



eXtragon: экспериментальная система для автоматического реферирования веб-документов

Павел Браславский
ИМАШ УрО РАН

Иван Колычев
УГТУ-УПИ

Извлечение предложений (sentence extraction)

- Значимость отдельных слов:
 - $TF*IDF$
 - Графическое выделение (HTML)
 - Присутствие в запросе
- Выделение словосочетаний *
- Положение предложения в тексте
- Слова запроса – нелинейный рост
- Длина предложения *
- Разнообразие реферата *

Дополнительные ресурсы

- HashTrie (SoftLab MIL-TEC Ltd.)
- mystem
- частотный словарь (1 000 000 → 333 605)

Этапы обработки

- Кодировка
- HTML → TXT
- Вычисление базовых (частотных) весов слов
- Вычисление базовых весов словосочетаний
- Вычисление весов ключевых слов:

$$K = 1 + K_B + K_U + K_I + K_T + K_H + K_Q$$

Вес предложения

$$P = L \cdot I \cdot e^{-\left(\frac{SL-OL}{P}\right)^2} \cdot \left(1 + \frac{2q^2}{QL}\right) \cdot \sum_{i=1}^{SL} W_i$$

L – учет положения предложения;

I – понижающий коэффициент для вопросительных предложений;

SL – длина предложения в словах;

QL – длина запроса в словах;

W_i – вес i -го слова в предложении;

q – количество слов запроса в предложении;

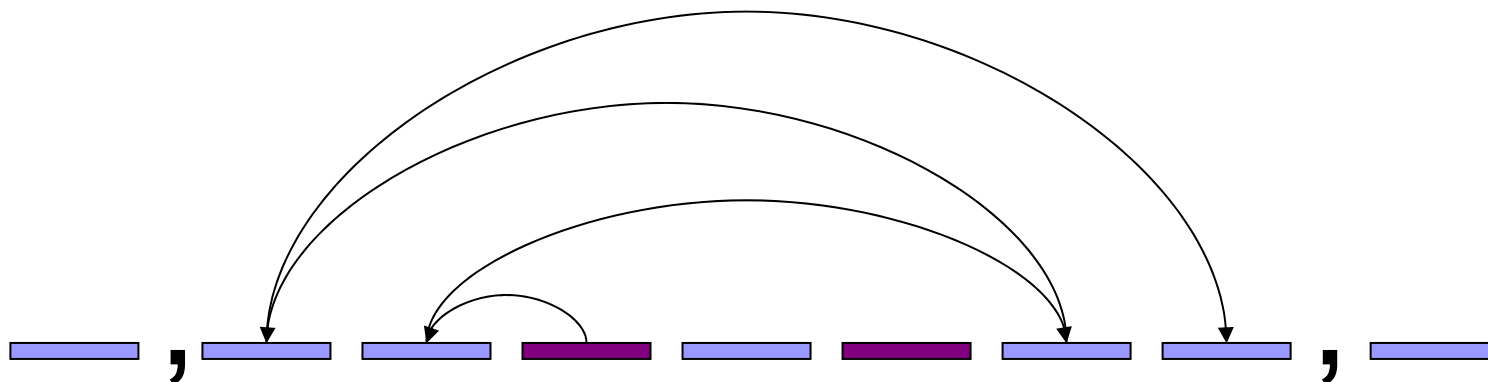
OL – «оптимальная» длина предложения для реферата;

P – коэффициент уменьшения веса предложения при отклонении от «оптимальной» длины.

Формирование реферата

- Сортировка предложений по весу
- Проверка разнообразия реферата
- Восстановление порядка (как в исходном тексте)

Формирование реферата (2)



Производительность

Задания РОМИП:

~21000 документов – 60 часов (2х30ч Р4)

Оптимизированная версия:

~1,2 с – 100 Кб

~15 с – 300 Кб

Оценки РОМИП

	relevant–				vital			
	OR_OR		AND_AND		OR_OR		AND_AND	
	A	E	A	E	A	E	A	E
Web	0,92	0,35	0,79	0,35	0,90	0,63	0,90	0,47
Legal	0,84	0,56	0,78	0,38	0,88	0,70	0,90	0,48
Всего	0,87	0,48	0,78	0,37	0,89	0,67	0,90	0,49

Проблемы

- Для Веб-документов

- ☐ оглавления, списки ссылок
- ☐ не учитывается структура страницы

- Для нормативно-правовых документов:

- ☐ Терминология (например, для Гражданского кодекса, часть 2 самые весомые слова: *статья, комментарий, кодекс; настоящий кодекс*)
- ☐ Особенности стиля и структуры документов (длинные предложения, пункты, статьи)

Преимущества

- Универсальность
- Простота настройки взвешивания слов и предложений
- Составление реферата даже при отсутствии слов запроса в документе
- Веса предложений могут быть вычислены на этапе индексирования и «подправлены» при обработке запроса

Перспективы развития системы

- Качество определения границ предложений
- Учет структуры документа
- Учет оформления страницы
- Исследование роли словосочетаний

- Оптимизация структур данных