Информационный поиск

Алексей Владыкин

СПбГУ ИТМО

1 апреля 2011

- Information Retrieval

- Поиск по запросу
- Ранжирование
- Навигация (классификация, кластеризация)
- Извлечение информации

Модель поиска

- Информационная потребность пользователя
- Поисковый запрос
- Представление запроса
- Представление коллекции документов
- Пертинентность / релевантность

Индексирование

- Токенизация
- Стемминг, лемматизация
- Удаление стоп-слов
- Присвоение весов
- Инвертированный индекс

Булевская модель

- Запрос — булева формула
- Документ — множество слов
- Релевантность дискретна — 0 или 1
- Необходимо внешнее ранжирование

Векторная модель

- Запрос и документ — вектора весов слов
- $\mathrm{TF}(\mathrm{T}, \mathrm{D})$ — количество вхождений слова T в документ D
- $\mathrm{DF}(\mathrm{T})$ — количество документов, содержащих слово T
- $\mathrm{IDF}(\mathrm{T}) = \log(\mathrm{N}/\mathrm{DF}(\mathrm{T}))$ — обратная частота слова T в коллекции
- Релевантность:

$$\mathrm{score}(\mathrm{Q}, \mathrm{D}) = \sum_{\mathrm{T} \in \mathrm{Q}} \mathrm{TF}(\mathrm{T}, \mathrm{D}) \times \mathrm{IDF}(\mathrm{T})$$

- В общем случае:

$$\mathrm{score}(\mathrm{Q}, \mathrm{D}) = \frac{\vec{\mathrm{Q}} \cdot \vec{\mathrm{D}}}{|\vec{\mathrm{Q}}| \cdot |\vec{\mathrm{D}}|}$$

Использование ссылочной структуры

- PageRank — модель случайного блуждания
- HITS — hubs & authorities
- Тематический индекс цитирования Яndex'а

Оценка качества поиска

- Полнота

$$\text{recall} = \frac{|\text{relevant} \cap \text{retrieved}|}{|\text{relevant}|}$$

- Точность

$$\text{precision} = \frac{|\text{relevant} \cap \text{retrieved}|}{|\text{retrieved}|}$$

- F-мера

$$\text{F} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Рекомендуемая литература


Baeza-Yates R., Ribeiro-Neto B.
Modern Information Retrieval.
Addison-Wesley, 1999.


Manning C. D., Raghavan P., Schütze H.
Introduction to Information Retrieval.
Cambridge University Press, 2008.