

**TH 3713**  
**CAPAIAN MAKLUMAT MULTIMEDIA**

---

**TOPIK 1**  
**Pengenalan kepada Capaian Maklumat**

**Objektif:**

1. Capaian Data
2. Capaian Maklumat
3. Sistem Capaian Maklumat
4. Proses Capaian Maklumat
5. Penilaian prestasi capaian

TH3713 20052006 2

**INFORMATION RETRIEVAL**

- Problems of information storage and retrieval attract attention since 1940s
- Vast amounts of information to which accurate and speedy access is becoming more difficult → problem of effective retrieval remains unsolved
- In principle, information storage and retrieval is simple. Suppose there is a store of documents and a person (user of the store) formulates a Q's (request a query) to which the answer is a set of documents satisfying the information need expressed by his question. He can obtain the set by reading all the documents in the store, retaining the relevant docs and discarding all the others. In a sense, this constitutes 'perfect' retrieval. A user either does not have the time or does not wish to spend the time reading the entire docs collection, apart from the fact that it may be physically impossible for him to do so

TH3713 20052006 3

Data is Different  
Types

| NAME  | AGE | SALARY | DATE JOINED |
|-------|-----|--------|-------------|
| ERRA  | 35  | 5000   | 5/2/05      |
| YUSRY | 53  | 10000  | 3/4/96      |

"Jules Verne wrote 20K Leagues Under The Sea and Around The World in 80 days. He died in 1905"

words

TH3713 20052006 4

And so are Search Statements

- SQL
  - SELECT Name
  - FROM Employee
  - WHERE Age
  - BETWEEN 30 AND 40
- Artificial Language
- Exact Description

TH3713 20052006 5

**Relevance**

- Relevance is the core concept in IR
  - But nobody has a good definition
  - Relevance = useful information
  - Relevance = related information
  - Relevance = new information
  - Relevance = interesting information
  - Relevance = ???
  - However we still want relevant *information*

TH3713 20052006 6

- Many automatic information retrieval systems are experimental.
- **Experimental IR** mainly carried out in a lab situation whereas **operational systems** are **commercially systems** which charge for the service they provide.
- The two systems are evaluated differently
- The real world IR systems are evaluated in terms of **'user satisfaction'** and the price user is willing to pay for their services.
- Experimental IR system are evaluated by **comparing** the retrieval experiments with **standards** specially constructed for the purpose.

TH3713 20052006 7

- IR → representation, storage, organization of, and access to **information items**
- Focus is on the **user information need**
- User information need:
  - Find all docs containing information on college tennis teams which: (1) are maintained by a USA university and (2) participate in the NCAA tournament.
  - When did the Buffalo Bills last win the Super Bowl?
  - ???
- Emphasis is on the retrieval of information (not data)

TH3713 20052006 8

DATA RETRIEVAL

- DR we are normally looking for an **exact match** -> checking to see whether an item is/not present in the file.
  - IR sometimes be of interest but more generally we want to find those items which **partially match** the request and then select from those a few of the best matching ones.
- Describe data in **DR is deterministic** but **IR is probabilistic**. In DR probabilistic do not enter into the processing
- Query language in DR – artificial, one with restricted syntax and vocabulary. In IR– use natural language.

TH3713 20052006 9

- In DR the query is generally a **complete specification** of what is wanted, in IR it is **incomplete**.
- In IR we are searching for the **relevant docs**.
- **DR is more sensitive to error** in the sense that an error in matching will not retrieve the wanted item which implies a total failure of the system.
  - In IR errors in matching generally do not affect performance of the system

TH3713 20052006 10

| DR vs IR            |               |                           |
|---------------------|---------------|---------------------------|
|                     | DR            | IR                        |
| Matching            | exact match   | partial match, best match |
| Model               | deterministic | probabilistic             |
| Query language      | artificial    | natural                   |
| Query specification | complete      | incomplete                |
| Items wanted        | matching      | relevant                  |
| Error response      | sensitive     | insensitive               |

Data retrieval

- which docs contain a set of keywords?
- Well defined semantics & structure
- a single erroneous object implies failure!

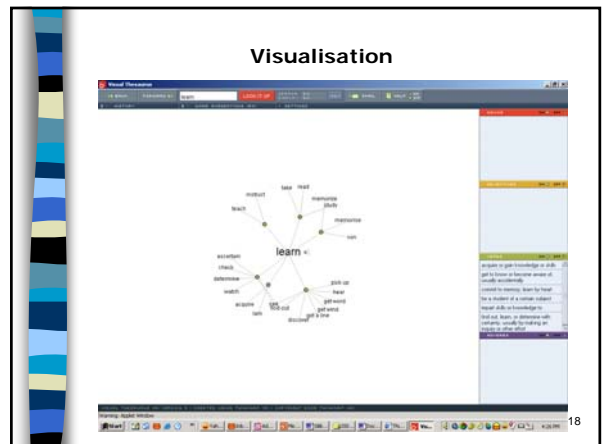
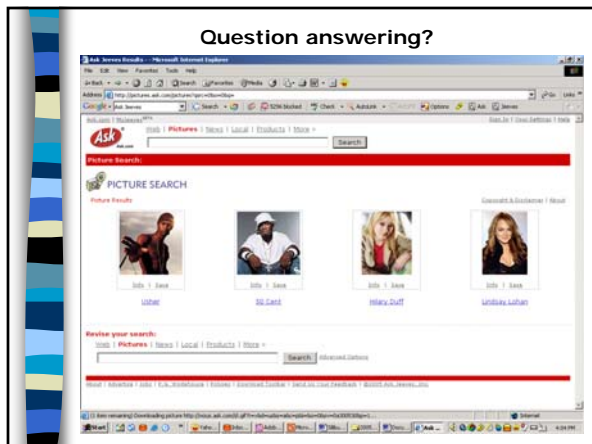
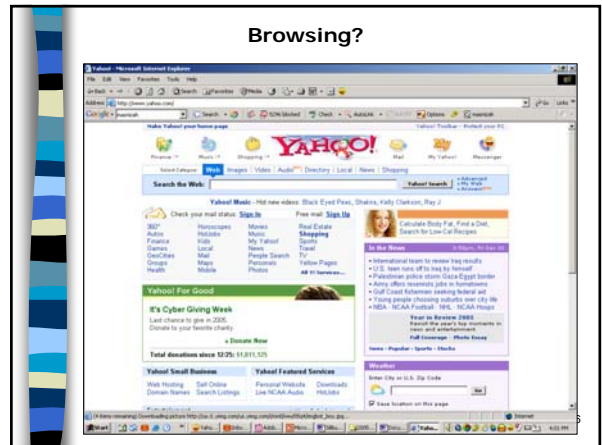
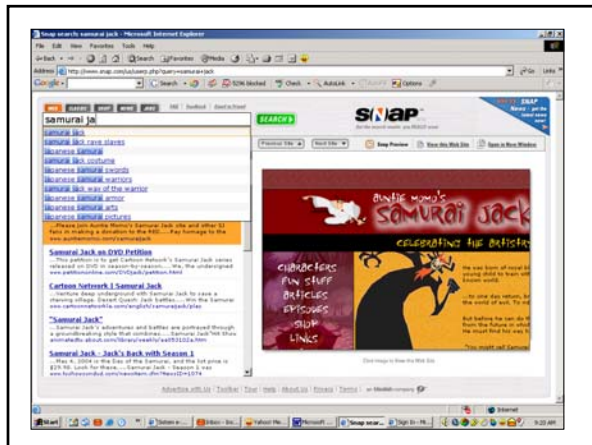
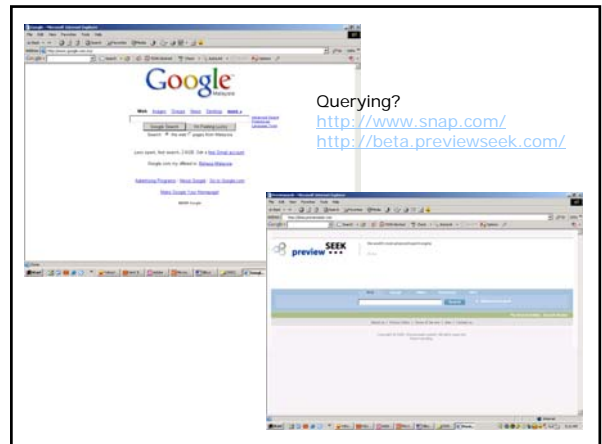
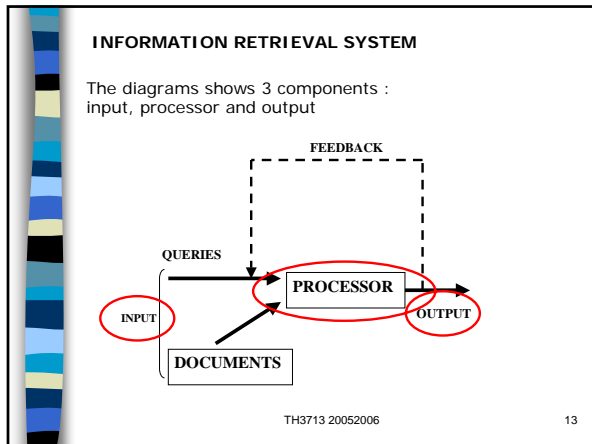
Information retrieval

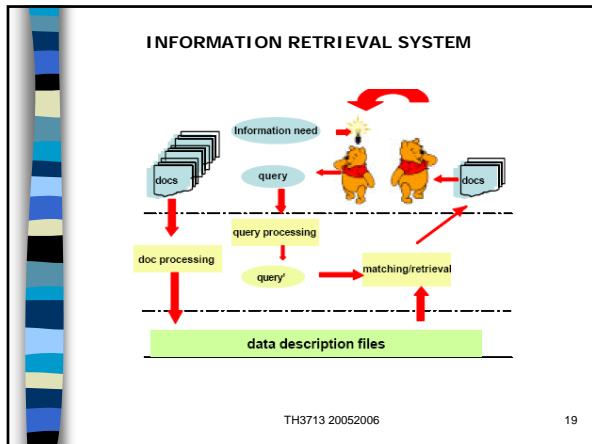
- information about a subject or topic
- deals with *unstructured text*
- semantics is frequently loose
- small errors are tolerated

IR system:

- interpret contents of information items
- generate a *ranking* which reflects relevance
- The primary goal of an IR system is to retrieve all documents which are relevant to a user query.

TH3713 20052006 12





- ### Why is this hard?
- Documents/images/video/speech/etc are complex
  - We need some representation but
    - Semantics
      - What words mean
      - context (how we use words)
    - Natural language
      - How we say things
  - Computers cannot deal with these easily  
eg Opinion
- TH3713 20052006 20

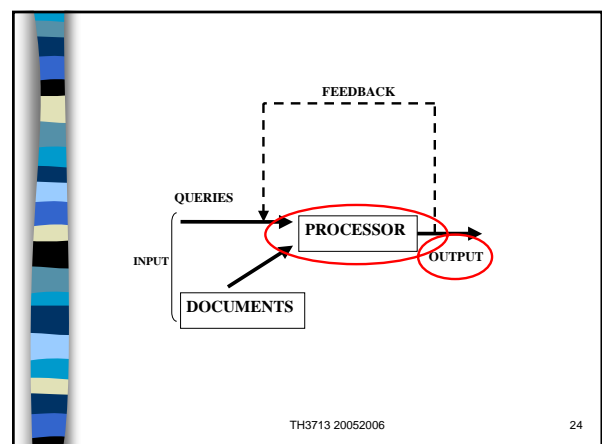
How do we describe this?

- Words, colours, shapes, text-extraction?

TH3713 20052006 21

- ### Why is this hard?
- Information needs must be expressed as a query
- But user's don't often know what they want
  - Problems
    - Verbalising information needs
    - Understanding query syntax
    - Understanding search engines
- SEARCH VS GUESS WHAT I MEAN?
- TH3713 20052006 22

- ### Why is this hard?
- Information is often dynamic
    - News
    - Web pages
    - Weather maps
    - Etc
  - And so are queries
    - Searchers may change information need while searching
  - IR must cope with change in data and searcher
- TH3713 20052006 23



- When the retrieval system is online, it is possible for the user to change his request during one search session, improving retrieval run. Such a procedure is called **FEEDBACK**.
- Processor** – retrieval process. Process involve structuring the information in some appropriate way(classifying etc). Also involve performing the actual retrieval function, that is executing the search strategy in response to a query.
- Output** – set of citations or documents numbers. In an operational system the story ends here, in an experimental system it leaves the evaluation to be done

TH3713 20052006 25

### Comparison of different information systems

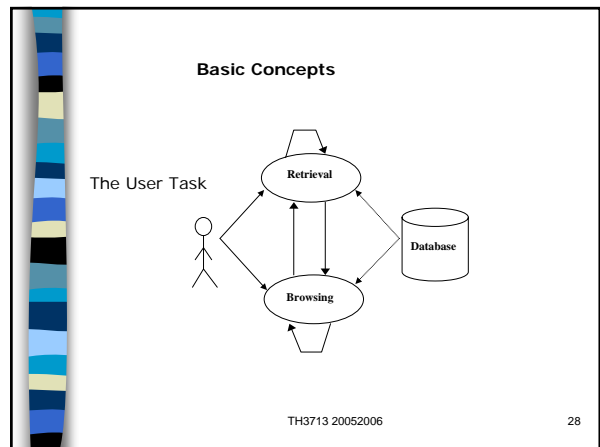
| Discipline | Data Object        | Primary Operation         | DB Size          |
|------------|--------------------|---------------------------|------------------|
| IR         | Document           | Retrieval (probabilistic) | Small – v. large |
| DBMS       | Table              | Retrieval (deterministic) | Small – v. large |
| AI         | Logical statements | Inference                 | small            |

TH3713 20052006 26

### Basic concept -IR

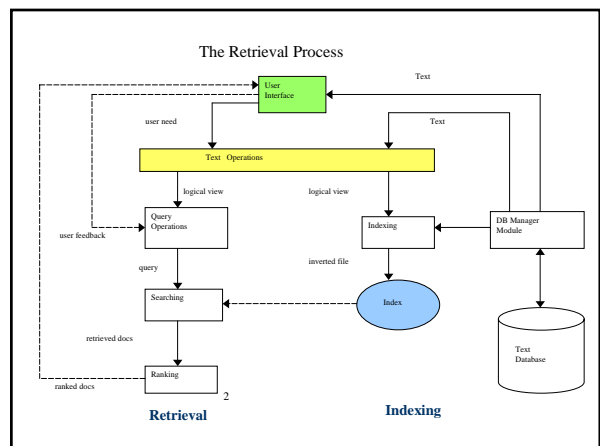
- Effective retrieval of information affected by:
  - User task
  - Logical view of document

TH3713 20052006 27



- User task**
  - User translate his information need into query – specify a set of words which convey the semantics
  - User searches useful information executing a *retrieval* task
  - Browsing* document– still a process of retrieving information, not searching; main objective not clearly defined and purpose might change during interaction with the system
- Logical view of the documents**
  - Documents in a collection is represented through a set of index terms or keywords.
  - Keyword might be extracted directly from the text or specified by human

TH3713 20052006 29



### The Retrieval Process

- Rajah
- First define the text database done by the manager of the db which specify
  - Document used
  - Operation to be perform on the text
  - Text model/structure and what elements can be retrieved
- Once the logical view of documents is defined, db manager (DB Manager Module) builds an index of the text
  - An index is a critical data structure – allow fast searching over large volume
  - Different index might be used, most popular one is [inverted file](#)

TH3713 20052006 31

### Inverted File

| Term        | Record | Frequency |
|-------------|--------|-----------|
| computer    | 1      | 3         |
| computer    | 3      | 5         |
| computing   | 2      | 1         |
| distributed | 2      | 1         |
| parallel    | 1      | 2         |
| system      | 2      | 1         |
| ...         | ...    | ...       |

TH3713 20052006 32

### The Retrieval Process

- Once document database is indexed, retrieval process can be initiated
- User specify [user need](#) then [query operation](#) is applied. Query is processed to obtain the retrieved documents
- Before sent to user, the retrieved documents are ranked according to relevance
- User examine the set ranked documents in the search. At this point, user might initiate a [user feedback cycle](#) → system uses the documents selected to change query formulation to get better representation

TH3713 20052006 33

### Effectiveness and efficiency

- Much R & D in IR aimed at improving the effectiveness and efficiency of retrieval
- Efficiency is usually measured in terms of the computer resources used such as CPU time.
- Effectiveness is measured in terms of [precision](#) and [recall](#)
  - [Precision](#) the ratio of the number of relevant documents retrieved to total number of documents retrieved
  - [Recall](#) is the ratio of the number of relevant documents retrieved to the total number of relevant documents (both retrieved and not retrieved)

TH3713 20052006 34

**Rajah Capaian Dokumen**

X - Bilangan dokumen yang relevan yang dicapai  
 Y - Bilangan dokumen yang tak relevan yang dicapai  
 R - Bilangan dokumen relevan dalam koleksi  
 NR - Bilangan dokumen yang tidak relevan dalam koleksi

Kejituan (**Precision**) =  $\frac{X}{X + Y}$       Dapatan semula (**Recall**) =  $\frac{X}{R}$

TH3713 20052006 35

### Retrieval Performance Evaluation

- Before final implementation of IR system, system evaluation is carried out
- DR system
  - response time and space required
- IR System
  - Beside time and space, require precision of the answer set (relevance ranking – DR system don't have)

TH3713 20052006 36

**Penilaian prestasi capaian (retrieval performance evaluation)**

- Dalam capaian maklumat tradisional, klasifikasi dokumen dijalankan berdasarkan jadual di bawah:

|               | <u>relevan</u>          | <u>tak relevan</u>     |
|---------------|-------------------------|------------------------|
| Retrieved     | A (correctly retrieved) | B (falsely retrieved)  |
| Not retrieved | C (missed)              | D (correctly rejected) |

TH3713 20052006 37

**KESIMPULANNYA??**