Information Retrieval System Capabilities

- Search Capabilities
- Browse Capabilities
- Miscellaneous Capabilities

1.5 Search Capabilities

- The objective of the search capability is to allow for a mapping between a user's specified need and the items in the information database that will answer that need.
- It can consist of natural language text in composition style and/or query terms with Boolean logic indicators between them.
- the concept "**weighting"** of search terms allow a user to indicate the importance of search terms in either a Boolean or natural language interface
- Weight holds significant potential for assisting in the location and ranking of relevant items.
- Given a natural language query statement, the importance of a particular search term is indicated by a value in parenthesis between 0.0 and 1.0 with 1.0 being the most important.

- The search statement may apply to the complete item or contain additional parameters limiting it to a logical division of the item (i.e., to a zone).
- Based upon the algorithms used in a system, many different functions are associated with the system's understanding of the search statement.
- The functions define the relationships between the terms in the search statement and the interpretation of a particular word.
- Example for the relationships are
    - Boolean,
    - Natural Language,
    - Proximity,
    - Contiguous Word Phrases,

- Fuzzy Searches

- Example for the interpretations are
  - Term Masking,
  - Numeric and Date Range,
  - Contiguous Word Phrases,
  - Concept/Thesaurus expansion

Boolean Logic

- allows a user to logically relate multiple concepts together to define what information is needed.
- Boolean functions apply to processing tokens identified anywhere within an item.
- Typical Boolean operators are **AND, OR,** and **NOT**.
- These operations are implemented using set intersection, set union and set difference procedures.
- Few systems introduced the concept of "exclusive or" but it is more complex query using the other operators and is not generally useful to users.
- A special type of Boolean search is called "M of N" logic.
- The user lists a set of possible search terms and identifies any item that contains a subset of the terms.
- For example, "Find any item containing any two of the following terms: "AA," "BB," "CC."
- This can be expanded into a Boolean search that performs an AND between all combinations of two terms and "OR"s the results together ((AA AND BB) or (AA AND CC) or (BB AND CC)).

SEARCH STATEMENT	SYSTEM OPERATION
COMPUTER OR PROCESSOR NOT MAINFRAME	Select all items discussing Computers and/or Processors that do not discuss Mainframes
COMPUTER OR (PROCESSOR NOT MAINFRAME)	Select all items discussing Computers and/or items that discuss Processors and do not discuss Mainframes
COMPUTER AND NOT PROCESSOR OR MAINFRAME	Select all items that discuss computers and not processors or mainframes in the item

Figure 2.1 Use of Boolean Operators

Proximity

- Proximity is used to restrict the distance allowed within an item between two search terms.
- The semantic concept is that the closer two terms are found in a text the more likely they are related in the description of a particular concept.
- Proximity is used to increase the precision of a search.
- The typical format for proximity is:
  - TERM1 within "m" "units" of TERM2
  - The distance operator "m" is an integer number and units are in Characters, Words, Sentences, or Paragraphs.

SEARCH STATEMENT	SYSTEM OPERATION
"Venetian" ADJ "Blind"	would find items that mention a Venetian Blind on a window but not items discussing a Blind Venetian
"United" within five words of "American"	would hit on "United States and American interests," "United Airlines and American Airlines" not on "United States of America and the American dream"
"Nuclear" within zero paragraphs of "clean-up"	would find items that have "nuclear" and "clean-up" in the same paragraph.

Figure 2.2 Use of Proximity

- A special case of the Proximity operator is the Adjacent (ADJ) operator that normally has a distance operator of one and a forward only direction.
- Another special case is where the distance is set to zero meaning within the same semantic unit.

Contiguous Word Phrases (CWP)

- a way of specifying a query term and a special search operator.
- A Contiguous Word Phrase is two or more words that are treated as a single semantic unit.
- Example
  - "United States of America." The four words that specify a search term representing a single specific semantic concept (a country).
  - "manufacturing" AND "United States of America" returns any item that contains the word "manufacturing" and the contiguous words "United States of America."

- CWP acts like a special search operator that is similar to the proximity (Adjacency) operator but allows for additional specificity.
- If two terms are specified, the contiguous word phrase and the proximity operator using directional one word parameters or the Adjacent operator are identical.
- Nested Adjacencies
  - For contiguous word phrases of more than two terms, the only way of creating an equivalent search statement using proximity and Boolean operators is via nested Adjacencies..
- Proximity and Boolean operators are binary operators
- but contiguous word phrases are an "N"ary operator where "N" is the number of words in the CWP.

- *Other names of CWP*
  - Contiguous Word Phrases are called **Literal Strings** in WAIS.
  - **Exact Phrases** in RetrievalWare.

- In WAIS multiple Adjacency (ADJ) operators are used to define a Literal String (e.g., “United” ADJ “States” ADJ “of” ADJ “America”).

Fuzzy Searches

- Fuzzy Searches provide the capability to locate spellings of words that are similar to the entered search term.
- used to compensate for errors in spelling of words.
- increases recall at the expense of decreasing precision (i.e., it can erroneously identify terms as the search term).
- In the process of expanding a query term fuzzy searching includes other terms that have similar spellings, giving more weight (in systems that rank output) to words in the database that have similar word lengths and position of the characters as the entered term.
- A Fuzzy Search on the term “computer” would automatically include the following words from the information database: “computer,” “compiter,” “conputer,” “computter,” “compute.”

Examples for Interpretations

Term Masking

- ability to expand a query term by masking a portion of the term and accepting as valid any processing token that maps to the unmasked portion of the term.
- The value of term masking is much higher in systems that do not perform stemming or only provide a very simple stemming algorithm.
- There are two types of search term masking:
    - fixed length
    - variable length.
- Fixed length masking
    - is a single position mask.
    - masks out any symbol in a particular position or the lack of that position in a word.
- Variable length “don’t cares”
    - allows masking of any number of characters within a processing token.

- The masking may be
  - in the front
  - at the end
  - at both front and end
  - imbedded.
- The first three of these cases are called suffix search, prefix search and imbedded character string search, respectively.
- The use of an imbedded variable length don't care is seldom used.
- If "*" represents a variable length don't care then the following are examples of its use: "*COMPUTER" Suffix Search

"COMPUTER*" Prefix Search

"*COMPUTER*" Imbedded String Search

SEARCH STATEMENT	SYSTEM OPERATION
multi$national	Matches"multi-national," "multinational," "multinational" but does not match "multi national" since it is two processing tokens.
computer	Matches,"minicomputer" "microcomputer" or "computer"
comput*	Matches "computers," "computing," "computes"
comput	Matches "microcomputers" , "minicomputing," "compute"

Figure 2.3 Term Masking

Numeric and Date Ranges

- Term masking is useful when applied to words, but does not work for finding ranges of numbers or numeric dates.
- To find numbers larger than "125," using a term "125*" will not find any number except those that begin with the digits "125."

Concept/Thesaurus Expansion

- Associated with both Boolean and Natural Language Queries
- ability to expand the search terms via Thesaurus or Concept Class database reference tool.
- thesaurus is typically a one-level or two-level expansion of a term to other terms that are similar in meaning.
- A Concept Class is a tree structure that expands each meaning of a word into potential concepts that are related to the initial term (e.g., in the TOPIC system).
- Concept classes are sometimes implemented as a network structure that links word stems (e.g., in the RetrievalWare system).
- An example of Thesaurus and Concept Class structures are shown in Figure 2.4 and Figure 2.5.

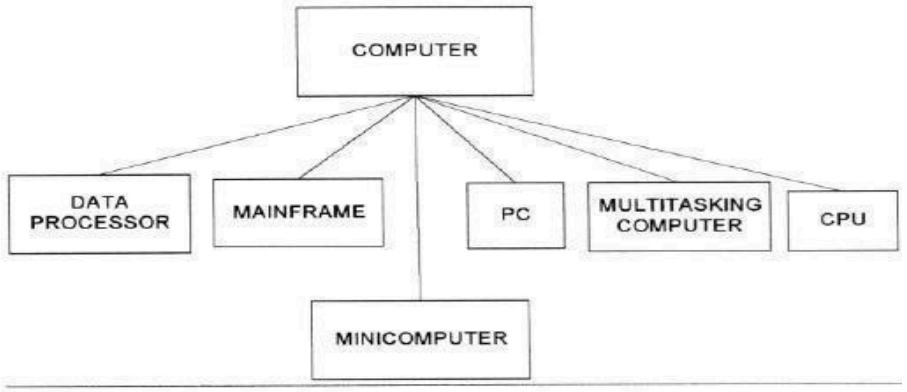

Figure 2.4 Thesaurus for term "computer"

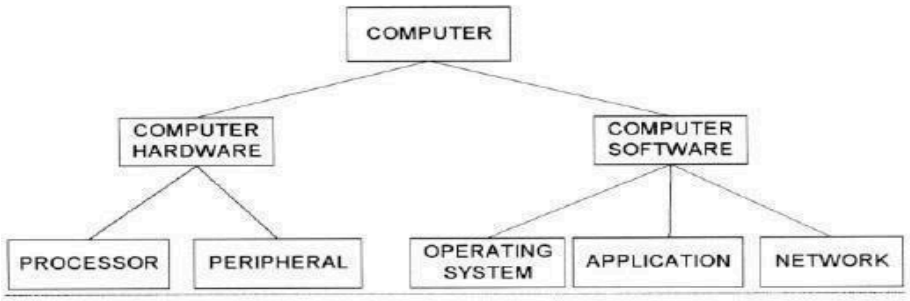

Figure 2.5 Hierarchical Concept Class Structure for "Computer"

- Thesauri are either *semantic* or *based upon statistics.*
- S*emantic thesaurus*
  - is a listing of words and then other words that are semantically similar.

- The problem with thesauri is that they are generic to a language.
- Can introduce many search terms that are not found in the document database.

- *Statistics based Thesauri*
  - uses the database or a representative sample of it to create statistically related terms.
  - It is conceptually a thesaurus in that words that are statistically related to other words by their frequently occurring together in the same items.
  - is very dependent upon the database being searched
  - may not be portable to other databases.

Natural Language Queries

- allows user to enter a prose statement instead of using Boolean.
- Difficulty in using negation with NL queries.
- Natural language interfaces improve the recall of systems with a decrease in precision when negation is required.