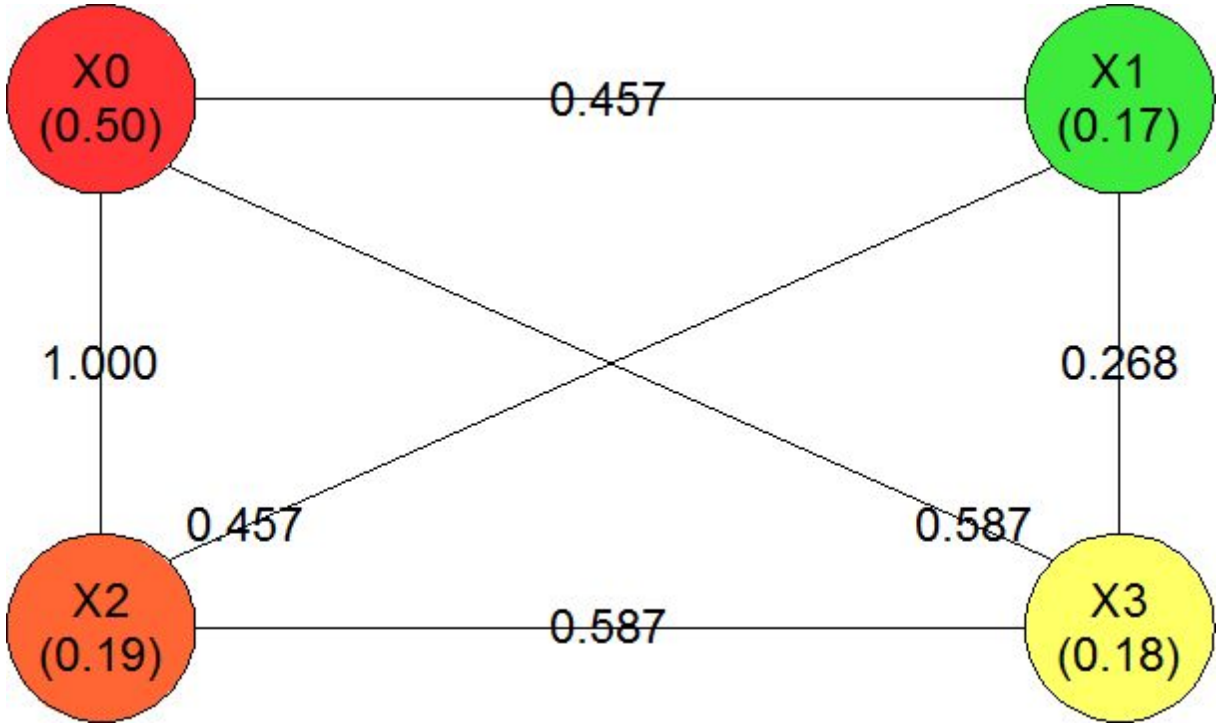
«Мама мыла раму»

Расчет вектора F:

$$f^0 = (0.00;0.00;0.00;0.00)$$
$$f^1 = (0.40;0.00;0.00;0.00)$$
$$f^2 = (0.40;0.09;0.12;0.09)$$
$$f^3 = (0.46;0.13;0.14;0.13)$$
$$f^4 = (0.48;0.16;0.17;0.16)$$
$$f^5 = (0.50;0.17;0.19;0.18)$$

«Мама мыла раму»

Граф связности текста:



Алгоритм

7. Можно также предположить, что связи между предложениями одного документа, а также связи между предложениями различных документов набора должны быть дифференцированы. В этом случае полагается, что :

$$W = W_{inner} + W_{\text{int}\,ra}$$

$$W = \lambda_1 \cdot W_{inner} + \lambda_2 \cdot W_{\text{int}\,ra}$$

Алгоритм усечения сходных предложений

- Необходимо исключить из рассмотрения предложения, повторяющие по своей структуре те, что уже попали в обзорный реферат

- Необходимо выполнить итоговою сортировку предложений в обзорном реферате

Алгоритм усечения сходных предложений

1. Инициализируются два множества A и B. Все предложения помещаются в B. Для каждого предложения B текущий ранг принимается равным f_i.

$$RankScore(x_i) = f_i^*, i = 1,2,..n$$

2. Предложения множества B сортируются в соответствии с их текущим рангом в порядке убывания.

Алгоритм усечения сходных предложений

3. Полагая, что предложение x_i имеет наивысший ранг, оно перемещается из B в A. Ранг оставшихся в B предложений рассчитывается как:

$$RankScore(x_j) = RankScore(x_j) - \omega \cdot \overline{S}_{j,i} \cdot f_i^*$$

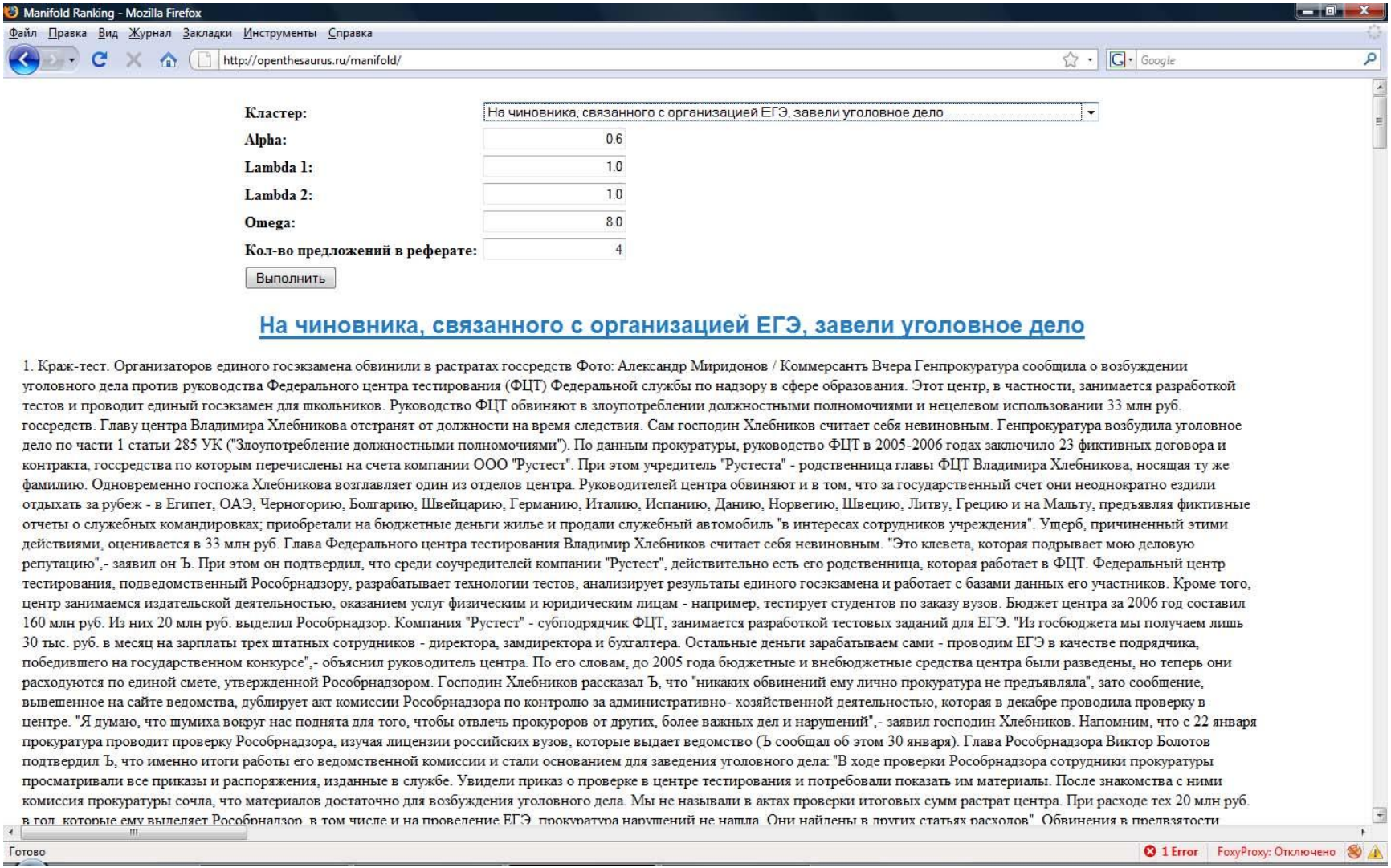
$$\overline{S} = D^{-1} \cdot W$$

$\omega > 0$ - фактор усечения сходных предложений

Реализация

- Web-интерфейс
- PHP
- Расширение ***php_math***
- MTL
- AOT
- Подбор параметров $\alpha, \lambda_1, \lambda_2, \omega$
- http://openthesaurus.ru/manifold/

On-line сервис



On-line сервис

Окончательный расчет

#	1	2	3	4
0	0.75105717775984	0.10581176200654	-0.28581128137615	-0.73590170149818
1	0.2972579363609	-0.1338486002092	-0.39550292545016	
2	0.73701954902901	0.15909507888126	-0.30858958666131	-0.71172033539189
3	0.28999836622042	-0.094496888082202		
4	0.75105717775984	0.10581176200654	-0.28581128137615	-0.73590170149818
5	0.3500190547552			

Сводный реферат

N	Предложение
5	Мама мыла раму тряпкой
3	Сегодня наша мама мыла раму, а папа смотрел телевизор все время.
1	Первого октября мама мыла раму окна
2	Сегодня наша мама мыла раму

0.00 сек.

Исходные данные

В качестве исходных данных для оценки работы алгоритма был взят набор кластеров новостной тематики, любезно предоставленный НИВЦ МГУ.

Пример аннотации

Для кластера «На севере Омской области выпал разноцветный снег» содержащего 8 документов (всего 61 предложение) был получен обзорный реферат из 4 предложений:

«*Представители властей заявили, что если вдруг выяснится, что разноцветный снег в Сибири выпал из-за промышленных выбросов, нарушителей привлекут к уголовной ответственности. Пока специалисты только говорят, что аномальные осадки не опасны для здоровья. Кроме того, необычный снег выпал в Томской и Тюменской областях. Вчера были обнародованы первые лабораторные исследования выпавшего 31 января в Омской области желто-оранжевого снега*».

Оценка

На основе ручных аннотаций, любезно предоставленных НИВЦ МГУ проведена оценка качества системы реферирования при помощи меры ROUGE.

$$ROUGE-N=\frac{\sum_{S\in \mathrm{Re}\, fSum}\ \sum_{n-gram\in S} Count_{match}(n-gramm)}{\sum_{S\in \mathrm{Re}\, fSum}\ \sum_{n-gram\in S} Count(n-gramm)}$$

Результаты оценки

$$\alpha = 0.6, \lambda_1 = 0.3, \lambda_2 = 1, \omega = 15$$

№	Тема кластера	ROUGE-1	ROUGE-2	ROUGE-3
1	В Гальском районе Абхазии неизвестные похитили главу районной избирательной комиссии	0.4286	0.2545	0.2037
2	Китай успешно запустил на орбиту навигационный спутник "Бэйдоу"	0.2581	0.0656	0.0333
3	Секретаршу из Coca-Cola признали виновной в краже секретов компании	0.5600	0.4286	0.3542
4	На севере Омской области выпал разноцветный снег	0.4655	0.3157	0.2500

Сравнение с DUC

	DUC 2003	**DUC 2005**	**Построенная система**
ROUGE-1	0.37332	0.38434	**0.42805**
ROUGE-2	0.07677	0.07317	**0.26610**

Итоги

- Алгоритм реализован в виде “on-line” Web-сервиса. Сводные рефераты могут быть получены «на лету»
- Алгоритм апробирован на русскоязычных новостных кластерах
- Произведена оценка качества работы алгоритма по мере ROUGE. Выполнено сравнение результатов с DUC
- Сформулирован список возможных направлений по улучшению качества аннотирования

Будущая работа

- Улучшить алгоритм распознавания и разрешения анафор
- Добавить в систему синонимию.
- Провести более основательную оценку качества работы системы на основе большего количества ручных рефератов.
- Реализовать Multi-Topic MDS для новостных кластеров
- Исследовать алгоритм для реферирования кластеров другой тематики (например, описания фильмов)