의료정보검색
Information Retrieval

정보검색이란

- Information Retrieval
- 원하는 정보를 찾는 것
- Data retrieval vs. Information retrieval

What is IR?

- Information Retrieval is a science which deals with the knowledge representation, storage, organization and access of information items.

Need for Information Retrieval

- 얼마나 많은 정보를 활용할 수 있는가?
- 정보는 과연 재활용할 수 있는 형태로 기록되어 있는가?
- Index, search등을 쉽게 할 수 있는가?
- 다른 의료기관에서도 같은 표현을 사용하고 있는가?
- 진료를 얼마나 지원해 줄 수 있는가?

NLM and Medline

- 10 million articles
- 3,500 journals since 1966
- PubMed, Internet Grateful Med
  - http://www.nlm.nih.gov/

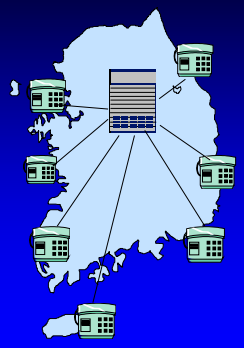
IR 관련된 기술분야

- Internet
- Search Engine
- Vocabulary System
- Information Modeling
- Filtering and Classification
- Natural Language Processing
- …….

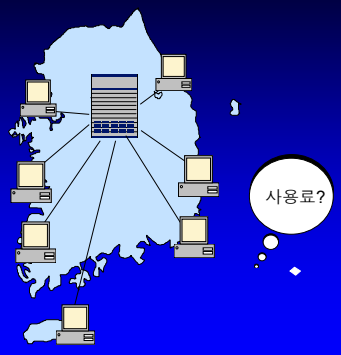
Internet

- TCP/IP를 사용하는 전세계적인 network
- public(not free, but open to everyone)
- carrier of electronic mail
- convenient to get free SW
- terabytes of information
- dynamic rerouting

Telephone network

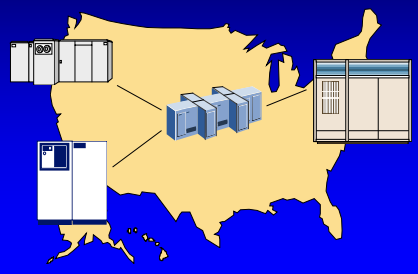


Another network



Network in early stage

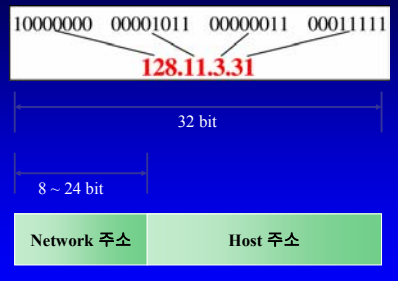
- 국방성 ARPAnet
- TCP/IP 통신프로토콜 사용



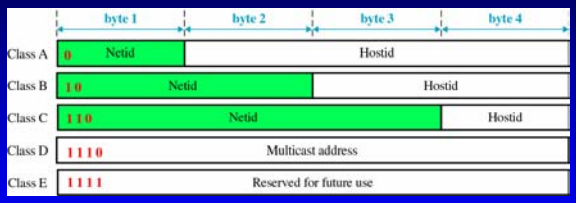
TCP/IP

- Protocol
  - rules of behavior
-
  - 한국 : 정지 신호 준비
  - 독일 : 출발 준비
- TCP/IP
  - 2 widely used network protocols : computer network 에 접속하기 위한 100 여 가지의 규약

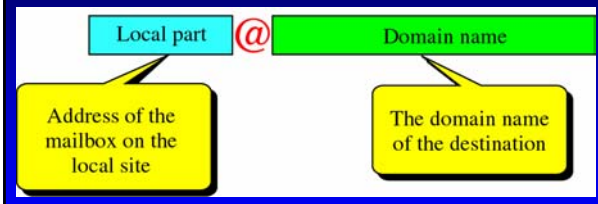
Internet Address



Internet Classes



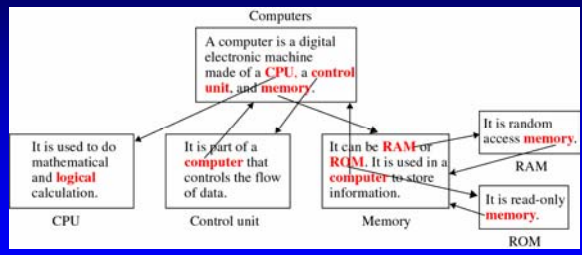
E-mail Address



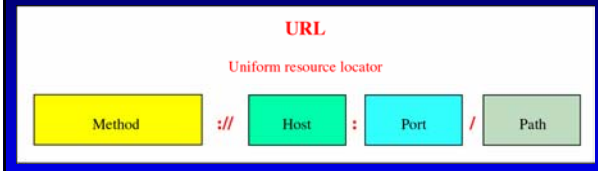
jinchoi@snu.ac.kr

URL

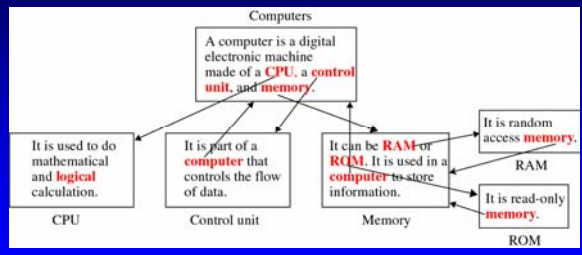
Hypertext



URL



Hypertext



DNS

- Domain Name System(DNS)
  - 전세계적으로 흩어져 있는 컴퓨터의 소재 정보를 관리하는 컴퓨터시스템
  - 도메인 이름을 IP주소로 변환하는 역할

www.kyobobook.co.kr ⟷ 203.112.118.101
www.cnn.com ⟷ 198.137.240.92

인터넷에서 정보 찾기

- Search engine 이용
- News group에 문의
- Mailing list 활용

Internet Search Engine

- www.yahoo.com
  - www.yahoo.co.kr
- www.altavista.com
  - www.altavista.co.kr
- www.excite.com
- www.lycos.com,
  - www.lycos.co.kr
- www.dreamwiz.com
- www.naver.com

Internet Search Engine

검색엔진의 특징별 비교

구분	장점	단점
디렉토리형 (주제형)	• 특정한 주제별로 정보를 찾을 때 유리하다. • 특별한 주제어나 키워드를 모를 경우 사용하기 편하다.	• 단계별 분류를 거쳐 정보를 찾게 되므로 잘못 들어설 경우 시간적 낭비가 크다.
로봇 에이전트형 (키워드형)	• 몇 개의 키워드를 통해 원하는 정보를 쉽게 찾을 수 있다.	• 1~2개 정도의 단어나 검색식이 부정확할 경우 정확한 찾기가 어렵다.
메타형	• 한번의 키워드 입력만으로 원하는 정보를 쉽게 찾을 수 있다.	• 다른 검색엔진들을 참조해야 하므로 검색속도가 느리다. • 특정한 검색엔진별로의 검색에서 실패할 경우도 많다.

AND search

- Search for Monet AND Renoir
- Search for +Monet +Renoir
- Search for Monet Renoir
  - "All the words" option

OR search

- Search for UPS U.P.S.
- Search for UPS OR U.P.S.
- Search for UPS U.P.S
  - "Any of the Words" option
- "foreign policy" vs foreign policy

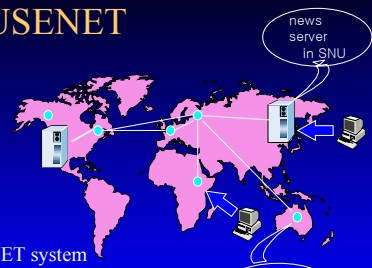
NOT search

- Search for "bugs life" -ants
- Search for "bugs life" NOT ants
- Search for "bugs life" AND NOT ants

Near Search

- Korea NEAR climate
  - Altavista (advanced search)
  - two terms within 10 words
- Korea NEAR climate
  - Lycos (advanced search)
  - two terms within 25 words

USENET



- USENET system
  광범위한 게시판 시스템
- news group
  USENET 에 개설된 토론 그룹을 말함

Newsgroup Search Engine

Mailing list

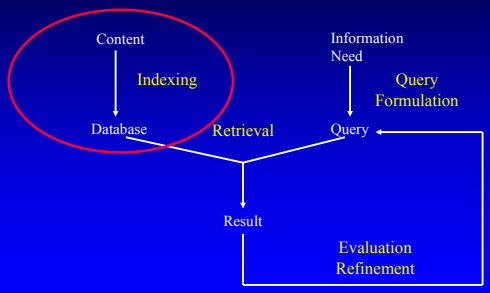
- automatic mailing programs
  - LISTSERV
  - Majordomo

IR Modeling

IR steps

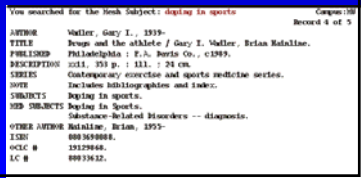
- Text processing
- Indexing
  - inverted file
  - signature file
- Organization in DB
- Query processing
- Evaluation

Information-Retrieval Process

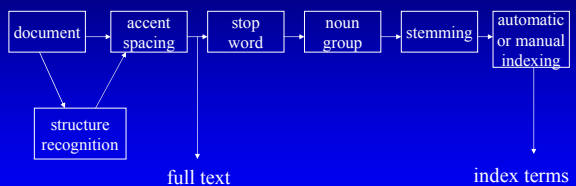


Index

- ✓ *Index content (bibliographic information) vs. Full-text content*
- ✓ *Item vs. Attribute (항목 vs. 속성, 예: 저자, 홍길동)*
- ✓ *Subjects vs. keyword search*
- ✓ *색인의 목적은 검색속도 향상. 순차검색 vs. 이진검색*
- ✓ *Bibliographic index vs. full-text index*

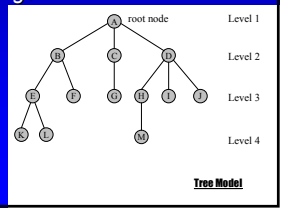


Indexing Process

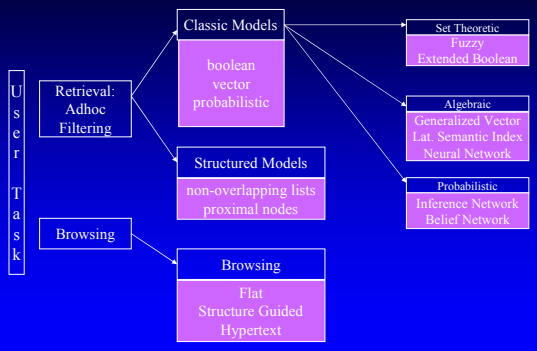


MeSH (Medical Subject Headings)

- ✓ *Controlled vocabulary*
- ✓ *15 trees*
- ✓ *Tree vs. graph*
- ✓ *Human indexer*
- ✓ *http://www.ncbi.nlm.nih.gov/*

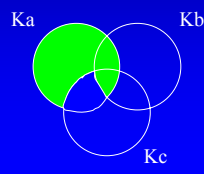


Classification of IR



Boolean Model

- query can be written in disjunctive normal form
- $q = k_a \wedge (k_b \vee \neg k_c)$
- $q_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$



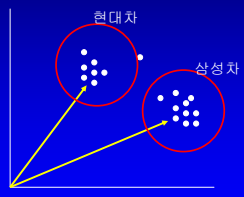
Vector Model and Weight function

K = {소나타, 2000cc, 자동변속, 흰색, …, k_t}

D_1 = {20, 20, 11, 5, … ,5}

D_2 = {20, 18, 12, 4, …, 5}

D_{30} = {0, 20, 12, 3, …,9}



weight terms are assumed to be mutually independent !

Boolean vs. Vector model

	Petroleum	Mexico	Oil	Texas	Refinery	Ship
Boolean	(1	1	1	0	1	0)
Vector	(2.8	1.6	3.5	3	3.1	1)

Retrieval Issues

- Indexing
  - inverted file
- Ranking
  - relevance에 따른 ranking
  - chronology에 따른 ranking
- Display

(aspirin, prevention)
(prevention, …)

(aspirin, attack, heat)

Item	Attribute (Doc. #)
Aspirin	1,5,6,9
Attack	3,6,7,8
Heart	4,6,7,10
Prevention	1,2,6,9

Indexing with Inverted File

Document	Text
1	Gold silver truck
2	Shipment of gold damaged in a fire
3	Delivery of silver arrived in a silver truck
4	Shipment of gold arrived in a truck

Number	Term	Times: Documents Words
1	a	<3: (2:7),(3:6),(4:6)>
2	arrived	<2: (3:4),(4:4)>
3	damaged	<1: (2:4)>
4	delivery	<1: (3:1)>
5	fire	<1: (2:7)>
6	gold	<3: (1:1),(2:3),(4:3)>
7	of	<3: (2:2),(3:2),(4:2)>
8	in	<3: (2:5),(3:5),(4:5)>
9	shipment	<2: (2:1),(4:1)>
10	silver	<2: (1:2),(3:3,7)>
11	truck	<3: (1:3),(3:8),(4:7)>>

Evaluation of IR

- Recall
  - 전체 찾아야 되는(원하는 내용을 담고있는) 문서 중 검색되어진 문서 숫자
- Precision
  - 검색된 문서 중 찾고자 하는 문서를 갖고 있는 비율

Word based indexing 문제점

- Context
  - meaning is affected by meaning of other words
    - high, blood, pressure
    - low *pressure* at *high* altitude increase red *blood* cell
- Polysemy
  - lead vs lead
- Synonymy
  - hypertension vs. high blood pressure
- Granularity
  - antibiotics, penicillin
- Focus of Content
  - Key word vs Plain word

The End