3.2.3 Bayesian Model

- Bayesian approach is based upon conditional probabilities (e.g., Probability of Event 1 given Event 2 occurred).
- This concept can be applied to the search function as well as to creating the index to the database.
- The objective of information systems is to return relevant items.
- Thus the Bayesian formula, is P(REL/DOCi, Queryj) (the probability of relevance (REL) to a search statement given a particular document and query).
- In addition to search, Bayesian formulas can be used in determining the weights associated with a particular processing token in an item.
- The objective of creating the index to an item is to represent the semantic information in the item.
- A Bayesian network can be used to determine the final set of processing tokens (called topics) and their weights.

- Figure 5.6 shows a simple view of the process where Ti represents the relevance of topic "i" in a particular item and Pj represents a statistic associated with the event of processing token "j" being present in the item.
- The "m" topics would be stored as the final index to the item.
- The statistics associated with the processing token are typically frequency of occurrence

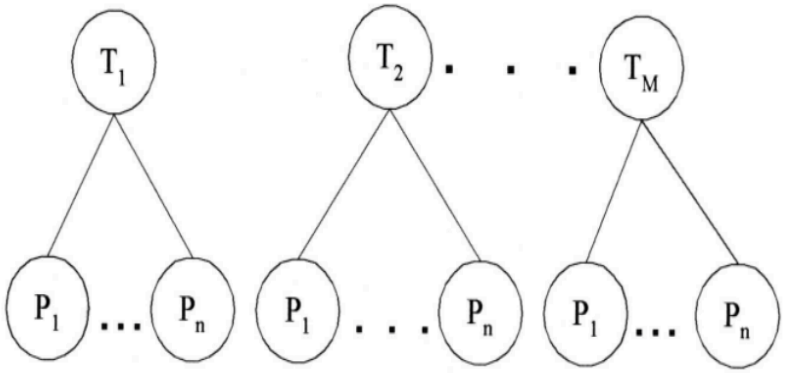

Figure 5.6 Bayesian Term Weighting

- Assumption of Binary Independence:
  - The topics and the processing token statistics are independent of each other. The existence of one topic is not related to the existence of the other topics. The existence of one processing token is not related to the existence of other processing tokens.

- Some topics are related to other topics and some processing tokens related to other processing tokens.

- For example, the topics of “Politics” and “Economics” are in some instances related to each other and in many other instances totally unrelated.

- There are two approaches to handling this problem.

- The first approach

    - assume that there are dependencies, but the errors introduced by mutual independence do not affect the determination of relevance of an item nor its relative rank associated with other retrieved items.

- A second approach
    - extend the network to additional layers to handle interdependencies.
    - additional layer of Independent Topics (ITs) can be placed above the Topic layer.
    - A layer of Independent Processing Tokens (IPs) can be placed above the processing token layer.

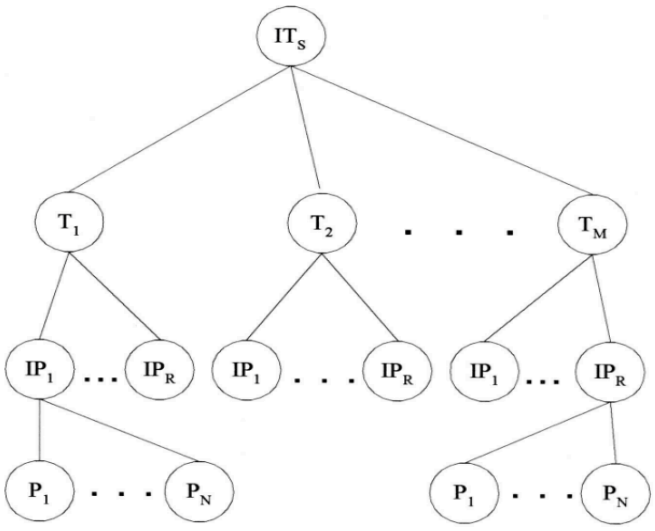

Figure 5.7 Extended Bayesian Network