W3

2016年8月11日 21:43

W3L1

- Information retrieval (IR)
  - information need -> search and query -relevant> system retrieves matched doc (Answers)
    - Search is much broader than the web and is used on vastly different scales. the former are the written expression of the latter. Search is one component, but not the only one, of the task of resolving an information need.
  - Distinguishes
    - emphasis on the user
    - The meaning or content of a document is of more interest
      - Doc not always txt
  - difference between data retrieval and information retrieval
    - Data retrieval
      - used to retrieve items based on facts that describe them
        - Data transformed into a representation suitable for manipulation by an algebraic query language prior to storage
          - Queries are represented in an algebraic language
        - information unambiguous
          - Atypical information cannot be represented or queried
    - Info retrieval
      - retrieve items based on their meaning
        - Stored Doc Not well formatted
          - real-world objects created for individual reasons
          - concerned with originally created document, not with a formal document representation (such as a list of keywords)
        - Users may not agree on the value
        - Doc rich and ambiguous
          - no conceivable automatic method for translating them into an algebraic form
        - Structured txt occasionally useful
  - Past: librarianship
  - The recent past
  - Present: search engine

W3L2

- judge the relevance of a document to a query
  - Boolean querying
    - diabetes = 110,risk = 011, diabetes AND risk = 010
    - only appropriate for heavyweight applications such as deep exploration of a collection.
    - No ranking
      - Similarity, relevance probability
  - Evidence of similarity
    - words-in-the-document strategy. some words more significant. term significance.
      - Translation, Query expansion, Relevance feedback may be used
  - Similarity measurement
    - vector-space model
      - each document d can be thought of as a vector
        - Documents with similar terms have points that are “nearby” in the n-space
          - $\langle w_{d,1}, w_{d,2}, \ldots, w_{d,t}, \ldots, w_{d,n} \rangle$
            - weight describing the importance of term t in d
            - TF-IDF
              - Inverse document frequency
              - In-document frequencY
                - Weights of terms in the documents: $w_{d,t}$ (TF)
                - Weights of terms in the query: $w_{q,t}$ (IDF)
            - $f_{d,t}$, the frequency of term t in document d.
            - $f_{q,t}$, the frequency of term t in the query.
            - f_t, the number of documents containing term t.
            - N, the number of documents in the collection.
            - n, the number of indexed terms in the collection.
            - $F_t = \sum_d f_{d,t}$, the number of occurrences of t in the collection.
            - $F = \sum_t F_t$, the number of occurrences in the collection.

            To link back to our heuristics: we wish to find documents d that have

            - Term t with low f_t, that is, are rare;
            - But t has high $f_{d,t}$, that is, is common in the document;
            - And $\sum_{w \in d} f_{d,w}$ is low, that is, the document is short.

In estimating topical similarity

- length unimportant
- angles matters
  - strategy
    - The Cosine measure (approximately, don't do math, aim at find the most relevant)

$$sim(q,d) = \frac{\sum_t w_{d,t} \times w_{q,t}}{|q||d|}$$

or

$$= \frac{\sum_{t \in q} \text{TF-IDF}_{d,t}}{|d|}$$

Many possible choices for TF-IDF models

- TF: $f_{d,t}$
- IDF: $\frac{N}{f_t}$
- TF-IDF: $f_{d,t} \times \frac{N}{f_t}$
- Document length: $\sqrt{\sum_t \text{TF}_{d,t}^2}$
  (or $\sqrt{\sum_t \text{TF-IDF}_{d,t}^2}$)

Or

- TF: $1 + \log_2 f_{d,t}$ (or 0)
- IDF: $\log_2(1 + \frac{N}{f_t})$ (or 0)
- TF-IDF: $(1 + \log_2 f_{d,t}) \times (\log_2(1 + \frac{N}{f_t}))$ (or 0)
- Document length: $\sqrt{\sum_t \text{TF}_{d,t}^2}$
  (or $\sqrt{\sum_t \text{TF-IDF}_{d,t}^2}$)

Or

- Okapi BM25

Language model

- per-term information
  - which document most closely models the distribution of the terms in the query

$$S(q,d) = \prod_{t \in q} \left(\frac{|d|}{|d|+\mu} \cdot \frac{f_{d,t}}{|d|} + \frac{\mu}{|d|+\mu} \cdot \frac{F_t}{F} \right)$$

$$\overset{rank}{=} \sum_{t \in q} \log \left(\frac{f_{d,t}}{|d|+\mu} + \frac{\mu}{|d|+\mu} \cdot \frac{F_t}{F} \right)$$

  - No IDF
  - term’s likelihood in the document
  - smoothed with information from the collection as a whole
- Evaluation Metrics
  - Accuracy X
    - returns many documents from the collection
  - Precision √ (suitably averaged across multiple queries)
  - Recall
    - often useless in an IR context
  - Precision@K √
  - Recall@K

Precision √ (suitably averaged across multiple queries)
Recall
    often useless in an IR context
Precision@K √
Recall@K
    usually not meaningful

too many

Average Precision: $\frac{1}{N}\sum_{k|d(k)\text{is relevant}} P@k$

where N is the total number of relevant documents for the query
(denominator of Recall)

Typically averaged over many queries: MAP (Mean Average Precision)

differences between evaluation in IR and approx search
    more results in IR
        The collection is larger
    IR multiple “correct” (relevant) results; Approx. Search only one
        collection larger and redundant
        User’s need can be met in different ways
        Accuracy isn’t meaningful
    IR results are ranked, Approx. Search typically not
        Boolean querying typically more like Approx. Search evaluation
        Approx. Search could be ranked, but typically many ties