Информационный поиск

Лидия Михайловна Пивоварова

Системы понимания текста

Введение

- Информационный поиск – поиск в большой коллекции документов, удовлетворяющих потребности пользователя, сформулированной в виде короткого запроса на естественном языке.

- Стремительный рост Интернета и успешное развитие информационно-поисковых систем привели к тому, что современный информационный поиск как дисциплина включает широкий круг вопросов, связанных со сбором, хранением, поиском и представлением самой разнообразной информации; сюда же естественным образом относятся многие задачи автоматической обработки текста.

Содержание

1. Индексирование
2. Модели информационного поиска
3. Оценка информационного поиска
4. Роль автоматической обработки текста в информационном поиске

Индексирование

- Поиск по большим коллекциям не может осуществляться в режиме реального времени.
- Для быстрого поиска коллекция предварительно обрабатывается и по ней строится **индекс**(ы) – набор атрибутов, которые упорядочены в удобном для поиска порядке.
- В случае полнотекстового поиска такими атрибутами являются слова (словосочетания), приведенные к нормальной форме.

Структура индекса

Рис. 1.3. Две части инвертированного индекса. Словарь обычно находится в памяти вместе с указателями на каждый список словопозиций, которые хранятся на диске

Процесс индексирования

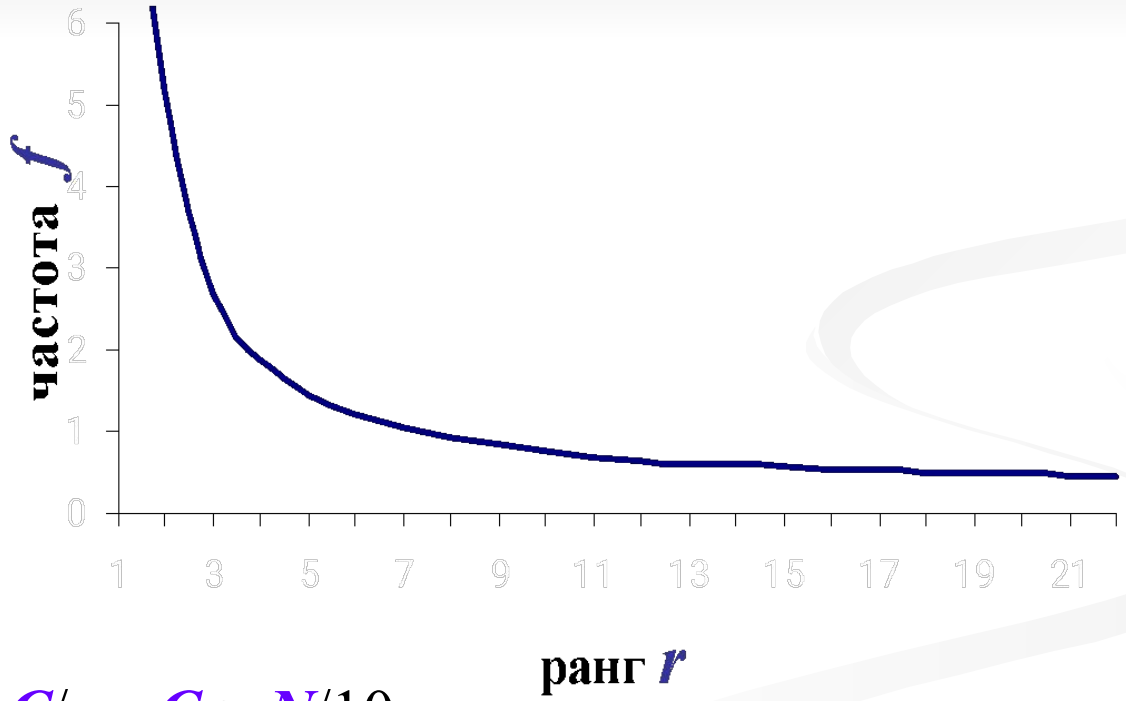
1. Анализ структуры – выделение заголовков, абзацев и т.п.; удаление html-разметки и т.д;
2. Токенизация – разбиение текста на слова, удаление знаков препинания;
3. Удаление стоп-слов - высокочастотных служебных слов (предлогов, союзов и т.п.);
4. Лемматизация – приведение слов к нормальной (например, словарной) форме;
5. Взвешивание

Взвешивание

- В индексе хочется учитывать не только сам факт вхождения слова в документ, но и «вес», т.е. информацию о частоте данного слова в документе.
- Однако саму по себе частоту использовать плохо, поскольку слова распределены в языке неравномерно: некоторые встречаются гораздо чаще других

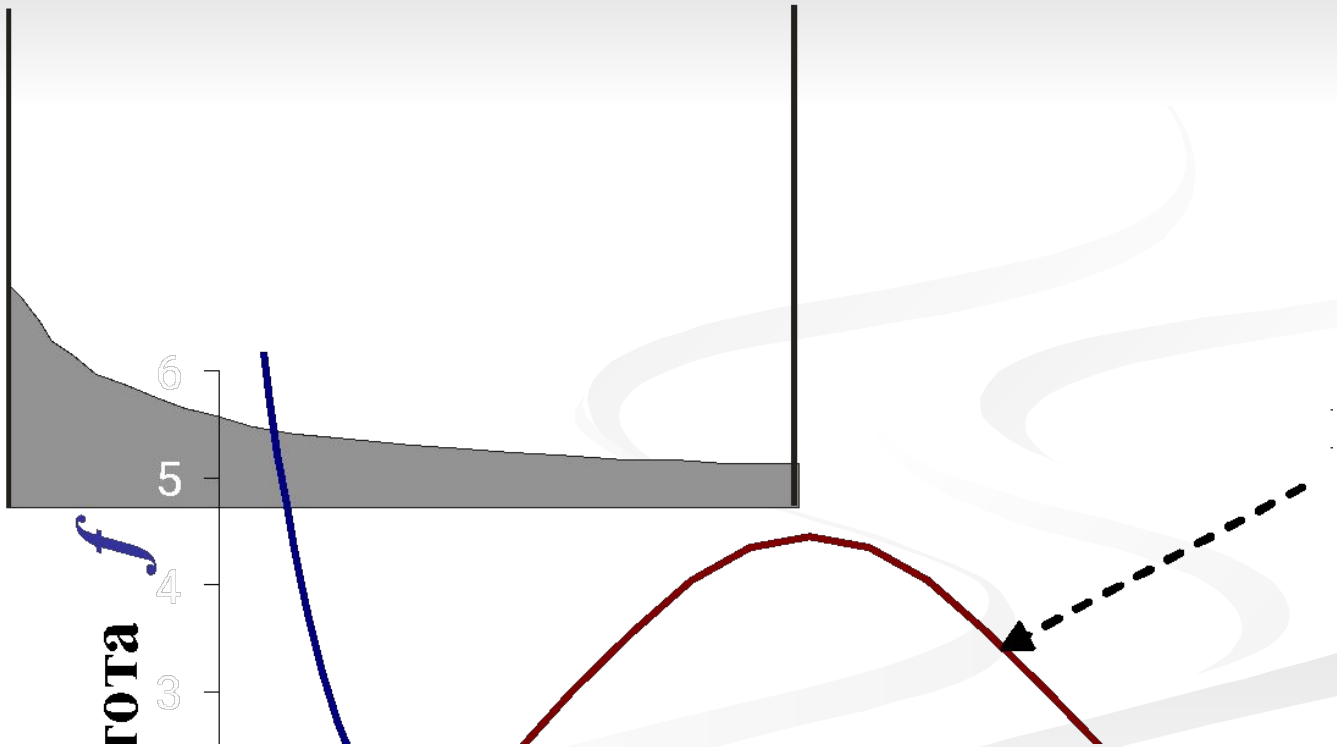
Закон Ципфа (Zipf)

Произведение частоты термина ***f*** на его ранг ***r*** остается примерно постоянной величиной



$f = C/r, \quad C \approx N/10$

Принцип Луна (Luhn)



Самые часто встречающиеся слова – не самые значимые!

Классический метод взвешивания: tf-idf

- tf – относительная частота слова в документе
- idf – обратная документальная частота (чем меньше в коллекции документов, в которые входит это слово, тем idf больше)

$$idf_t = \log \frac{N}{df_t}$$

Вес слова в документе: $\text{tf–idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t.$

В современных поисковых системах используются более сложные варианты взвешивания.

Содержание

1. Индексирование
2. Модели информационного поиска
3. Оценка информационного поиска
4. Роль автоматической обработки текста в информационном поиске

Булева модель

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Рис. 1.1. Матрица инцидентности "термин–документ". Элемент матрицы (t,d) равен 1, если пьеса в столбце d содержит слово из строки t, и 0, если не содержит

- Запрос: булево выражение: Brutus AND Caesar AND NOT Calpurnia
- Ответ: 110100 AND 110111 AND 101111 = 100100
- Плюс: простота; минус: отсутствие ранжирование

Векторная модель

- Коллекция из n документов и m различных терминов представляется в виде матрицы mxn, где каждый документ – вектор в m-мерном пространстве.
- Веса терминов можно считать по разному: частота, бинарная частота (входит – не входит), tf*idf…
- Порядок слов не учитывается (bag of words)
- Матрица очень большая (большое число различных терминов в гетерогенной коллекции).
- В матрице много нулей

Векторная модель

- Близость запроса к документу: косинусная мера близости

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

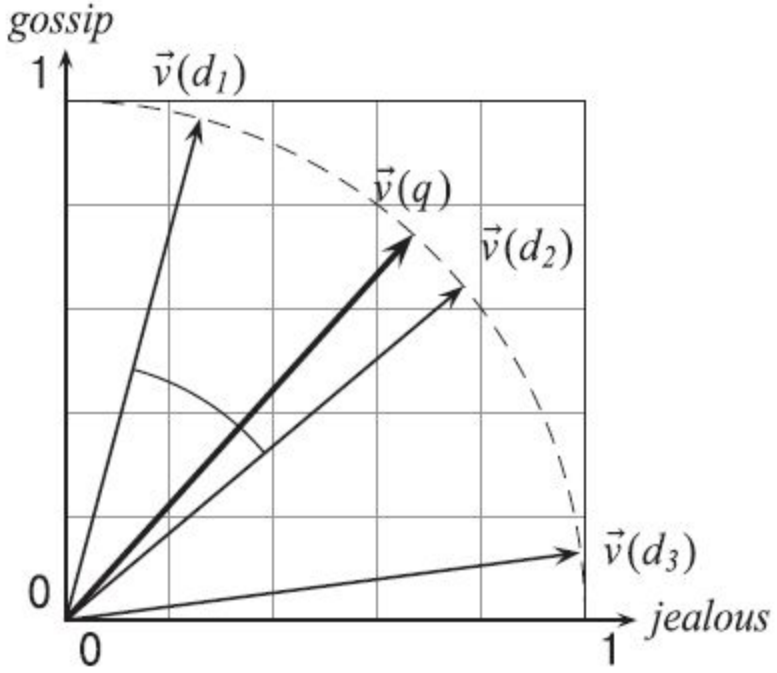


Рис. 6.11. Косинусная мера сходства: $sim(d_1, d_2) = \cos\theta$

Вероятностные модели

- Вероятность вычисляется на основе теоремы Байеса:

$$P(R \mid d) = \frac{P(d \mid R) \cdot P(R)}{P(d)}$$

- ***P(R)*** – вероятность того, что случайно выбранный из коллекции документ ***D*** является релевантным
- ***P(d|R)*** – вероятность случайного выбора документа ***d*** из множества релевантных документов
- ***P(d)*** – вероятность случайного выбора документа ***d*** из коллекции ***D***

Вероятностные модели

- Решающее правило заключается в максимизации следующей функции:

$$S(d) = \frac{\Pr(d \mid R)}{\Pr(d \mid \overline{R})}$$

Содержание

1. Индексирование
2. Модели информационного поиска
3. Оценка информационного поиска
4. Роль автоматической обработки текста в информационном поиске

Оценка информационного поиска

	Релевантные	Нерелевантные
Найденные	tp	fp
Ненайденные	fn	tn

Полнота (recall):
$R = tp / (tp+fn)$

Точность (presicion):
$P = tp / (tp+fp)$

F-мера:

$$F_\alpha = \frac{(1+\alpha)RP}{\alpha P + R}$$

Оценка требует большой коллекции размеченных документов, т.е. огромного труда асессоров.
Большое продвижение дают конференции-соревнования: TREC, РОМИП и т.д.

Аккуратность (accuracy):
$A = (tp + tn) / (tp + tn + fp + fn)$

Содержание

1. Индексирование
2. Модели информационного поиска
3. Оценка информационного поиска
4. Роль автоматической обработки текста в информационном поиске

Уровни анализа языка

- Морфологический анализ
  – признан необходимым для информационного поиска, особенно для флективных языков (например, русского); сюда же относится предсказательная морфология (для незнакомых слов), а также исправление опечаток.

- Синтаксический анализ
  – уже из самого понятия “bag of words” следует, что синтаксис здесь практически не используется; исключения: линейный порядок слов, именные группы, сборка терминологических словосочетаний.

- Семантический анализ
  – в классическом информационном поиске как правило не используется; некоторые элементы лексической семантики применяются при расширении запросов, индексировании документов и составлении каталогов.

Источники

1. J. Savoy, E. Gaussier Information Retrieval // Handbook of natural language processing, Second Edition Editor(s): Nitin Indurkhya; Fred J. Damerau, Goshen, Connecticut, USA – 2010 – pp. 455-484
2. К. Д. Маннинг, П. Рагхаван, Х. Шютце Введение в информационный поиск – Вильямс, 2011
3. А.В. Сычев Информационно-поисковые системы - http://company.yandex.ru/academic/class2006/sychev.xml