2.4 Automatic Indexing

- Advantages of human indexing
  - ability to determine concept abstraction
  - ability to judge the value of a concept.
- Disadvantages of human indexing
  - Cost
  - processing time
  - consistency

- *Automatic indexing*
  - Capability for the system to automatically determine the index terms to be assigned to an item.
  - More complex processing is required when emulate a human indexer and determine a limited number of index terms.
  - No additional indexing costs versus the salaries and benefits regularly paid to human indexers.
  - Requires only a few seconds or less of computer time based upon the size of the processor and the complexity of the algorithms to generate the index.
  - Consistency in the index term selection process as indexing is performed automatically by an algorithm.
  - Indexes from automated indexing fall into two classes:
    - unweighted.
    - weighted
- *Unweighted indexing system*
  - index term in a document and its word location(s) are kept in the searchable data structure.
  - Queries against unweighted systems are based upon Boolean logic and the items in the resultant Hit file are considered equal in value.
  - The last item presented in the file is the first item to be relevant to the user's information need.

- *Weighted indexing system*

- weight of the index term is based upon a function associated with the frequency of occurrence of the term in the item.
- values for the index terms are normalized between zero and one.
- The higher the weight, the more the term represents a concept discussed in the item.
- query process uses the weights along with any weights assigned to terms in the query to determine a rank value.

- 2.4.1 Indexing by Term
- 2.4.2 Indexing by Concept
- 2.4.3 Multimedia Indexing

Indexing by Term

- There are two major techniques for creation of the index terms:
  - statistical
  - natural language
- Statistical techniques
  - based upon vector models and probabilistic models with a special case being Bayesian models.
  - calculation of weights in those models use statistical information such as the frequency of occurrence of words and their distributions in the searchable database.

- Natural language techniques
  - perform more complex parsing to define the final set of index concepts.
  - weighted systems are discussed as vectorized information systems.

- Vector model (Example. SMART system)
  - The system emphasizes weights as a foundation for information detection and stores these weights in a vector form.
  - Each vector represents a document
  - each position in a vector represents a different unique word (processing token) in the database.

- The value assigned to each position is the weight of that term in the document.
- A value of zero indicates that the word was not in the document.
- Queries can be translated into the vector form.
- Search is accomplished by calculating the distance between the query vector and the document vectors.

- Probabilistic model.
  - Bayesian approach is the most successful model in this area.
  - It is based upon the theories of evidential reasoning (drawing conclusions from evidence).
  - The Bayesian approach could be applied as part of index term weighting by calculating the relationship between an item and a specific query.
  - A Bayesian network is a directed acyclic graph
  - Each node represents a random variable.
  - Arcs between the nodes represent a probabilistic dependence between the node and its parents.

  - Figure shows the basic weighting approach for index terms or associations between query terms and index terms.

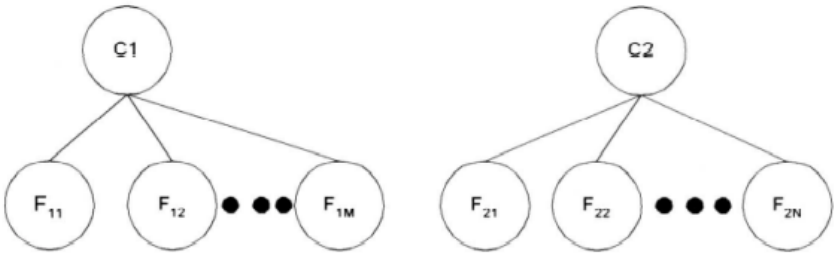


Figure 3.3 Two-level Bayesian network

- The nodes C1 and C2 represent "the item contains concept Ci"
- F nodes represent "the item has feature Fij (e.g., words)

- The network interpreted as C representing concepts in a query
- F representing concepts in an item.
- The goal is to calculate the probability of Ci given Fij.

- To perform that calculation two sets of probabilities are needed:

- The prior probability P(Ci) that an item is relevant to concept C.
- The conditional probability P(Fij/Ci) that the features Fij where j=1,m are present in an item given that the item contains topic Ci.

- The automatic indexing task is to calculate the posterior probability $P(C_i/F_{i1}, \dots, F_{im})$ the probability that the item contains concept Ci given the presence of features Fij

- The Bayes inference formula that is used is:

$$P(C_i/F_{i1}, \dots, F_{im}) = P(C_i)\, P(F_{i1}, \dots, F_{im}/C_i) \backslash P(F_{i1}, \dots, F_{im}).$$

- If the goal is to provide ranking as the result of a search by the posteriors, the Bayes rule can be simplified to a linear decision rule:

$$g(C_i/F_{i1, \dots,} F_{im}) = \Sigma_k\, I(F_{ik})\, w(F_{ik}, C_i)$$

- where $I(F_{ik})$ is an indicator variable that equals 1 only if F_{ik} is present in the item(equals zero otherwise) and w is a coefficient corresponding to a specific feature/concept pair.
- function g is the sum of the weights of the features
- w is weight corresponding to each feature (index term)
- w produces a ranking in decreasing order that is equivalent to the order produced by the posterior probabilities.

- DR-LINK (Document Retrieval through Linguistic Knowledge) system
  - define indexes to items via natural language processing.
  - processes items at the morphological, lexical, semantic, syntactic, and discourse levels.
  - Each level uses information from the previous level to perform its additional analysis.
  - The discourse level is abstracting information beyond the sentence level and can determine abstract concepts.
  - This allows the indexing to include specific term as well as abstract concepts such as time.

Indexing by Concept

- Indexing by term treats each occurrence as a different index.

- Then uses thesauri or other query expansion techniques to expand a query to find the different ways the same thing has been represented.

- *basis for concept indexing*
  - There are many ways to express the same idea and increased retrieval performance comes from using a single representation.
- Concept indexing determines a related set of concepts based upon a test set of terms
- uses those concepts for indexing all items
- Latent Semantic Indexing
  - Indexing the latent semantic information in items.

- *MatchPlus system.*
  - Example for concept indexing.
  - neural networks facilitates machine learning of concept/word relationships.
  - goal is to determine word relationships (e.g., synonyms) and the strength of these relationships and use that information in generating context vectors, from the corpus of items.

  - Two neural networks are used.
    - One neural network learning algorithm generates stem context vectors that are sensitive to similarity of use.
    - another one performs query modification based upon user feedback.
  - *context vectors*
    - Word stems, items and queries that are represented by high dimensional (atleast 300 dimensions) vectors.
    - Each dimension in a vector could be viewed as an abstract concept class.

Multimedia Indexing

- The automated indexing takes place in multiple passes of the information versus just a direct conversion to the indexing structure.

- The first pass in most cases is a conversion from the analog input mode into a digital structure.
- Then algorithms are applied to the digital structure to extract the unit of processing of the different modalities that will be used to represent the item.
- In an abstract sense this could be considered the location of a processing token in the modality.
- This unit will then undergo the final processing that will extract the searchable features that represent the unit.

- Indexing video or images can be accomplished
  - at the raw data level (e.g., the aggregation of raw pixels),
  - the feature level distinguishing primitive attributes such as color and luminance
  - at the semantic level where meaningful objects are recognized (e.g., an airplane in the image/video frame).

- An example is processing of video.
- The system (e.g., Virage) will periodically collect a frame of video input for processing.
- It might compare that frame to the last frame captured to determine the differences between the frames.
- If the difference is below a threshold it will discard the frame.
- For a frame requiring processing, it will define a vector that represents the different features associated with that frame.
- Each dimension of the vector represents a different feature level aspect of the frame.
- The vector then becomes the unit of processing in the search system.
- This is similar to processing an image.
- Semantic level indexing requires pattern recognition of objects within the images.

- analog audio input

  - system will convert the audio to digital format and determine the phonemes associated with the utterances.

- The phonemes will be used as input to a Hidden Markov Search model that will determine with a confidence level the words that were spoken.
- A single phoneme can be divided into four states for the Markov model.
- It is the textual words associated with the audio that becomes the searchable structure.

- In addition to storing the extracted index searchable data, a multimedia item also needs to store some mechanism to correlate the different modalities during search.
- There are two main mechanisms that are used
  - Positional
  - Temporal
- Positional
  - is used when the modalities are scattered in a linear sequential composition.
  - For example a document that has images or audio inserted, can be considered a linear structure and the only relationship between the modalities will be the just a position of each modality.
- Temporal
  - based upon time because the modalities are executing concurrently.
  - The typical video source off television is inherently a multimedia source.
  - It contains video, audio, and potentially closed captioning.

- Synchronized Multimedia Integration Language (SMIL).
  - creation of multimedia presentations are becoming more common using the Synchronized Multimedia Integration Language (SMIL).
  - It is a mark-up language designed to support multimedia presentations that integrate text (e.g.,from slides or free running text) with audio, images and video.
  - time is the mechanism that is used to synchronize the different modalities.
  - indexing must include a time-offset parameter versus a physical displacement.