Интеллектуальные информационные системы

Тема 4

Выделение ключевых слов. Реферирование

Павел Исаакович Браславский
pb@imach.uran.ru
весенний семестр 2006

План

- Выделение ключевых слов
  - Графические подсказки, структура, морфология
  - Статистика в тексте: закон Ципфа
  - Глобальная статистика (TF*IDF)
- Реферирование
  - Подходы: генерация vs *экстракция*
  - Положение в тексте
  - Сигнальные слова и фразы
  - Связи между предложениями (анафора, симметричное реферирование)

Ключевые слова – зачем?

Ключевые слова (*keywords*) – это «семантическая выжимка», грубый «смысловой портрет» документа.

- Индексирование
- Тематическая классификация
- Интерфейсы (подсветка, *browsing*)

Даже если ключевые слова указаны автором!

Структура и оформление

Заголовок: Методы извлечения ключевых слов

Текст: Важные слова часто выделены в тексте графически – подчеркиванием, *курсивом*, **полужирным шрифтом**, КАПИТАЛИЗАЦИЕЙ, а также другими способами.

HTML → <title>, <h1>…<h3>, <u>, <i>, <b>.

Частота внутри документа

Чем чаще, тем лучше

(после удаления стоп-слов!!!)

+ морфология: именные части речи (существительные и прилагательные)

Закон Ципфа

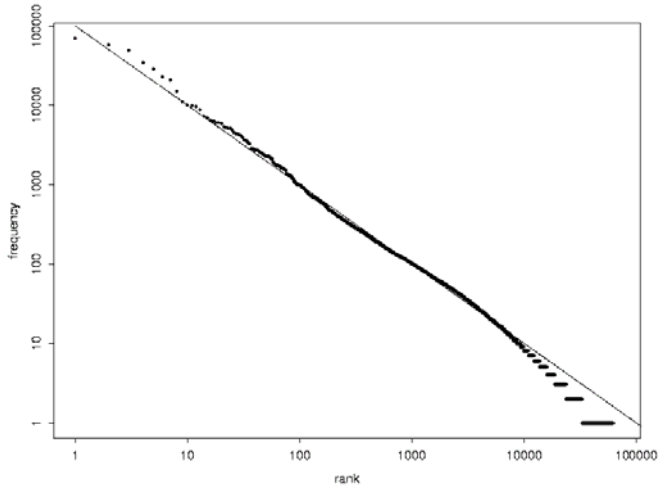
Частота i-го термина в тексте обратно пропорциональна его рангу (порядковому номеру):

$$f_i = \frac{c}{i^\theta} \qquad c = \frac{1}{\sum_{i=1}^{m} \frac{1}{i^\theta}}$$

$\theta = 1{,}5..2.0$

Упрощение: $\theta = 1 \rightarrow c \sim 1/\ln m$

Закон Ципфа – иллюстрация



Manning&Raghavan, 2005

+ глобальная статистика

(уже хорошо знакомый) подход *TF*IDF*

Но: специфика коллекции/документа
Пример: Гражданский кодекс РФ,
самое весомое слово – *статья*

Автоматическое реферирование

Automatic summarization

- Аналитика (сокращение объема)
- Выдача МП: решение о релевантности
- Много документов → один реферат
- Критика/оценка

Два подхода

- Текст → семантическое представление («понимание») → генерация реферата

(abstract generation)

- Выбор значимых фрагментов, обычно предложений

(sentence extraction)

Слова

- Вес предложения пропорционален весу входящих в него слов.
- После некоторого порога возможно растет нелинейно (*keyword burst*)
- Не забывать про графическое выделение, присутствие в запросе и т.д.

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

Предложения

- Начало и конец документа обычно содержат «хорошие» предложения (если текст «хороший»)
- Нормализация по длине предложения (длинные предложения набирают вес просто за счет длины)
- Слишком короткие предложения часто неинформативны
- Сигнальные фразы и слова: *можно сделать вывод*…, *необходимо подчеркнуть*… (ограничение по жанру, сложно поддерживать словарь)
- Вопросительные предложения скорее всего не подходят

Мегаформула ☺

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

Браславский, Колычев 2005

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;

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

Текст

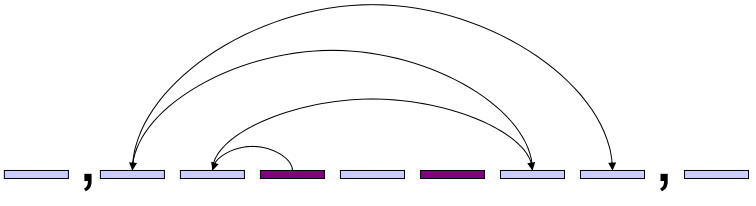
- Симметричное реферирование: взаимосвязи между предложениями по ключевым словам
- Анафоричские ссылки – убирать предложения с «висящими» отсылками или присоединять предыдущие.

Вася съел 46 пельменей. Это его рекорд.

Постобработка реферата

- Восстановление первоначального порядка
- Требование разнообразия реферата
- Сокращение до заданной длины
- Сглаживание

Отсечение «лишнего»



Более изощренные методы

In Mr. Garland's vision, that might mean charging all users a flat media fee, paid through their Internet service providers, which in turn would pay the studios.

(=27)

In Mr. Garland's vision, **that might mean charging all users a flat media fee**, paid through their Internet service providers, which in turn would pay the studios.

(=10)