Information Retrieval

**Describe the components of IR models.**

Information retrieval model: (D,Q,F,R)

D is a set of document representations (Document)

Q is a set of queries (Query)

F is a framework for modeling document representations, queries and their relationships (Framework)

R associates a real number to document-query-pairs (Ranking)

**Explain and exemplify the basic principles of boolean/vector/probabilistic model for IR (different components).**

Boolean Model: queries: Boolean (and,or,not),queries are translated to DNF

Vector Model: introduce weights in document vectors, which represent importance of the terms for describing the document content. Wij = fij \* idfi

Sim(d,q) = d\*q/|d|\*|q|

Probabilistic Model:

Weights are binary (Wij =0, or Wij = 1)

The system use the users return back information to refine the description of the ideal answer set.

**Describe the advantages and drawbacks of the boolean/vector/probabilistic model.**

Boolean Model:

Advantage: based on set theory and Boolean algebra, simple

Drawbacks: binary decisions, no notion of partial match

Vector Model:

Advantage: term weighting improves retrieval performance, partial matching, ranking according to similarity.

Drawbacks: assume mutually independent terms

Probabilistic Model:

Advantage: ranking of documents with respect to probability of being relevant

Drawbacks: initial guess about relevance, all weight are binary, mutually independence assumption

**Explain TF-IDF.**

**TF:**

term frequency freqi j: how often does term ki occur in document dj?

normalized term frequency: fi,j = freqi,j / maxl freql,j

**IDF**

N = total number of documents

ni = number of documents in which ki occurs inverse document frequency idfi: log (N / ni)

**TF-IDF**

wi,q = (0.5 + 0.5 fi,q) x idfi