Search Internet (1/2)

Information Literacy for Academic Studies

Instructor: Chenhui Chu
Email: chu@i.kyoto-u.ac.jp

Teaching Assistant: Yikun Sun
E-mail: sun@nlp.ist.i.kyoto-u.ac.jp

Course Description

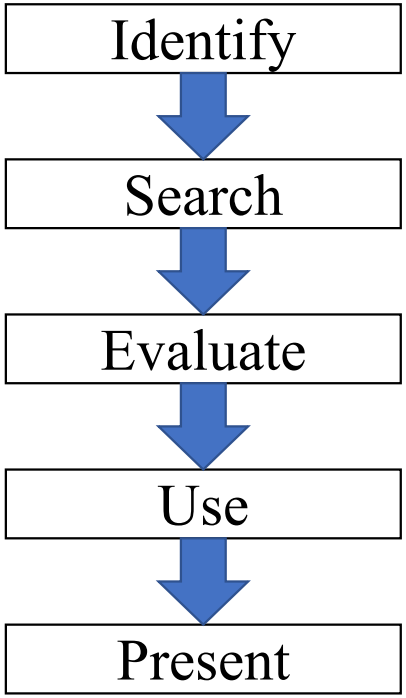
- This course is designed to train you to be able to

effectively identify / search / evaluate / use / present the information for decision making / problem solving in your academic studies.

- This course focuses on the abilities of autonomous and life-long learning which is essential in today's society.

Information Literacy (IL)

- Identify the problem and the information needs, and determine the extent.
- Develop a search strategy which can access the needed information effectively and efficiently.
- Evaluate the information obtained and its sources critically.
- Extract, summarize and analyze the information into your knowledge base, and effectively accomplish the task.
- Write a paper and give a presentation. Do use information ethically and legally (citation).



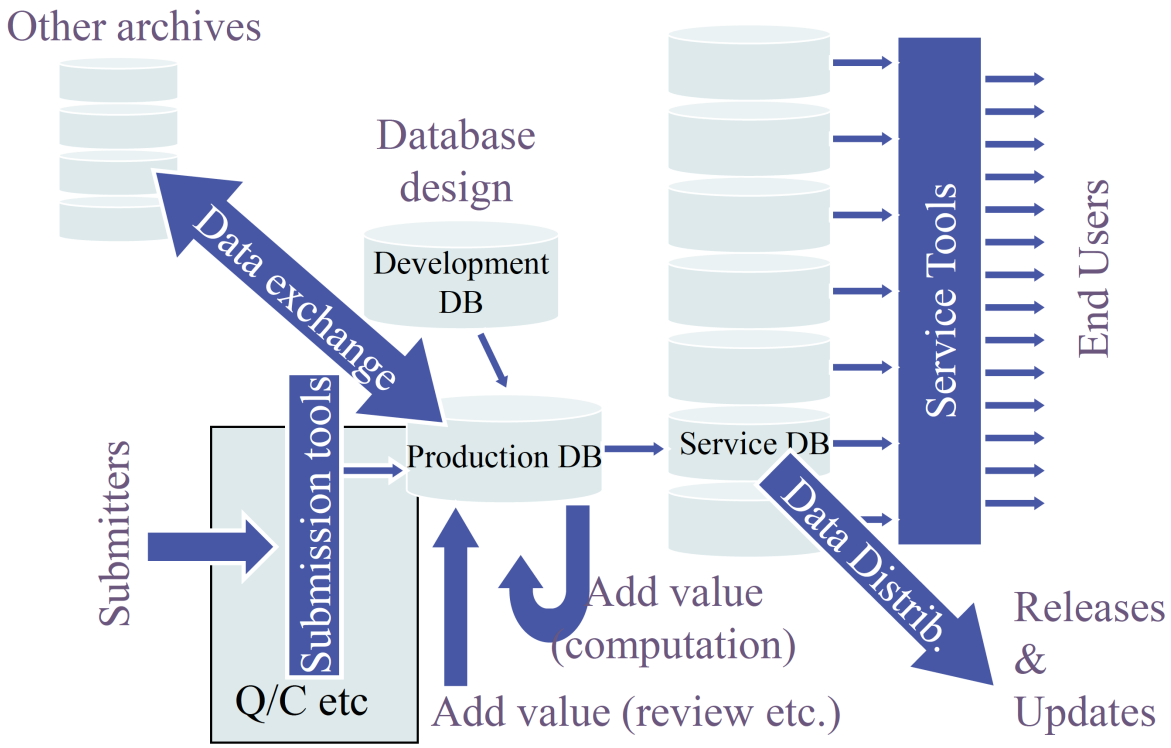
Outline of this Course

- Basic concepts of information literacy
- Study strategies (2/2)
- Searching in library
- Searching databases
- Searching internet (1/2)
- Evaluating sources (3 weeks)
- Referring sources and academic integrity (2 weeks)
- Presenting information (2 weeks)

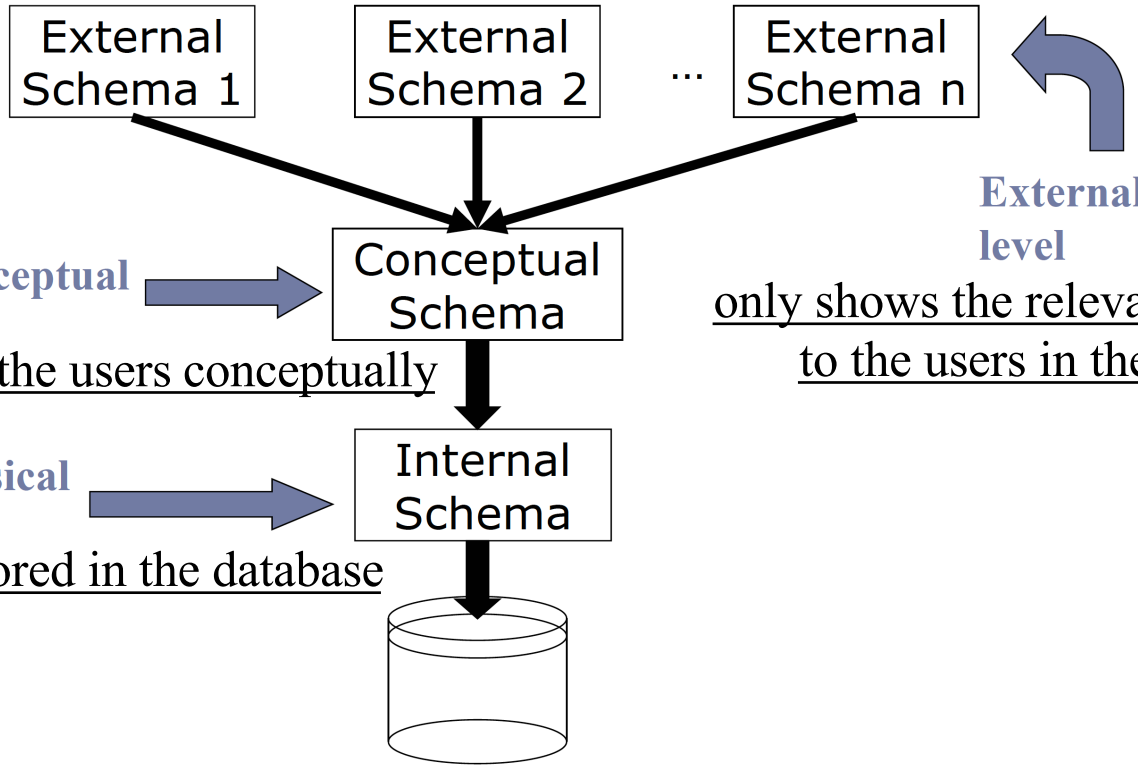
What is a Database

- A database is a large, integrated collection of data.
- A database contains a model of something!
- A database management system (DBMS) is a software system designed to store, manage and facilitate access to the database.
- A database
  - manages Very Large Amounts of Data
  - supports efficient access to Very Large Amounts of Data
  - supports concurrent access to Very Large Amounts of Data
  - supports secure, atomic access to Very Large Amounts of Data

Today, Database Systems are Ubiquitous



Three-level Architecture



External level

only shows the relevant database content to the users in the form of views

Conceptual level

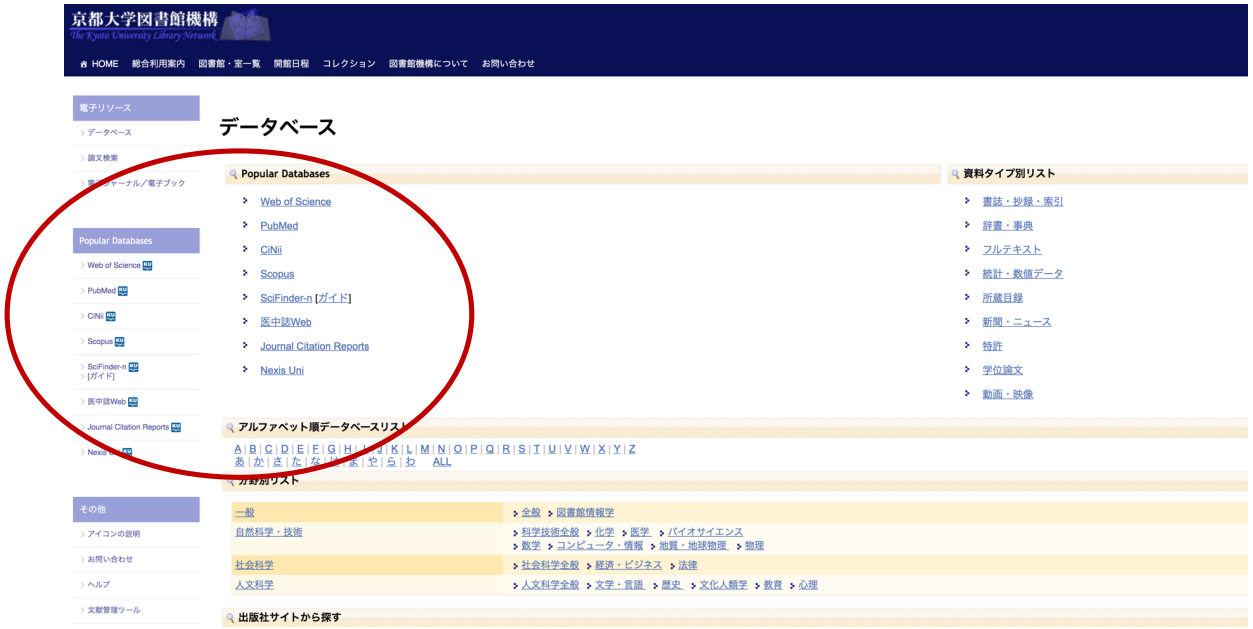
how the database appears to the users conceptually

Physical level

how data is actually stored in the database

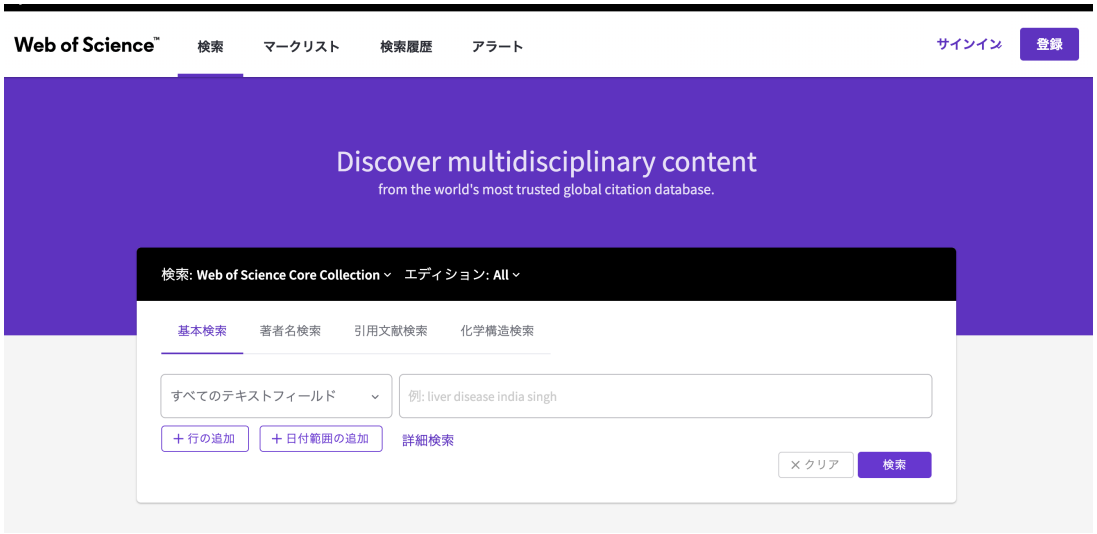
Library Database (1/2)

- All of our library databases can be accessed from our library webpage.
- Choosing the right one for your topic is the first step in using them.
- These are some of the large products that cover a wide range of topics. Think of them sort of as the supermarket of databases.



Library Database (2/2)

- Some More Specific Databases



Boolean Search

- Boolean operators are words such as **and**, **or**, and **not** that you use to combine search terms.
- The operator you use will either broaden or narrow the results of your search.

Operator	Use	Example
AND	limits your search	AI AND physics AI AND physics AND art
OR	expands your search	AI OR physics AI OR physics OR art
NOT	excludes specific terms	AI NOT NLP

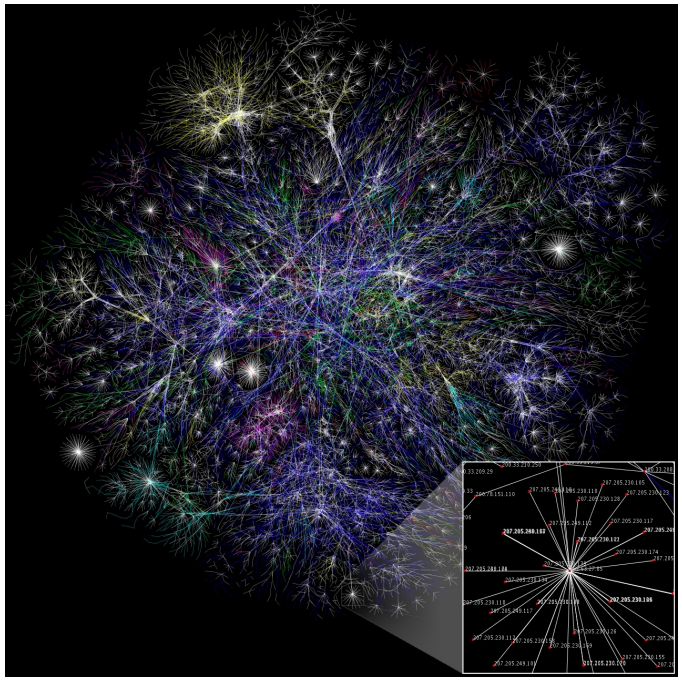
Task 5

- List the database software/products you ever used in your PC and/or Internet. What did you use them for?
- What is the relation between database and Internet?

Can you share your discussion of task 5?

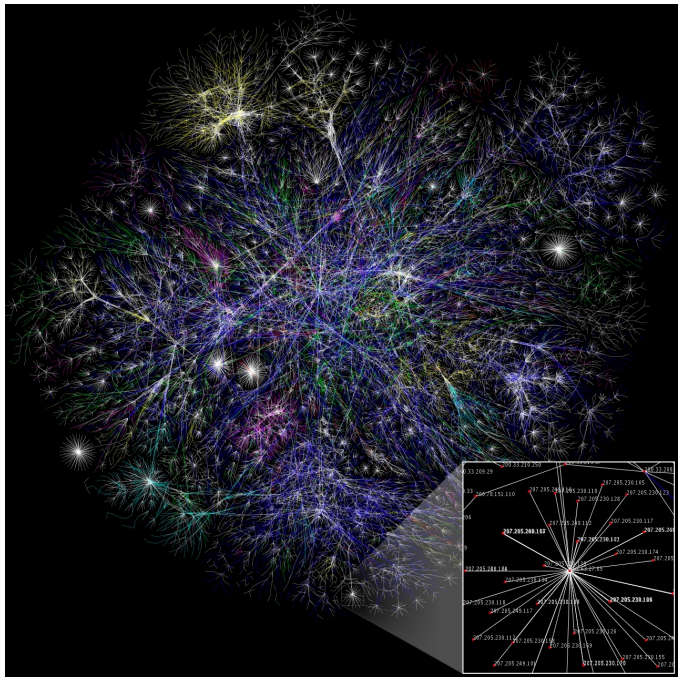
Internet (1/2)

- Internet is a worldwide network of computers that are connected by cables.
- It allows people all across the world to communicate with each other instantaneously, and has revolutionized how we share and access information.
- It consists of several different "protocols," such as: email, the World Wide Web, FTP, and Telnet.



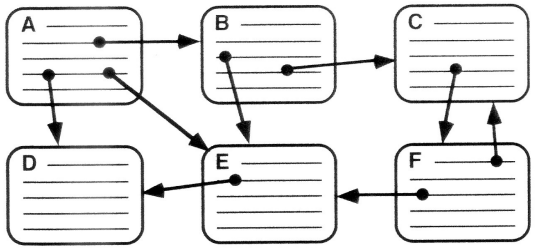
Internet (2/2)

- No one person or group is in charge of the Web. It's a very democratic way of disseminating information.
  - Every one has a voice, but also make junk.
- The Web does NOT have everything.
  - People often believe that it has everything, but only a small fraction of the world of information is available (for free).
  - Many useful information doesn't exist in digital format.



World Wide Web (1/2)

- A system of interlinked hypertext documents that run on and are accessed via the Internet.
- With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them via hyperlinks.
- Hypertext is text displayed on a computer display or other electronic devices with references (hyperlinks) to other text which the reader can immediately access, or where text can be revealed progressively at multiple levels of detail.



World Wide Web (2/2)

- Hyperlink is a reference to data that the reader can directly follow either by clicking or by hovering or that is followed automatically.
- A hyperlink points to a whole document or to a specific element within a document.

- The web page in WWW can be static; or could be a web page generated instantly and dynamically by computer programs extracting information from a specialized database.

URL

- Stands for Uniform Resource Locator.
- Is simply the address for a web page.
- Often tell you a lot, and aid you in your searching and surfing the web.

How to Read a URL

http://www.i.kyoto-u.ac.jp/en/index.html

- http:// is called the “protocol”, always be either http:// or https://.
- www is the name of the server, can be other names “mail”, “news”, “blog”.
- ac.jp is the domain name, identifies the organization.
- i.kyoto-u.ac.jp/en/ After the domain name is a bunch of sections separated by slashes (/). Each slash represents a different directory or folder where the page is located.
- index.html is the name of the file, or web page, that you're looking at. The name index is often used as a default for an entire folder.

URL Can Help You

- Know whether a web form is secure or not (http vs. https)
- Figure out what kind of organization is hosting the site.
  - This can help in evaluating the information found.
- Find what you were looking for.
  - If you get lost within a web site, it is possible to backtrack by eliminating each folder until you find something useful.

You Can Guess a URL:

- Type “www” (most browsers don't require you to type in http://)
- Add the name or abbreviation of the organization (ex. Microsoft, google)
- Add the domain name:
  - .co if it is a company
  - .ac if it is a school
  - .go if it is a government agency
  - .org if it is an organization

Web Domains

- It tells you who is sponsoring the site.
- The last part of the domain name is called the top-level domain code, always say: what kind of organization this is or what country the website is from.

Code	Type of Organization
.com, .co	Business or commercial enterprise
.edu, .ac	Educational institution
.gov, .go	Government agency
.org	Non-profit organization
.net	Network provider (Internet Service Providers)
.mil	Military organization

Code	Country
.us	United States
.uk	United Kingdom (Great Britain)
.au	Australia
.fr	France
.mx	Mexico
.se	Sweden
.ca	Canada
.de	Germany
.jp	Japan
.cn	China
.it	Italy

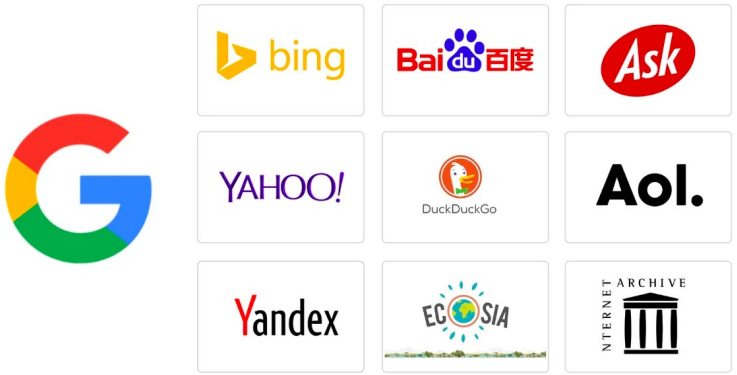
Search Internet

- The Internet contains a lot of information, but it is not indexed like the library's catalog.
- Searching the Internet requires part skill, part luck, and a little bit of art.
- Search Tools on the Internet
  - 1. Search engines: Uses a computer program (called web-crawler) to navigate through the web and collect information about web pages.
  - 2. Subject directories / web directories: Manual entry and classification
  - 3. Invisible web / deep web: Includes dynamic electronic databases that are not searchable through search engines.

Search Engines

- Search Engine is a tool enabling document search, with respect to specified keywords, in the Web and returns a list of documents where the keywords were found.

- They maintain a large index for a huge number of Internet sites by retrieving each individual web pages.



Components of Web Search Engine

- 1. User Interface
- 2. Parser
- 3. Web Crawler
- 4. Database
- 5. Ranking Engine

User Interface

- It is the part of Web Search Engine interacting with the users and allowing them to query and view query results.

Parser (1/2)

- It is the component providing term (keyword) extraction for both sides.
  - The parser determines the keywords of the user query and all the terms of the Web documents which have been scanning by the crawler.
- Term extraction procedure includes the following sub-procedures:
  - 1. Tokenization: A process of converting a sequence of characters into a sequence of tokens
    - E.g., sum = 3 + 2;
    - Tokenized and represented by this table:

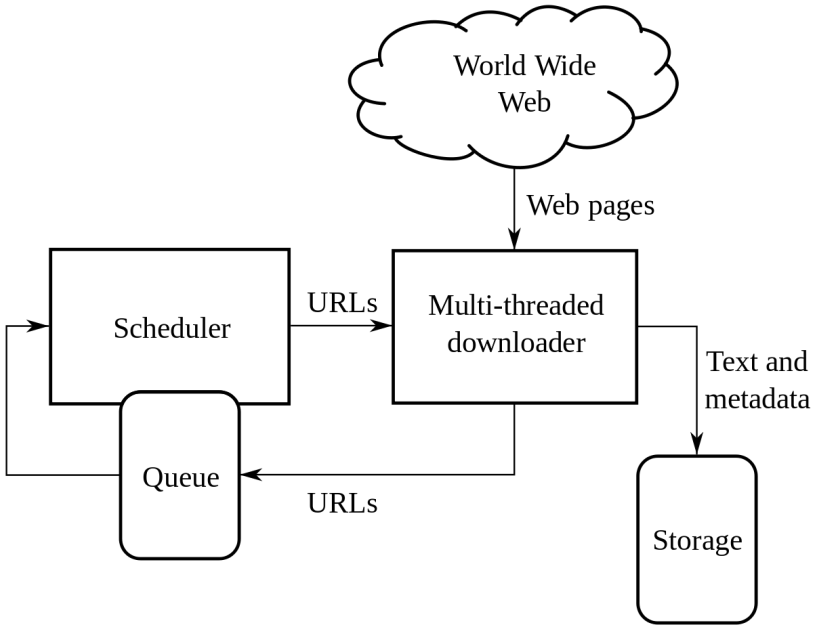
Lexeme	Token type
sum	**Identifier**
=	**Assignment operator**
3	**Integer literal**
+	**Addition operator**
2	**Integer literal**
;	**End of statement**

Parser (2/2)

- Term extraction procedure includes the following sub-procedures:
  - 2. Stemming
    - Normalizing words with respect to their different syntactical variations and converting to the same root form.
    - e.g. “runs”, “runner”, “running” converted to “run”.
    - Recall of IR systems is improved but precision may degrade (e.g., “cope”, “cop” → “cop”).
  - 3. Stop word handling
    - Stop words - terms with little semantic meaning or little discriminative power that usually occur very frequently in the language, e.g. “a”, “about”, “that”, “there”, etc.
    - Removing stop words helps to reduce the size of the index (about 20%-30%)

Web Crawler

- A web crawler is a relatively simple automated program, or script, that methodically scans or “crawls” through Internet pages to create an index of the data it is looking for.
- Alternative names for a web crawler include web spider, web robot, crawler, and automatic indexer.



Work Mode of Web Crawler

- 1. When a web crawler visits a web page, it reads the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich meta tags.
- 2. Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information.
- 3. Lastly, the website is included in the search engine's database and its page ranking process.

Database

- It is the component that all the text and metadata specifying the web documents scanned by the crawler.

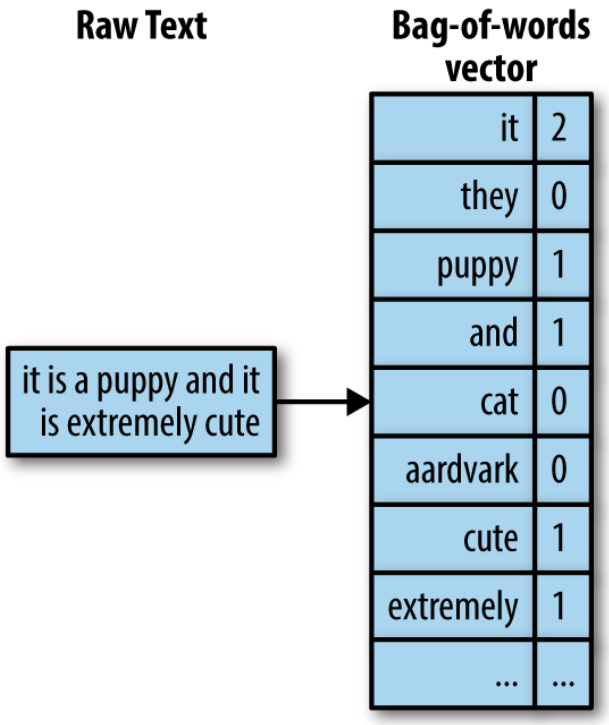
Ranking Engine

- The component is mainly the ranking algorithm operating on the current data, which is indexed by the crawler, to be able to provide some order of relevance, for the web documents, with respect to the user query.

- Three components:
  - 1. Documents representation
  - 2. Indexing
  - 3. Ranking algorithm

Documents Representation (1/2)

- Use “Bag of Words”
- Terms are considered to be independent from each other
- Term position in the sentence is irrelevant
- No order is considered between the terms in sentences
  - “Mary is taller than John” and “John is taller than Mary” have the same representation

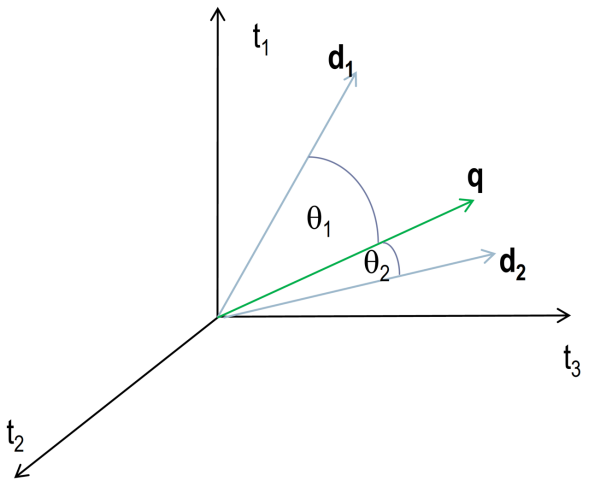


Documents Representation (2/2)

- Vector Space Model
  - Document represented as a vector of the dimension equal to the number of vocabulary V in the document collection
- Each element of a vector corresponds to a given term and its value reflects the weight given to the term
  - Several different weighting schemes for document terms $d_j = \lfloor w_{1,j}, w_{2,j}, ..., w_{|V|,j} \rfloor$
  - Query is also represented as a term vector $q = \lfloor w_{1,q}, w_{2,q}, ..., w_{|V|,q} \rfloor$
  - Similarity between the query and document vector representations reflects document relevance for the query

Query vs. Documents Similarity Measure

- V-dimensional vector space where document terms are axes of the space.
  - Documents and query are represented as vectors in this space.
  - Best approach is to use angle between the query and document vectors as a measure of document relevance
- Similarity measures
  - Dot product: $Sim(D,Q) = \sum (a_i * b_i)$
  - Cosine: $Sim(D,Q) = \frac{\sum_i (a_i * b_i)}{\sqrt{\sum_i a_i^2 + \sum_i b_i^2}}$
  - Dice: $Sim(D,Q) = \frac{2\sum_i (a_i * b_i)}{\sum_i a_i^2 + \sum_i b_i^2}$
  - Jaccard: $Sim(D,Q) = \frac{\sum_i (a_i * b_i)}{\sum_i a_i^2 + \sum_i b_i^2 - \sum_i (a_i * b_i)}$



Indexing

- Sequentially searching through document collection for a given query is ineffective
- Large search engines use inverted indices
  - Data structures optimized for fast search and retrieval
  - Enable locating relevant pages through binary search on the ordered list of terms
- Inverted index
  - Associates terms with page pointers
  - Term frequency in a page can be recorded
  - Positional index contains also term offsets in pages

Inverted Index

d_1 Kyoto(1) is(2) in(3) Kansai(4).

d_2 Kyoto(1) was(2) the(3) capital(4) of(5) Japan(6) and(7) is(8) in(9) Kansai(10).

d_3 Tokyo(1) is(2) the(3) largest(4) city(5) in(6) Japan(7).

d_4 Tokyo(1) and(2) Kyoto(3) are(4) located(5) in(6) Japan(7). Tokyo(8) is(9) bigger(10) than(11) Kyoto(12).

Simple inverted index

bigger: d_4
capital: d_2
city: d_3
japan: d_2 d_3 d_4
kansai: d_1 d_2
kyoto: d_1 d_2 d_4
largest: d_3
located: d_4
Tokyo: d_3 d_4

Inverted index with position and count information

bigger: [d_4,10,1]
capital: [d_2,4,1]
city: [d_3,5,1]
japan: [d_2,6,1] [d_3,7,1] [d_4,7,1]
kansai: [d_1,4,1] [d_2,10,1]
kyoto: [d_1,1,1] [d_2,1,1] [d_4,<3,12>,2]
largest: [d_3,4,1]
located: [d_4,5,1]
tokyo: [d_3,1,1] [d_4,<1,8>,2]

Query: *japan city*
→d_3

Query: *japan kyoto*
→d_2 d_4 as relevant documents and d_4 is the best one

Query: Kyoto kansai
→ d_1 d_2 as relevant documents and d_1 is the best one

Stop words not indexed

Search with Inverted Index

- For a given query composed of n terms the search using the inverted index is done as follows:
  - 1. Vocabulary search in the inverted index to locate the documents containing any of the query terms.
  - 2. Merging results to find documents that contain all query terms and documents that contain only subset of terms.
  - 3. Ranking documents depending on their coverage of query terms and other features such as page importance (e.g., PageRank), timestamp, etc.

Task 6

- Discuss how a search engine works
  - It can be a wrap up of what you learned from this lecture
  - Also, please try to discuss beyond what you learned from this lecture

- Submit your discussion report in pdf named as [**student id_name**] via PandA by next lecture