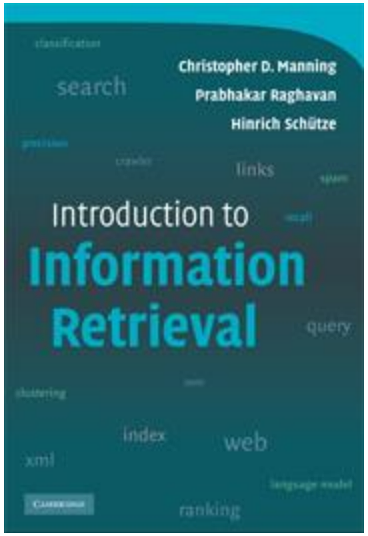
Information Retrieval and Organisation

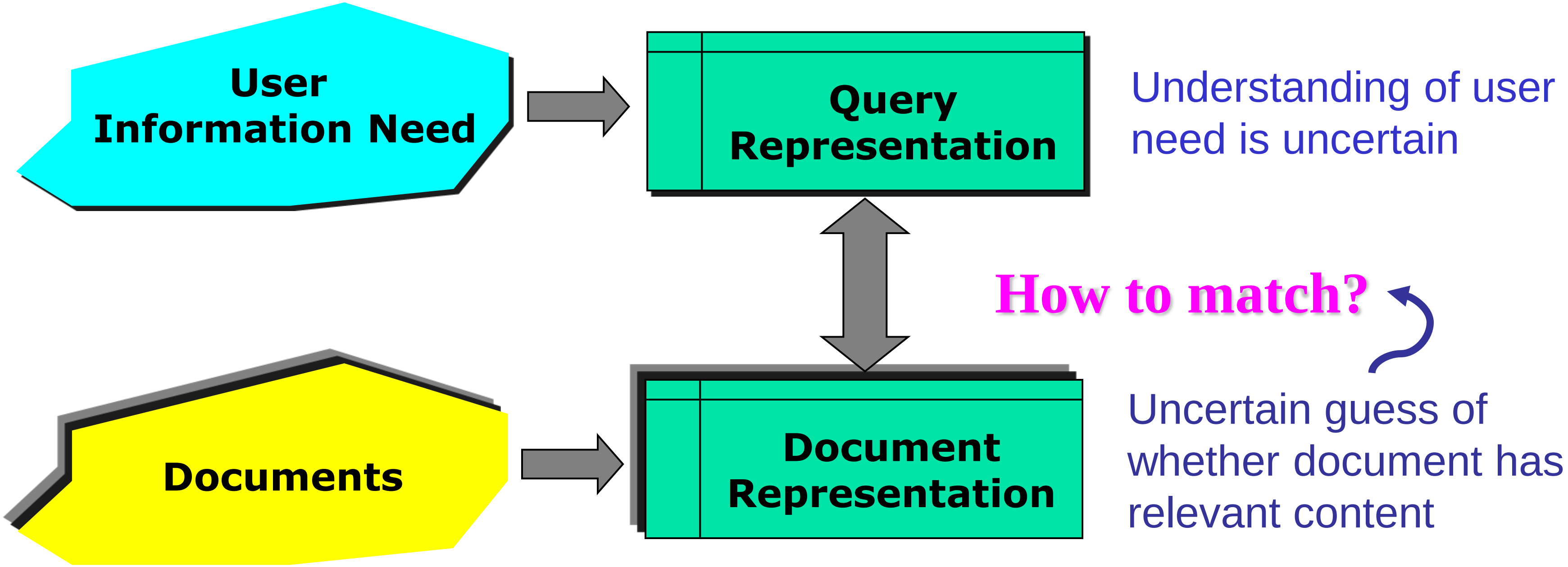


Chapter 11

Probabilistic Information Retrieval

Dell Zhang
Birkbeck, University of London

Why Probabilities in IR?


User Information Need
Query Representation
Understanding of user need is uncertain
How to match?
Documents
Document Representation
Uncertain guess of whether document has relevant content

In IR systems, matching between each document and query is attempted in a semantically imprecise space of index terms.

Probabilities provide a principled foundation for uncertain reasoning.
Can we use probabilities to quantify our uncertainties?

Why Probabilities in IR?

- Problems with vector space model
  - Ad-hoc term weighting schemes
  - Ad-hoc basis vectors
  - Ad-hoc similarity measurement
- We need something more principled!

Probability Ranking Principle

- The document ranking method is the core of an IR system
  - We have a collection of documents. The user issues a query. A list of documents needs to be returned.
  - In what order do we present documents to the user? We want the “best” document to be first, second best second, etc….

Probability Ranking Principle

"If a reference retrieval system's response to each request is a ranking of the documents in the collection **in order of decreasing <u>probability of relevance</u>** to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."

van Rijsbergen (1979:113-114)

$$P(R = 1|d, q)$$

Probability Ranking Principle

- **Theorem**. The PRP is optimal, in the sense that it minimizes the expected loss (also known as the Bayes risk) under 1/0 loss.
  - Provable if all probabilities are known correctly.

Appraisal

- Probabilistic methods are one of the ***oldest*** but also one of the currently ***hottest*** topics in IR.
  - Traditionally: neat ideas, but they've never won on performance.
  - It may be different now. For example, the Okapi BM25 term weighting formulas have been very successful, especially in TREC evaluations.

Okapi BM25

Retrieval Status Value

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d / L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

IDF(t)

The document length of d

The average document length for the collection

The parameters k_1, b should ideally be tuned on a validation set.
The good values in practice are $1.2 \le k_1 \le 2$; $b = 0.75$.

Well-Known UK Researchers

Karen Sparck Jones

Stephen Robertson

Keith van Rijsbergen