Основы обработки текстов, лекция 11: Информационный поиск

Ярослав Недумов

17 ноября 2019

Что такое информационный поиск?

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).[1]

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

С чего бы вообще что-то искать?[2]

- Неполнота в модели мира

[2] Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and theory. Journal of documentation.

Как помочь?

- Понять, что же пользователь хочет узнать
  - помочь сформулировать запрос
  - или переформулировать уже внутри системы
- Найти ответ в базе знаний
  - собрать такую базу
  - сохранить в подходящем для поиска виде
- Отобразить ответ в понятном пользователю виде

Несколько дат из истории информационного поиска

- 1945 – Vannevar Bush's "As we may think"
- 1971 – MEDLINE
- 1975 – Salton, G., Wong, A., & Yang, C. S. A vector space model for automatic indexing.
- 1990-ые – появление веба, Yandex, Google
- 1994 – Gey, F. C. Inferring probability of relevance using the method of logistic regression.
- 2006 – Marchionini, G. "Exploratory search: from finding to understanding."

SIGIR 2020

- Yu, S., Liu, J., Yang, J., Xiong, C., Bennett, P., Gao, J. and Liu, Z., 2020. Few-Shot Generative Conversational Query Rewriting.
- Morik, M., Singh, A., Hong, J. and Joachims, T., 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank.
- Zhang, F., Mao, J., Liu, Y., Xie, X., Ma, W., Zhang, M. and Ma, S., 2020, July. Models Versus Satisfaction: Towards a Better Understanding of Evaluation Metrics.

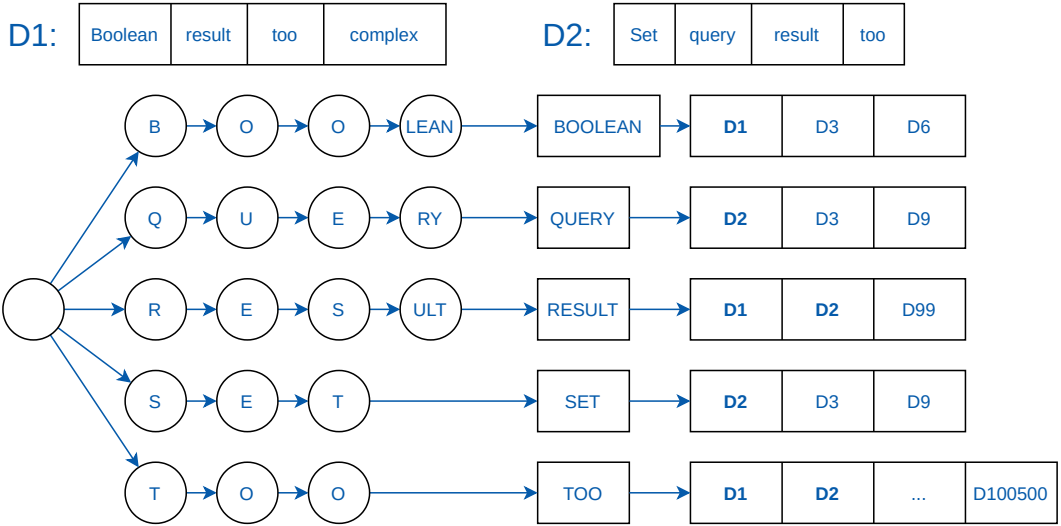
Модели информационного поиска

- Булевы модели
- Векторные модели
- Вероятностные модели
- Модели, основанные на машинном обучении

Простая булева модель

- Запрос: набор независимых терминов, связанных через AND, OR и NOT
- Документ: множество терминов (слов или фраз)
- Релевантность: документ или подходит, или нет

Простейший обратный индекс



Достоинства и недостатки булевой модели

Достоинства и недостатки булевой модели

+ Очень быстрая
+ Можно четко специфицировать какой именно документ требуется вернуть

Достоинства и недостатки булевой модели

+ Очень быстрая
+ Можно четко специфицировать какой именно документ требуется вернуть

- Очень сложно сформулировать запрос так, чтобы получить разумное количество результатов
- Нормальные люди в принципе не умеют в теорию множеств

Простая векторная модель

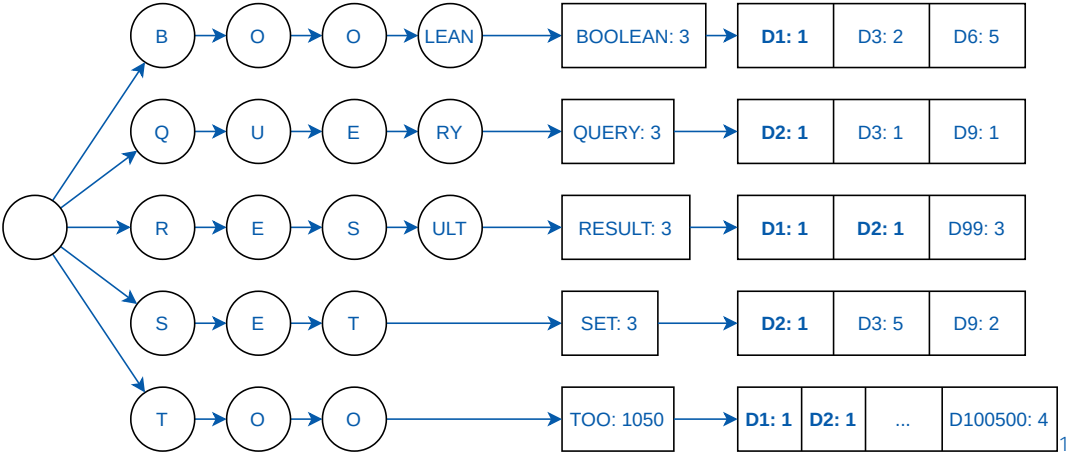
- Запрос: мешок слов
- Документ: мешок слов
- Релевантность: косинус

Обратный индекс с частотами

D1: Boolean | result | too | complex

D2: Set | query | result | too

N: 100500



Достоинства и недостатки векторной модели

Достоинства и недостатки векторной модели

+ Все еще очень быстрая
+ Можно ранжировать результаты

Достоинства и недостатки векторной модели

+ Все еще очень быстрая
+ Можно ранжировать результаты

- Все еще не различает порядка

Непростая вероятностная модель: Okapi BM25[3]

$$RSV_d = \sum_{t \in q} log \left[\left[\frac{(|VR_t| + \frac{1}{2})/(|VNR_t| + \frac{1}{2}))}{(df_t - |VR_T| + \frac{1}{2})/(N - df_t - |VR| + |VR_t| + \frac{1}{2})} \right] \times \frac{(k_1 + 1)tf_{td}}{k_1((1-b) + b(L_d/L_{ave})) + tf_{td}} \times \frac{(k_3 + 1)tf_{tq}}{(k_3 + tf_{tq}} \right] \quad (1)$$

[3] https://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html

Достоинства и недостатки BM25

+ Все еще очень быстрая
+ Обоснована теоретически
+ Учитывает обратную связь от пользователей
+ SotA практически до последнего времени

- Сложно добавить учет дополнительных свойств документов
- Все еще мешок слов (но на самом деле не обязательно)

Источники признаков для ранжирования

- Метаданные документов
- Ссылки между документами
- Данные о пользователе (персонализация)
- Данные о сессии (контекстуализация)

Learning to rank

А давайте используем для ранжирования машинку?

- Pointwise – предсказываем релевантность документа
- Pairwise – выбираем более релевантный документ для пары
- Listwise – сортируем сразу список целиком

Достоинства и недостатки Learning to rank

+ Можно учесть множество параметров
+ Не надо изобретать эвристики

- Сложно собирать тренировочные данные
- Медленно работает, поэтому используется вместе с более простыми моделями

Обратная связь от пользователей

- Явная: разметка результатов
- Неявная слежка за пользователями:
  - перехват кликов
  - оценка времени просмотра страницы
  - и другая аналитика

Проблемы со сбором обратной связи

- Разных запросов слишком много
- Поисковая потребность остается неявной
- Люди не умеют определять релевантность

Оценка качества: nDCG

$$CG_p = \sum_{i=1}^{p} rel_i \qquad (2)$$

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \qquad (3)$$

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p} \qquad (4)$$

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{\log_2(i+1)} \qquad (5)$$

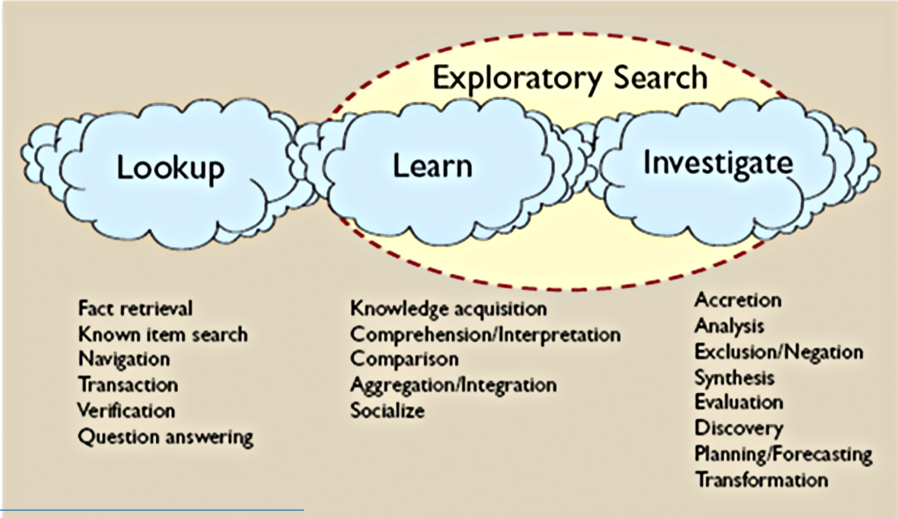
Помощь при составлении запроса

- подсказки
- исправление опечаток
- поиск похожих

Отображение результатов

- 10 blue links + релевантные фрагменты документов ~~+ реклама~~
- Фасеты
- Кластеры
- + Экзотика

Степень осознания потребности может быть разной[4]

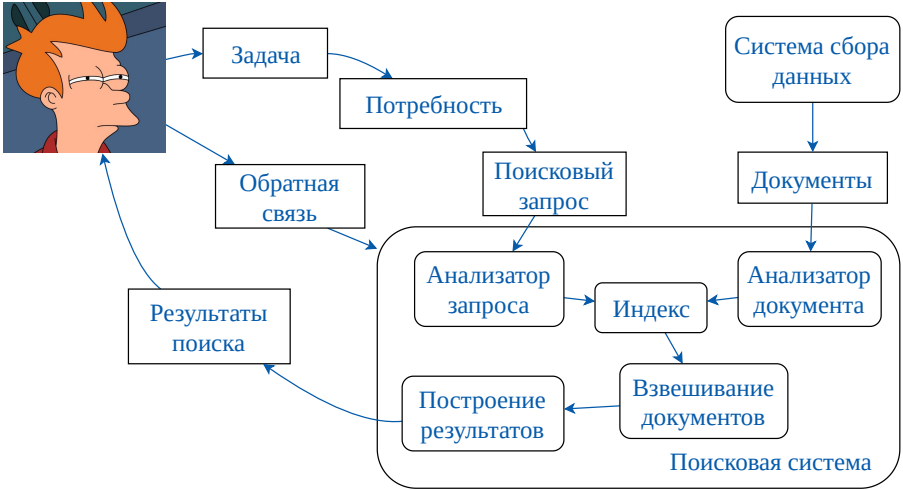


[4] Marchionini, Gary. "Exploratory search: from finding to understanding."Communications of the ACM 49, no. 4 (2006): 41-46.

Практикум

- PubMed
- Bing Images
- Навигатор Кинопоиска
- SemanticScholar
- Яндекс Маркет

TIL


Задача
Потребность
Система сбора данных
Обратная связь
Поисковый запрос
Документы
Анализатор запроса
Индекс
Анализатор документа
Результаты поиска
Построение результатов
Взвешивание документов
Поисковая система