**Functional Overview:**

* A total information storage and retrieval system is composed of four major functional processes: Item Normalization, Selective Dissemination of Information, Document Database Search, Index Database Search.

1. **Item Normalization:**

* The first step in any integrated system is to normalize the incoming items to a standard format. Item normalization provides logical restructuring of the item.
* Additional operations during item normalization are needed to create a searchable data structure: identification of processing tokens (e.g., words), characterization of the tokens, and stemming (e.g., removing word endings) of the tokens.
* The processing tokens and their characterization are used to define the searchable text from the total received text. Standardizing the input takes the different external formats of input data and performs the translation to the formats acceptable to the system.
* A system may have a single format for all items or allow multiple formats. One example of standardization could be translation of foreign languages into Unicode. Every language has a different internal binary encoding for the characters in the language. One standard encoding that covers English, French, Spanish, etc. is ISO-Latin.
* To assist the users in generating indexes, especially the professional indexers, the system provides a process called Automatic File Build(AFB).
* Multi-media adds an extra dimension to the normalization process. In addition to normalizing the textual input, the multi-media input also needs to be standardized.
* There are a lot of options to the standards being applied to the normalization. If the input is video the likely digital standards will be either MPEG-2, MPEG-1, AVI or Real Media. MPEG (Motion Picture Expert Group) standards are the most universal standards for higher quality video where Real Media is the most common standard for lower quality video being used on the Internet.
* The next process is to parse the item into logical sub-divisions that have meaning to the user. This process, called “Zoning,” is visible to the user and used to increase the precision of a search and optimize the display.
* A typical item is subdivided into zones, which may overlap and can be hierarchical, such as Title, Author, Abstract, Main Text, Conclusion, and References.
* The zoning information is passed to the processing token identification operation to store the information, allowing searches to be restricted to a specific zone.
* Next, a Stop List/Algorithm is applied to the list of potential processing tokens. The objective of the Stop function is to save system resources by eliminating from the set of searchable processing tokens those that have little value to the system.

1. **Selective Dissemination of Information:**

* The Selective Dissemination of Information (Mail) Process provides the capability to dynamically compare newly received items in the information system against standing statements of interest of users and deliver the item to those users whose statement of interest matches the contents of the item.
* The Mail process is composed of the search process, user statements of interest (Profiles) and user mail files. As each item is received, it is processed against every user’s profile.
* A profile contains a typically broad search statement along with a list of user mail files that will receive the document if the search statement in the profile is satisfied.
* Selective Dissemination of Information has not yet been applied to multimedia sources.

1. **Document Database Search:**

* The Document Database Search Process provides the capability for a query to search against all items received by the system.
* The Document Database Search process is composed of the search process, user entered queries (typically ad hoc queries) and the document database which contains all items that have been received, processed and stored by the system.
* Typically items in the Document Database do not change (i.e., are not edited) once received.

1. **Index Database Search:**

* When an item is determined to be of interest, a user may want to save it for future reference. This is in effect filing it.
* In an information system this is accomplished via the index process.
* In this process the user can logically store an item in a file along with additional index terms and descriptive text the user wants to associate with the item.
* The Index Database Search Process provides the capability to create indexes and search them.
* There are 2 classes of index files:

1. Public Index files
2. Private Index files

* Every user can have one or more Private Index files leading to a very large number of files. Private Index files typically have very limited access lists.
* Public Index files are maintained by professional library services personnel and typically index every item in the Document Database. There is a small number of Public Index files.
* These files have access lists (i.e., lists of users and their privileges) that allow anyone to search or retrieve data.

**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 (referred to as terms in this book) with Boolean logic indicators between them.
* One concept that has occasionally been implemented in commercial systems (e.g., RetrievalWare), and holds significant potential for assisting in the location and ranking of relevant items, is the “weighting” of search terms.
* Given the following natural language query statement where 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.

1. **Boolean Logic**

* Boolean logic allows a user to logically relate multiple concepts together to define what information is needed. Typically the Boolean functions apply to processing tokens identified anywhere within an item.
* The typical Boolean operators are AND, OR, and NOT. These operations are implemented using set intersection, set union and set difference procedures.
* 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 the search statement.

1. **Proximity**

* Proximity is used to restrict the distance allowed within an item between two search terms.
* The semantic concept is that the close search terms in either a Boolean or natural language interface.
* 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 the search statement.

1. **Contiguous Word Phrases**

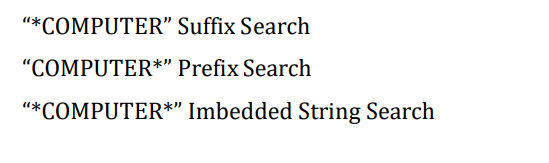
* A Contiguous Word Phrase (CWP) is both a way of specifying a query term and a special search operator.
* A Contiguous Word Phrase is two or more words that are treated as a single semantic unit.
* An example of a CWP is “United States of America.” It is four words that specify a search term representing a single specific semantic concept (a country) that can be used with any of the operators discussed above.
* Thus a query could specify “manufacturing” AND “United States of America” which returns any item that contains the word “manufacturing” and the contiguous words “United States of America.”
* Contiguous Word Phrases are called Literal Strings and Exact Phrases in RetrievalWare.

1. **Fuzzy Searches**

* Fuzzy Searches provide the capability to locate spellings of words that are similar to the entered search term.
* This function is primarily used to compensate for errors in spelling of words.
* Fuzzy searching increases recall at the expense of decreasing precision (i.e., it can erroneously identify terms as the search term).
* Search Google or type a URL Photo by Olga SyrovaIn 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.

1. **Term Masking**

* Term masking is the 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 and variable length. Sometimes they are called fixed and variable length “don’t care” functions.
* Fixed length masking is a single position mask. It 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.



1. **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.”

**Browse Capabilities**

* Once the search is complete, Browse capabilities provide the user with the capability to determine which items are of interest and select those to be displayed.
* There are two ways of displaying a summary of the items that are associated with a query: line item status and data visualization

1. **Ranking**

* Typically relevance scores are normalized to a value between 0.0 and 1.0. The highest value of 1.0 is interpreted that the system is sure that the item is relevant to the search statement.
* In addition to ranking based upon the characteristics of the item and the database, in many circumstances collaborative filtering is providing an option for selecting and ordering output.
* Collaborative filtering has been very successful in sites such as AMAZON.COM MovieFinder.com, and CDNow.com in deciding what products to display to users based upon their queries.
* Rather than limiting the number of items that can be assessed by the number of lines on a screen, other graphical visualization techniques showing the relevance relationships of the hit items can be used.

1. **Zoning**

* Related to zoning for use in minimizing what an end user needs to review from a hit item is the idea of locality and passage based search and retrieval.

1. **Highlighting**

* Most systems allow the display of an item to begin with the first highlight within the item and allow subsequent jumping to the next highlight.
* The highlighting may vary by introducing colours and intensities to indicate the relative importance of a particular word in the item in the decision to retrieve the item.

**IRS AND DBMS**

* IRS is a s/w that consists of all features and functions to manipulate information items, whereas DBMS handles structured data. Using structured data a user can send a request and receive desired information.
* It is tabulated and displayed in the form of a report for easy-to-use purpose.
* An IRS provides the iterative search process to its users and guides them in finding all the required items (known as relevance items).
* The results are provided to the users in relevance to ranked order.
* Though DBMS it is easy to implement it doesn’t provide features like ranking and relevance feedback. Hence both DBMS and IRS must be integrated.
* Commercial database companies must implement these systems that perform this integration. Some examples are INQUIRE DBMS (lasts over 15 years) and Oracle DBMS.
* This can perform integration through an embedded capability called CONVECTIS.
* The RetrievalWare provided by INFORMIX DBMS handles the structured data and information associated with IRS.