3.2 Statistical Indexing

- *Statistical indexing*
    - Uses frequency of occurrence of events to calculate a number that is used to indicate the potential relevance of an item.
    - Probabilistic systems
        - calculate a probability value (Probabilistic Weighting)
    - Bayesian Model and Vector (Weighting) approaches
        - Calculate a relative relevance value (e.g., confidence level).

3.2.1 Probabilistic Weighting

- The probabilistic approach is based upon direct application of the theory of probability to information retrieval systems.
- This is summarized by the Probability Ranking Principle (PRP) and its possible effect (Plausible Corollary):
- HYPOTHESIS: If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data is available for this purpose, then the overall effectiveness of the system to its users is the best obtainable on the basis of that data.
- PLAUSIBLE COROLLARY: The most promising source of techniques for estimating the probabilities of usefulness for output ranking in IR is standard probability theory and statistics.
- *Issues*
    - Probabilities are usually based upon a binary condition; an item is relevant or not.
    - In information systems the relevance of an item is a continuous function from non-relevant to absolutely useful.
    - The output ordering by rank of items based upon probabilities, even if accurately calculated, may not be as optimal.
    - The domains in which probabilistic ranking are suboptimal are so narrowly focused as to make this a minor issue.
- *advantage of the probabilistic approach*
    - it can accurately identify its weak assumptions and work to strengthen them.
    - There are many different areas in which the probabilistic approach may be applied.

- Logistic Regression method
  - starts by defining a "Model 0" system which exists before specific probabilistic models are applied.
  - In a retrieval system there exist query terms and document terms which have a set of attributes from the query (e.g., counts of term frequency in the query) from the document (e.g., counts of term frequency in the document) and from the database (e.g., total number of documents in the database divided by the number of documents indexed by the term).

- Logistic Reference model
  - uses a random sample of query-document-term for which binary relevance result have been made from training sample.
  - Log O is the logarithm of the odds (logodds) of relevance for term tk which is present in document Dj and query Qi.
  - The logarithm that the ith Query is relevant to the jth Document is the sum of the log odds for all terms.

$$\log(O(R \mid Q_i, D_j)) = \sum_{k=1}^{q} [\log(O(R \mid Q_i, D_j, t_k)) - \log(O(R))]$$

  - where O(R) is the odds that a document chosen at random from the database is relevant to query Qi.
  - The coefficients are derived using logistic regression which fits an equation to predict a dichotomous independent variable as a function of independent variables that show statistical variation.
  - The inverse logistic transformation is applied to obtain the probability of relevance of a document to a query:

$$P(R \mid Q_i, D_j) = 1 \setminus (1 + e^{-\log(O(R \mid Q_i, D_j))})$$

  - The coefficients of the equation for logodds is derived for a particular database using a random sample of query-document-term-relevance quadruples and used to predict odds of relevance for other query-document pairs.
  - Additional attributes of relative frequency in the query (QRF), relative frequency in the document (DRF) and relative frequency of the term in all the documents (RFAD) were included, producing the following logodds formula:

$$Z_j = \log(O(R \mid t_j)) = c_0 + c_1\log(QAF) + c_2\log(QRF) + c_3\log(DAF) + c_4\log(DRF)$$

$$+ c_5\log(IDF) + c_6\log(RFAD)$$

  - where QAF, DAF, and IDF were previously defined.
  - QRF = QAF/ (total number of terms in the query),
  - DRF = DAF/(total number of words in the document) and
  - RFAD = (total number of term occurrences in the database)/ (total number of all words in the database).

- Logistic Inference method

- applied to the test database along with the Cornell SMART vector system which uses traditional term frequency, inverse document frequency and cosine relevance weighting formulas.
- The logistic inference method outperformed the vector method.
- The index that supports the calculations for the logistic reference model contains the O(R) constant value along with the coefficients.
- Additionally, it needs to maintain the data to support DAF, DRF, IDF and RFAD.
- The values for QAF and QRF are derived from the query.

term frequency, inverse document frequency :

Numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

Given a corpus D, a term t_i and a document $d_j \in D$, we denote the number of occurrences of t_i in d_j by tf_{ij}. This is referred as the term frequency.

The inverse document frequency for a term t_i is defined as

idfi=log|D||d:ti∈d|

where $|D|$ is the number of documents in our corpus, and $|\{d : t_i \in d\}|$ is the number of documents in which the term appears. If the term t_i appears in every document of the corpus, idf_i is equal to 0. The fewer documents the term t_i appears in, the higher the idf_i value.