studocu

Information Retrieval Systems Capabilities

Design and analysis of software systems (BVRIT Hyderabad College of Engineering for Women)

Scan to open on Studocu

Studocu is not sponsored or endorsed by any college or university
Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Information Retrieval System Capabilities

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

IRS Capabilities

- The capabilities in the information retrieval systems are:

1. Search Capabilities :Querying
2. Browse Capabilities: Browsing
3. Miscellaneous capabilities: Query Refinement

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

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.
- The search capabilities address both Boolean and Natural Language queries.
- User can communicate a description of the needed information to the system.
- Based upon the algorithms used in a system many functions are associated with search statement.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Search Capabilities

- These functions define the relationships between the terms in the search statement.

  -Boolean, Natural Language Queries

  -Proximity Constraints

  -Contiguous Word Phrases

  -Fuzzy Searches

- They also used for interpretation of a particular word.

  -Term Masking

  -Numeric and Data Ranges

  -Concept or Thesaurus expansion

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Boolean Logic

- Boolean logic allows the user to combine words and phrases into *search statements*.
- Operators used are AND, OR, NOT (sometimes XOR).
- These operations are implemented using the corresponding set operations intersection, union and difference.
- Default Precedence is NOT, AND, OR. We use parentheses to specify the order of Boolean operations.
- M of N: Finds any document containing M of the terms T1,…,TN.
- In a Boolean search, synonyms and related terms are not searched on.
- Implied Boolean operators use the plus (+) and minus (-) symbols in place of the full Boolean operators, AND and NOT.

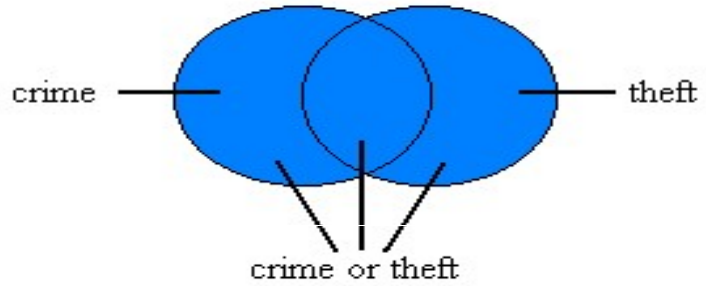
This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

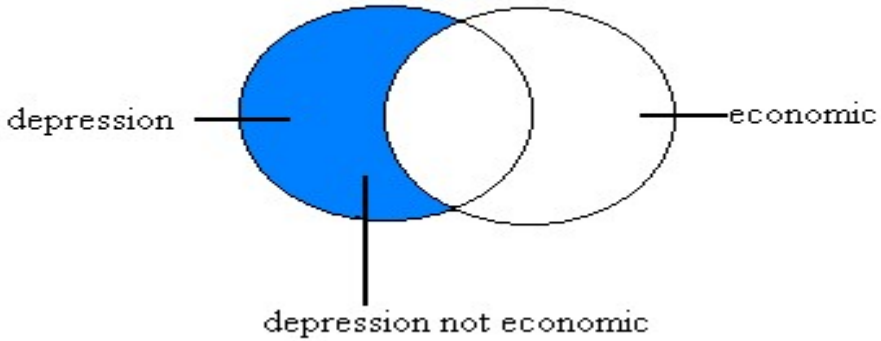
Boolean Logic

Search Statement: drug abuse and teenagers

Search Statement: crime or theft



Search Statement: depression not economic



Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Boolean Logic

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

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Problem with boolean logic

- Boolean queries often result in either too few (=0) or too many (1000s) results.
  - AND gives too few; OR gives too many
  - AND will **narrow** a search but OR will **broaden** a search.

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Natural Language Queries

- Natural Language Queries allow a user to enter a prose statement without any special syntax or format.
- The longer the prose, the more accurate file results returned.
- Natural language search takes into account the meanings of the words in your query
- In Natural Language Queries it is difficult to specify negation in the search statement.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Proximity Constraints

- Proximity is used to restrict the distance allowed within an item between two search terms.
- Proximity specifications increase the precision of the search.
- General Format: *TERM1 within m units of TERM2*

  UNIT may be character, word, paragraph, etc.
- Sometimes the proximity relationship contains Direction operator to specify which term should appear first.
- 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.

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Proximity Constraints

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

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Contiguous Word Phrases(CWP)

- A Contiguous Word Phrase is two or more words that are treated as a single semantic unit.
  - "United States of America".
- CWP is N-ary (not Boolean)operator and it Cannot be expressed as Boolean query.
- If only two are specified, then CWP reduces to the adjacent proximity operator.
- Also called literal strings or exact phrases.

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Fuzzy (Approximate)Search

- Fuzzy Searches provide the capability to locate spellings of words that are similar to the entered query terms.
- Fuzzy matching compensates for spelling errors, especially when documents were scanned-in and then subjected to optical character recognition (OCR).
- Increased recall (more documents qualify because new terms may be matched) at the expense of decreased precision.
- Example: COMPUTER may match COMPITER, CONPUTER, etc.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Term masking

- Term masking is the ability to expand a query term by masking a portion of the term and accepting as valid.
- 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 but a 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, or imbedded.

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Term masking

- **Example:** the term MULTI$NATIONAL will be matched by “multi-national” or “multinational”(but not by “multi national” since it is a sequence of twoterms)

- Suffix: *WARE will match terms that end with “ware”.
- Prefix: WARE* will match terms that begin with “ware”.
- Imbedded: *WARE* will match terms that contain “ware”.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Numeric and Date Ranges

- Term masking is useful when applied to words, but does not work for finding ranges of numbers or numeric dates.
- User can enter **inclusive** or **infinite ranges** as a part of query.
  - **Numeric query terms:**
  1. >125 (matches all numbers greater that 125)
  2. <=223(matches all documents less than or equal to 223)
  3. 125-425 (matches all numbers between 125 and 425).
  - <=125(matches all numbers less than or equal to125.
  - **Date query terms**: 28/01/13 - 28/01/14 (matches all dates between 28 jan 2013 and 28 jan 2014).

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

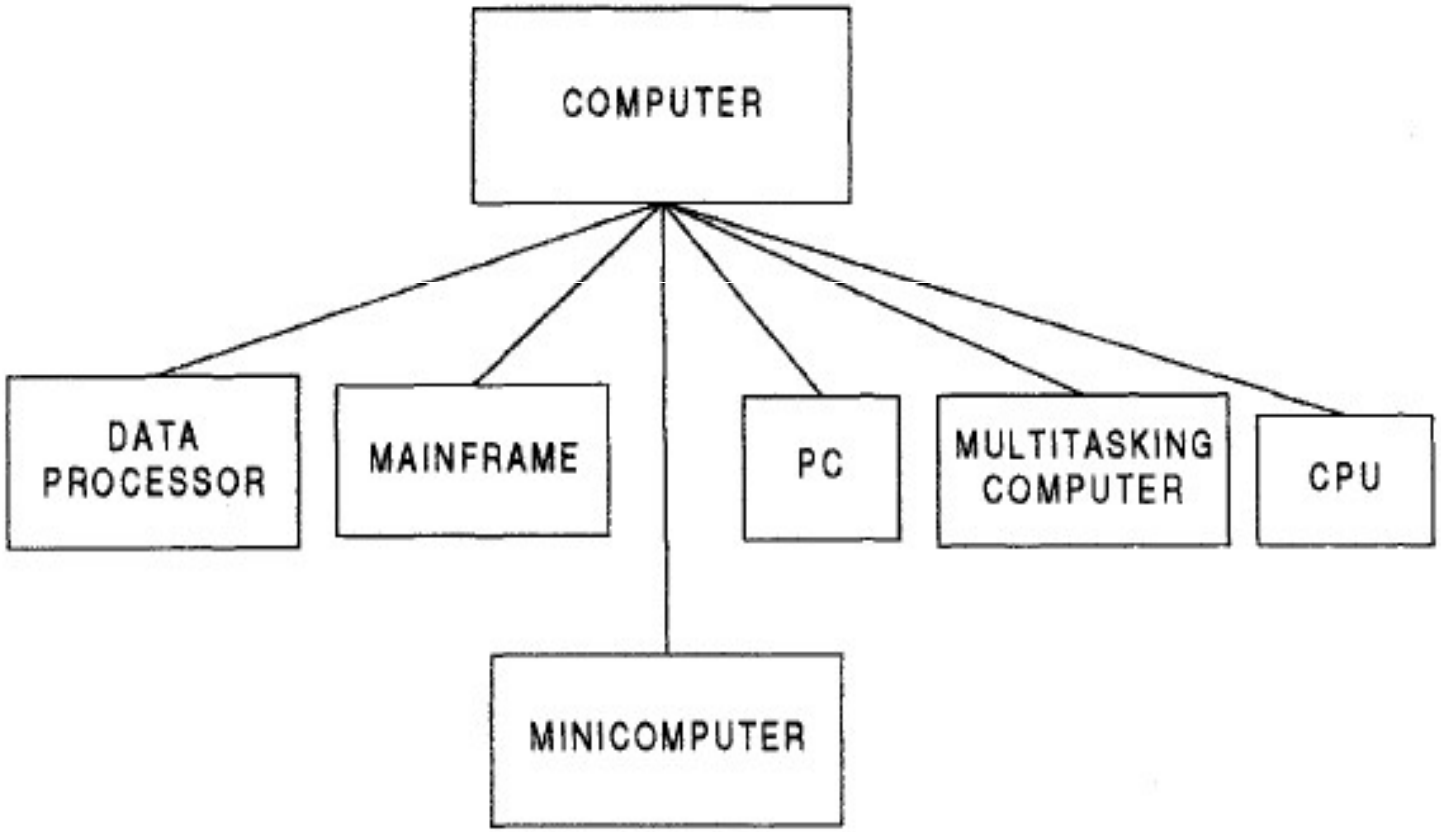
Concept/Thesaurus Expansion

- The ability to expand the search terms via Thesaurus or Concept classes.
- A 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
- A hierarchy(tree) of concepts is called **Concept hierarchy.**

This document is available free of charge on studocu

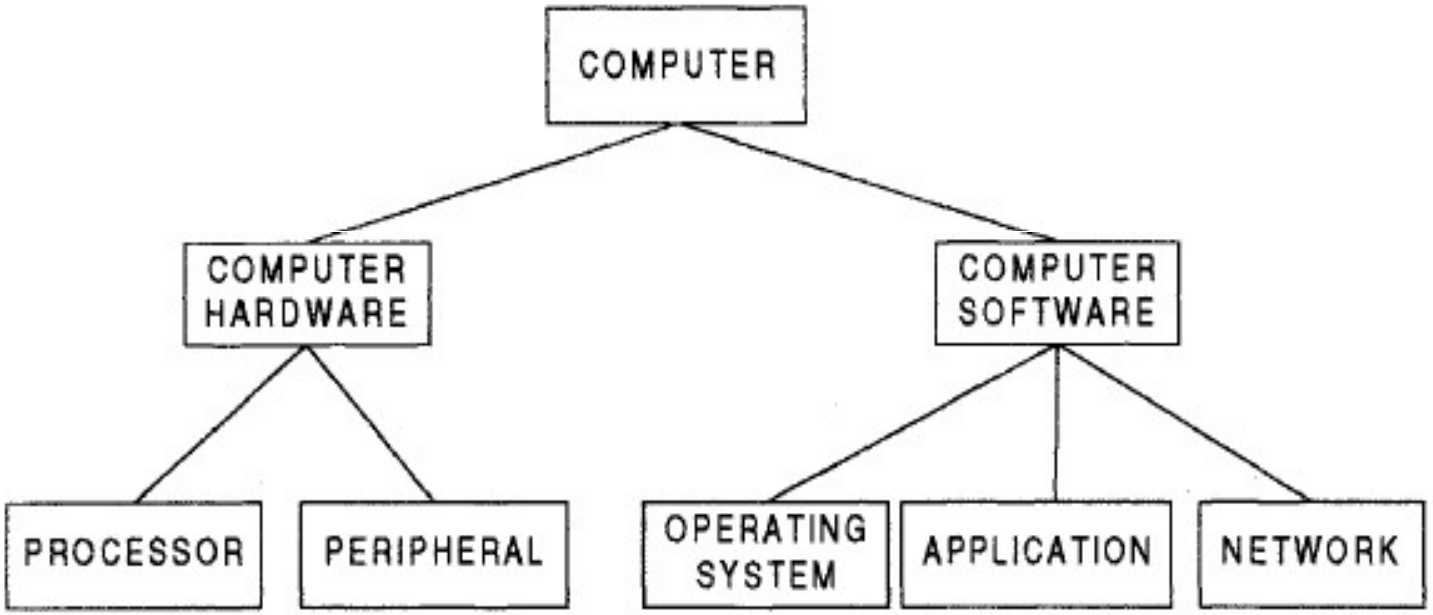
Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Thesaurus for term "computer"



Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Concept Hierarchy for term "Computer"'



This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Thesaurus -Types

- Thesauri are either semantic or based upon statistics.
- **Semantic thesaurus**:Groups together terms that are similar in meaning (a single level concept hierarchy).

- **Statistical thesaurus:**Groups together terms that are statistically related(occur together in the same documents).

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Concept/Thesaurus Expansion

- Replacing a query term by an ancestor(more general) term increases recall and decreases precision.
- Replacing a query term by a descendant (more specific)term decreases recall and increases precision.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Browse Capabilities

- Determine which items are of interest and select those to be displayed.
- There are two ways of displaying a summary of the retrieved items:
  - Line item status
  - Data visualization
- Powerful browsing capabilities are particularly important when precision is low.

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Browse Capabilities

- Browse capabilities can assist the user in focusing on items that have the highest likelihood in meeting his need.

  -Ranking

  -Zoning

  -Highlighting

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Ranking

- In Boolean systems all retrieved documents equally meet the query criteria.
- Documents are displayed in arbitrary, or in sorted order(alphabetically by title or chronologically by date).
- With the introduction of ranking based systems the status summary displays the relevance score(0-1) associated with the item along with a brief descriptor of the item.
- In these systems the retrieved documents are sorted by **relevance**. This allows the user to determine at what point to stop reviewing items.

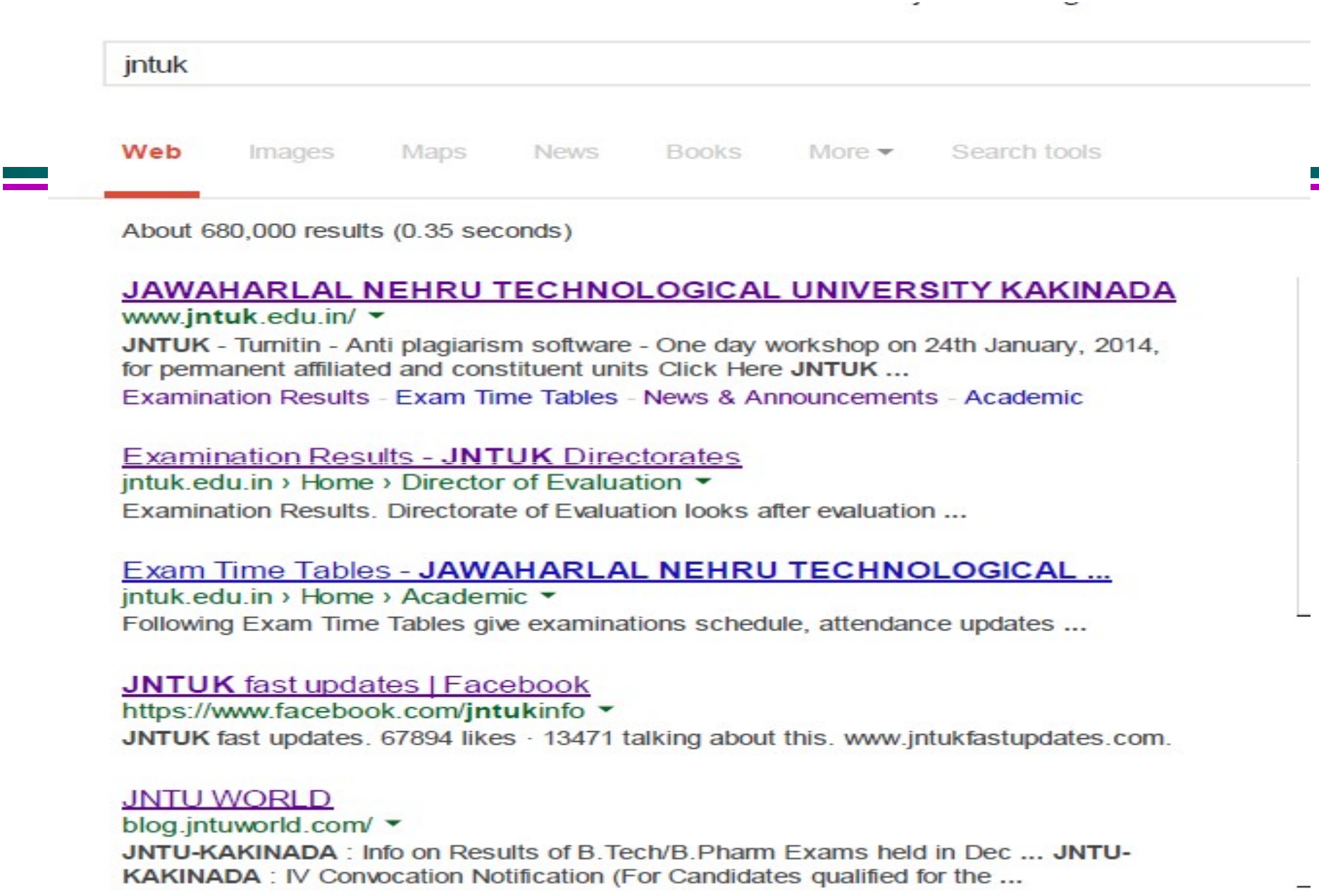
Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Zoning

- Assists the user to see the *minimum information* needed to determine the relevant items.
- Limited display screen sizes require selectability of what portions of an item a user needs to see.
- For example, display of the **Title** and **Abstract** may be sufficient information for a user to view Journal Paper.
- Limiting the display of each item to zones allows multiple items to be displayed on a single display screen.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)



Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Highlighting

- Lets the user quickly focus on the potentially relevant parts of the text to scan for item relevance.
- Different strengths of highlighting indicates how strongly the highlighted word participated in the selection of the item.
- 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 colors and intensities to indicate the relative importance of a particular word in the item.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Miscellaneous Capabilities

- Facilitate the user's ability to input queries, reducing the time it takes to generate the queries, and reducing *the probability of entering a poor query.*
- Vocabulary browse
- Iterative search and Search History Log
- Relevance feedback
- Canned Query

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Vocabulary browse

- Provides knowledge on the processing tokens available in the database.
- provides the capability to display in alphabetical sorted order words from the document database
- Logically, all unique words (processing tokens) in the database are kept in sorted order along with a count of the number of unique items in which the word is found.
- The user can enter a word or word fragment and the system will begin to display file dictionary around the entered text.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Vocabulary browse

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Vocabulary Browse List with entered term "comput"

TERM	OCCURRENCES
compromise	53
comptroller	18
compulsion	5
compulsive	22
compulsory	4
comput	
computation	265
compute	1245
computen	1
computer	10,800
computerize	18
computes	29

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Iterative Search and Search History Log

- Process of refining the results of a previous search to focus on relevant items.
- The results of the previous search can be used as a constraining list to create a new query.
- The search history log is the capability to display all the previous searches that were executed during the current session.
- search history logs are also used as starting points for new searches,

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Relevance feedback

- The old query is replaced by a new query
- The new query is a transformation of the old query, reflecting feedback about the relevance of the documents retrieved by the first query.

This document is available free of charge on studocu

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)

Canned Query

- The capability to name a query and store it to be retrieved and executed during a later user session is called canned or stored queries.
- Allows users to store previously-used queries and incorporate additional search criteria to retrieve data that is currently needed.
- Queries that start with a canned query are significantly larger.

Downloaded by Saidulu chittimalla (chittimallasaidulu3@gmail.com)