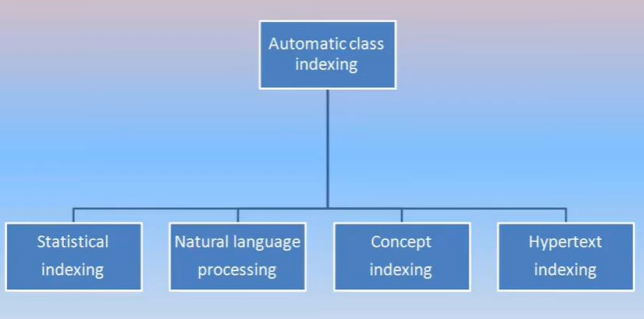
**Automatic Indexing**

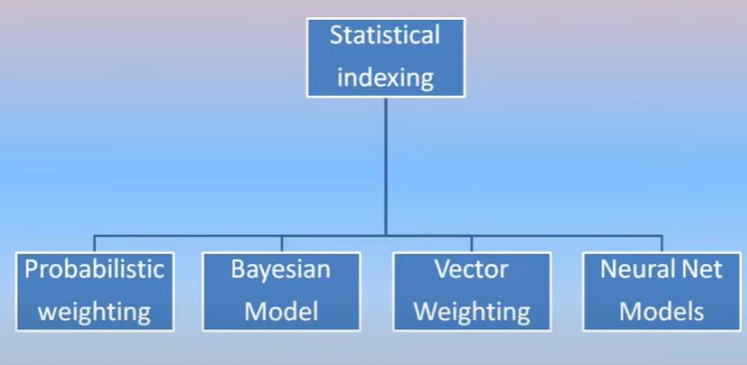
* Automatic indexing is the capability for the system to automatically determine the index terms to be assigned to an item.
* The transformation from received item to searchable data structure is called indexing.
* Indexing was originally called as cataloguing. It was oldest technique to identify the contents of items
* c = a + b item Semantics: variable1 assignment operator variable2 arithmetic operator additional operator variable3
* The index can be based on the
  + full text of the item
  + automatic or manual generation of a subset of terms
  + natural language representation of the item
  + or abstraction to concepts in the item.
* **Advantages of Automatic Indexing**
  + Reduced cost
  + Reduced processing time
  + Improved consistency

Classes of Automatic Indexing:



1. **Statistical Indexing:**

* Statistical strategies cover the broadest range of indexing techniques and are the most prevalent in commercial systems.
* The basis for a statistical approach is use of frequency of occurrence of events.



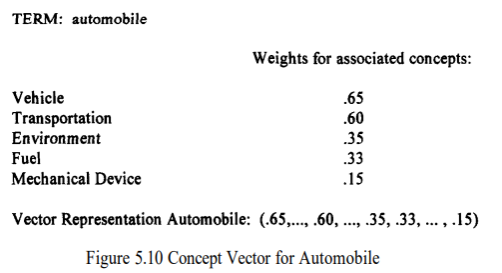
* **Probabilistic indexing** stores the information that are used in calculating a probability that a particular item a particular query.
* **Bayesian and vector approaches** store information used in generating a relative confidence level of an item’s relevance to a query.
* **Neural networks** are dynamic learning structures.

1. **Natural Language**

* Natural Language approaches perform the similar processing token identification as in statistical techniques, but then additionally perform varying levels of natural language parsing of the item.
* Natural language systems attempt to introduce a higher level of abstraction indexing on top of the statistical processes.
* This improves the precision of searches, reducing the number of false hits a user reviews.
* The semantic information is extracted.
* The simplest output of this process results in generation of phrases that become indexes to an item.
* Natural Language processing
  + Index Phrase Generation
  + Natural Language processing

1. **Concept Indexing**

* Concept indexing uses the words within an item to correlate to concepts discussed in the item.
* Concept indexing is a statistical technique whose goal is to determine a canonical representation of the concepts.
* Recall is improved.
* It can be used with concept classes using neural networks.
* An example of applying a concept approach is the convectis system.
* The convectis system uses neural network algorithms and terms in a similar context of other terms.
* A term can have different weights associated with different concepts as described.

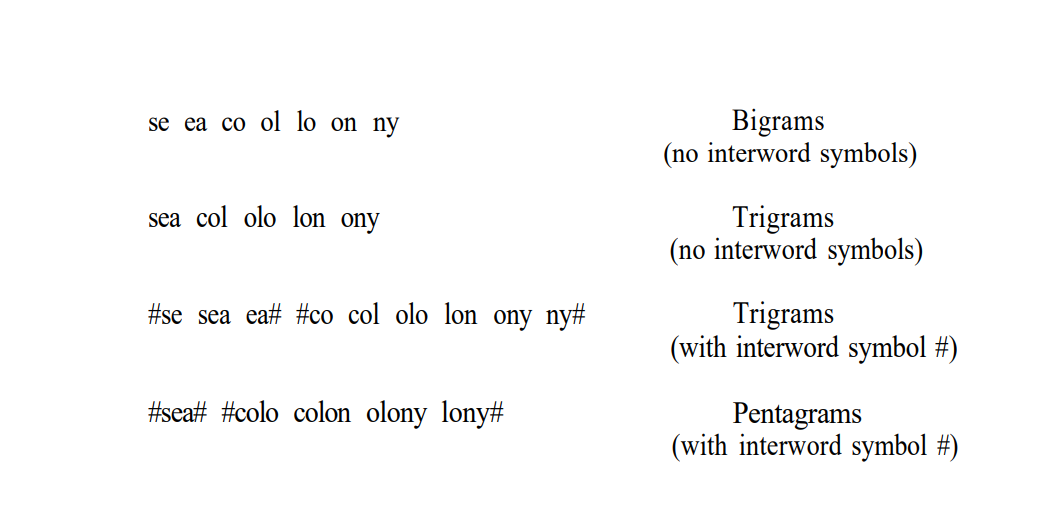


1. **Hypertext Linkage**

* A new class of information representation, described as the hypertext data structure, is evolving on the Internet.
* Hypertext data structures are generated manually.
* Hypertext linkage are creating an additional information retrieval dimension.
* Traditional items can be viewed as two dimensional constructs.
* Hypertext linkages are placed in a separate class because an algorithm to search items that include linkages has to address dependencies between items.

**N-gram Data structure**

* N-Grams can be viewed as a special technique for conflation (stemming) and as a unique data structure in information systems.
* N-Grams are a fixed length consecutive series of “n” characters.
* The symbol # is used to represent the interword symbol which is anyone of a set of symbols (e.g., blank, period, semicolon, colon, etc.).
* Examples of bigrams, trigrams and pentagrams are given in Figure 4.7 for the word phrase “sea colony.”



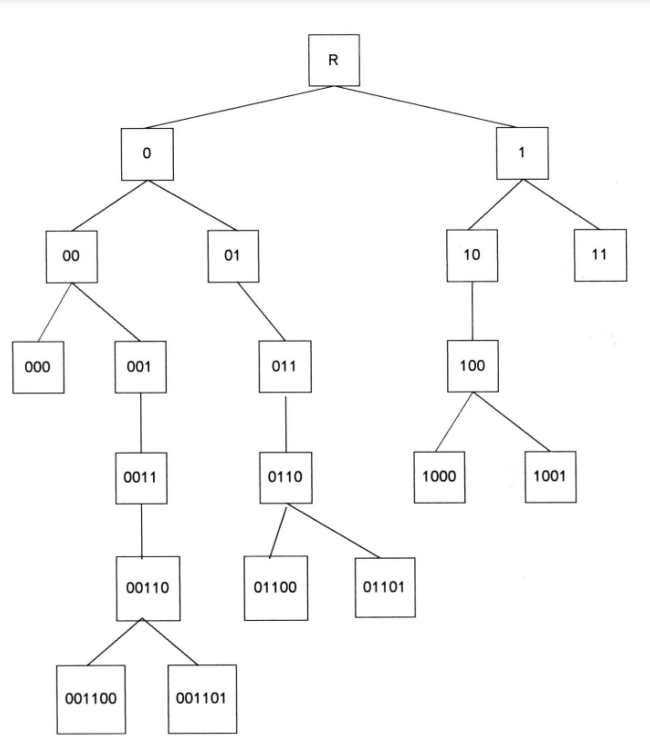
* An n-gram is a data structure that ignores words and treats the input as a continuous data, optionally limiting its processing by interword symbols.
* The data structure consists of fixed length overlapping symbol segments that define the searchable processing tokens. These tokens have logical linkages to all the items in which the tokens are found.
* The advantage of n-grams is that they place a finite limit on the number of searchable tokens.

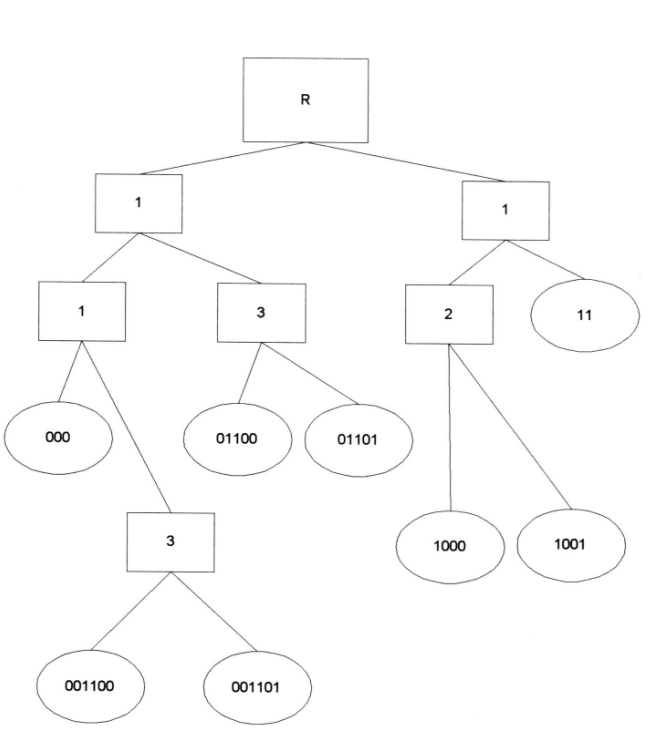


* The maximum number of unique n-grams that can be generated, MaxSeg, can be calculated as a function of n which is the length of the n-grams, and λ which is the number of processable symbols from the alphabet (i.e., non-interword symbols).
* Although there is a savings in the number of unique processing tokens and implementation techniques allow for fast processing on minimally sized machines, false hits can occur under some architectures.

**PAT Data Structure**

* Using n-grams with interword symbols included between valid processing tokens equates to a continuous text input data structure that is being indexed in contiguous “n” character tokens.
* A different view of addressing a continuous text input data structure comes from PAT trees and PAT arrays.
* The input stream is transformed into a searchable data structure consisting of substrings.





Compact trie.