Search Internet (2/2)

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

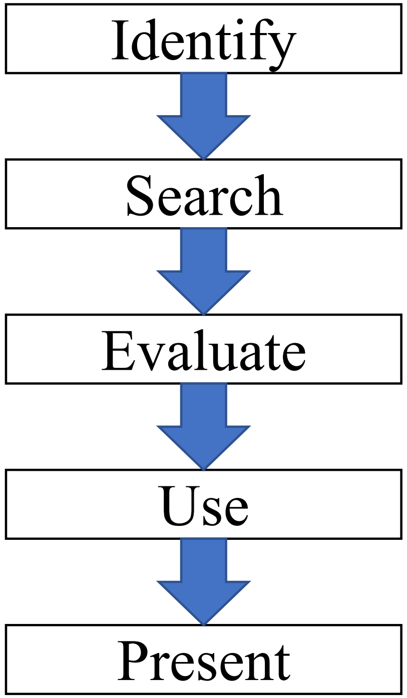
- This course is designed to train you to be able to

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).

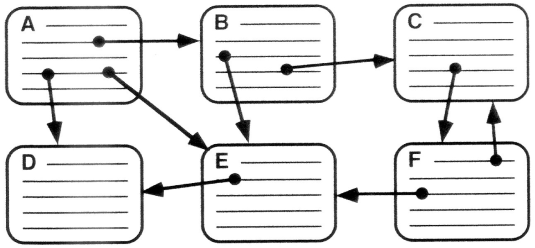

Identify
Search
Evaluate
Use
Present

Outline of this Course

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

World Wide Web

- 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.



How to Read a URL

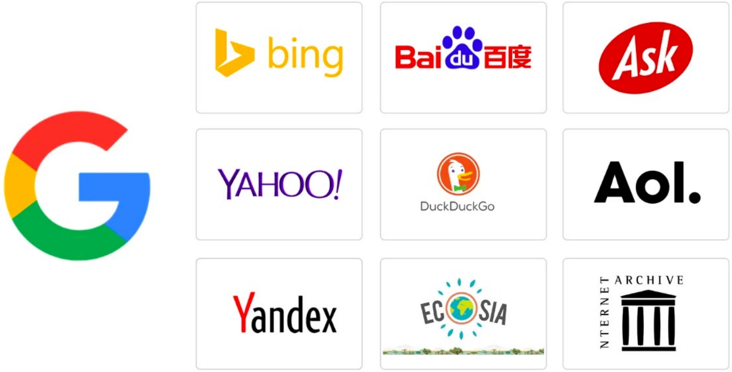
http://www.i.kyoto-u.ac.jp/en/index.html

- 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.

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

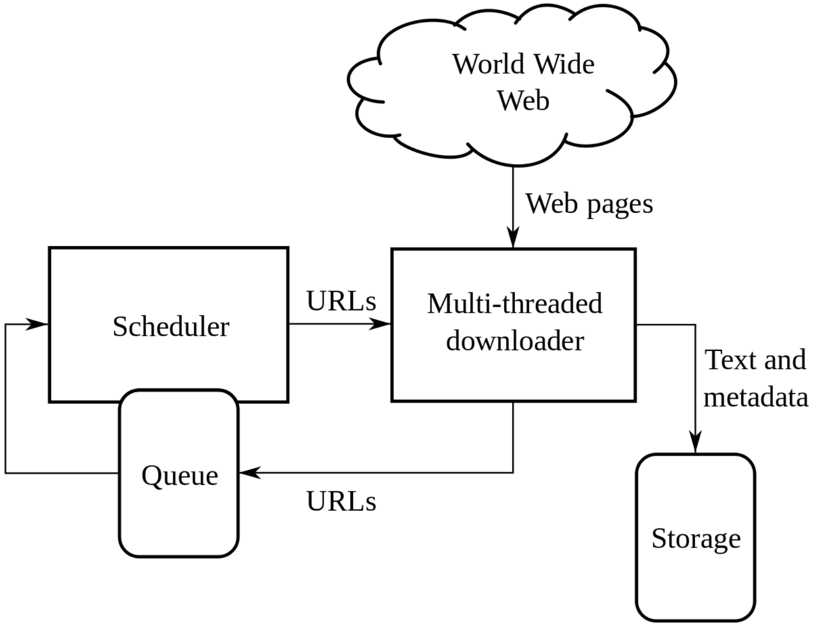
Parser

- 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

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.



Inverted Index

d_1 $\overset{1}{\text{Kyoto}}$ $\overset{2}{\text{is}}$ $\overset{3}{\text{in}}$ $\overset{4}{\text{Kansai.}}$

d_2 $\overset{1}{\text{Kyoto}}$ $\overset{2}{\text{was}}$ $\overset{3}{\text{the}}$ $\overset{4}{\text{capital}}$ $\overset{5}{\text{of}}$ $\overset{6}{\text{Japan}}$ $\overset{7}{\text{and}}$ $\overset{8}{\text{is}}$ $\overset{9}{\text{in}}$ $\overset{10}{\text{Kansai.}}$

d_3 $\overset{1}{\text{Tokyo}}$ $\overset{2}{\text{is}}$ $\overset{3}{\text{the}}$ $\overset{4}{\text{largest}}$ $\overset{5}{\text{city}}$ $\overset{6}{\text{in}}$ $\overset{7}{\text{Japan.}}$

d_4 $\overset{1}{\text{Tokyo}}$ $\overset{2}{\text{and}}$ $\overset{3}{\text{Kyoto}}$ $\overset{4}{\text{are}}$ $\overset{5}{\text{located}}$ $\overset{6}{\text{in}}$ $\overset{7}{\text{Japan.}}$ $\overset{8}{\text{Tokyo}}$ $\overset{9}{\text{is}}$ $\overset{10}{\text{bigger}}$ $\overset{11}{\text{than}}$ $\overset{12}{\text{Kyoto.}}$

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 with other students about how a search engine works
  - It can be a wrap up of what you learnt from this lecture
  - Also, please try to discuss beyond what you learnt from this lecture

Can you share your discussion of task 6?

Ranking

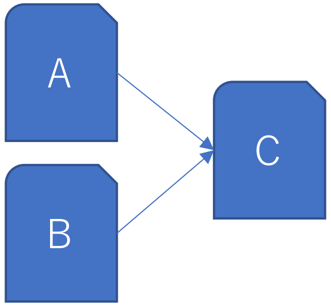
- There are so many web servers in the Internet and numerous web pages on each of them.
- It is so important for any web search engine to rank the pages with the aim of providing more useful data, by listing the pages containing the data at higher places, to the searcher about the searched keyword or subject.
- To be able to provide desired ordering for the web pages: A page ranking algorithm is the technique utilizing some valuable metrics about the web pages and ordering the pages accordingly.

PageRank Algorithm (Google)

- The “PageRank” algorithm, proposed by founders of Google Sergey Brin and Lawrance Page, is one of the most common page ranking algorithms.
- The algorithm uses the linking (citation) info occurring among the pages as the core metric in ranking procedure.
- Existence of a link from page p1 to p2 may indicate that the author is interested in page p2.
- The PageRank metric PR(p), defines the importance of page p to be the sum of the importance of the pages that point to p.

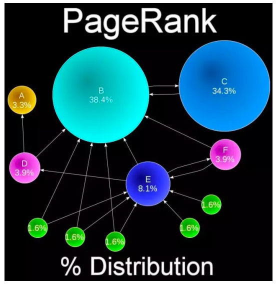
Link Structure of the Web

- 310 million web pages → 3.7 billion links
- Backlinks and Forward links:
  - A and B are C's backlinks
  - C is A and B's forward link



- Intuitively, a webpage is important if it has a lot of backlinks

PageRank Algorithm (1/3)



- Consider pages T1,…,Tn, which link to a page A, and let C(Ti) be the total number of links going out of page Ti. d is the damping factor which is set between 0 and 1. Then, PageRank of page A is given by:

$$PR(A) = (1-d) + d(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)})$$

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages, PageRanks will be 1.

PageRank Algorithm (2/3)

- Damping factor d makes sense because users will only continue clicking on links for a finite amount of time before they get distracted and start exploring something completely unrelated.
- With the remaining probability (1-d), the user will click on one of the C(Ti) links on page Ti at random.
- Damping factor is usually set to 0.85. So it is easy to infer that every page distributes 85% of its original PageRank evenly among all pages to which it points.

PageRank Algorithm (3/3)

- The PR of each page depends on the PR of the pages pointing to it.
- PR(Ti) does not influence PR(A) uniformly.
- PR(Ti) is weighted by the No. of forward link on Ti. Thus, more C(Ti), less PR(A).

$$PR(A) = (1 - d) + d(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)})$$

Problems of the PageRank Algorithm

- It is a static algorithm that, because of its cumulative scheme, popular pages tend to stay popular generally.
- Popularity of a site does not guarantee the desired information to the searcher so relevance factor also needs to be included.
- In Internet, available data is huge and the algorithm is not fast enough.
- It should support personalized search that personal specifications should be met by the search result.

Tips in Searching (1/4)

- Use quotation marks “” to locate an entire string.
  - eg. “Chenhui Chu" will only return results with that exact string.
- Mark essential words with a +
  - If a search term must contain certain words or phrases, mark it with a + symbol. eg: + “chenhui chu” conference will return all results containing “chenhui chu”.
- Negate unwanted words with a -
  - You may wish to search for the term bass, pertaining to the fish and be returned a list of music links as well. To narrow down your search, try: bass - music. This will return all results with “bass” and NOT “music”.

Tips in Searching (2/4)

- site:www.cwire.org
    - This will search only pages which reside on this domain.
- related:www.cwire.org
    - This will display all pages which Google finds to be related to your URL
- link:www.cwire.org
    - This will display a list of all pages which Google has found to be linking to your site. Useful to see how popular your site is.

Tips in Searching (3/4)

- filetype: Search for a specific file type. Try it with .mp3, .mpg or .avi if you like.
- allinurl: Google will restrict the results to those with all of the query words in the url. For instance, [allinurl: google search] will return only documents that have both "google" and "search" in the url.
- inurl: Google will restrict the results to documents containing that word in the url. For instance, [inurl:google search] will return documents that mention the word "google" in their url, and mention the word "search" anywhere in the document (url or no).
  - Note there can be no space between the "inurl:" and the following word.

Tips in Searching (4/4)

- allintitle: Google will restrict the results to those with all of the query words in the title. For instance, [allintitle: google search] will return only documents that have both “google” and “search” in the title.
- intitle: Google will restrict the results to documents containing that word in the title. For instance, [intitle:google search] will return documents that mention the word “google” in their title, and mention the word “search” anywhere in the document (title or no).
    - Note there can be no space between the “intitle:” and the following word.
- allinlinks: Searches only within links, not text or title.
- allintext: Searches only within text of pages, but not in the links or page title.

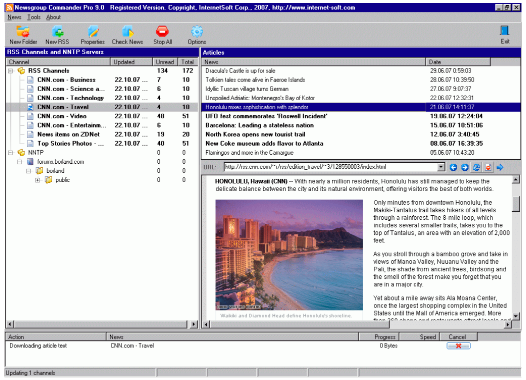
Questions

- Can search engines find everything on the Internet?
  - No. Search engines cannot index the pages in the invisible web / deep web:
  - Pages which are not linked to by other pages.
  - Dynamic Web pages based on responses to database queries.
  - Sites that require registration or otherwise limit access to their pages.

- Do search engines always put the relevant pages at the top?
  - No. Search engines put advertisers' pages at the top, called paid placement, sponsored links or sponsored listing.

Newsgroups (1/4)

- Newsgroups are discussion forums available through Internet services providers.
  - Each group is usually dedicated to a certain discussion topic. The topic is usually reflected by a unique name.
- They are not part of the WWW and users should use special software called a newsreader to access them.
  - Mozilla / Thunderbird
  - Microsoft Outlook
  - Google Groups


Newsgroup Commander Pro 9.0 Registered Version. Copyright, InternetSoft Corp., 2007, http://www.internet-soft.com
News
Tools
About
New Folder
New RSS
Properties
Check News
Stop All
Options
Exit
RSS Channels and NNTP Servers
Articles

Channel	Updated	Unread	Total
RSS Channels		134	172
CNN.com - Business	22.10.07 ...	7	10
CNN.com - Science a...	22.10.07 ...	6	10
CNN.com - Technology	22.10.07 ...	4	10
CNN.com - Travel	22.10.07 ...	4	10
CNN.com - Video	22.10.07 ...	48	51
CNN.com - Entertainm...	22.10.07 ...	6	10
News items on ZDNet	22.10.07 ...	19	20
Top Stories Photos - ...	22.10.07 ...	40	51
NNTP		0	0
forums.borland.com		0	0
borland		0	0
public		0	0

News	Date
Dracula's Castle is up for sale	29.06.07 0:59:03
Tolkien tales come alive in Faeroe Islands	28.06.07 10:39:50
Idyllic Tuscan village turns German	27.06.07 9:07:37
Unspoiled Adriatic: Montenegro's Bay of Kotor	22.06.07 12:32:31
Honolulu mixes sophistication with splendor	21.06.07 14:11:37
UFO fest commemorates 'Roswell Incident'	**19.06.07 12:24:04**
Barcelona: Leading a stateless nation	**15.06.07 10:51:06**
North Korea opens new tourist trail	**12.06.07 3:40:45**
New Coke museum adds flavor to Atlanta	**08.06.07 16:39:35**
Flamingos and more in the Camargue	05.06.07 10:43:20

URL: http://rss.cnn.com/~r/rss/edition_travel/~3/128550003/index.html

HONOLULU, Hawaii (CNN) -- With nearly a million residents, Honolulu has still managed to keep the delicate balance between the city and its natural environment, offering visitors the best of both worlds.

OAHU VISITORS BUREAU

Waikiki and Diamond Head define Honolulu's shoreline.

Only minutes from downtown Honolulu, the Makiki-Tantalus trail takes hikers of all levels through a rainforest. The 8-mile loop, which includes several smaller trails, takes you to the top of Tantalus, an area with an elevation of 2,000 feet.

As you stroll through a bamboo grove and take in views of Manoa Valley, Nuuanu Valley and the Pali, the shade from ancient trees, birdsong and the smell of the forest make you forget that you are in a major city.

Yet about a mile away sits Ala Moana Center, once the largest shopping complex in the United States until the Mall of America emerged. More

Action	News	Progress	Speed	Cancel
Downloading article text	CNN.com - Travel	0 Bytes		☒

Updating 1 channels

Newsgroups (2/4)

- The newsgroups are usually provided by Internet service providers (ISP). The administrator determines what newsgroups its users can access, and how long the discussion messages are kept.
- Users use a newsreader to read or post discussion messages. Posts are grouped into threads. A thread is just the chain of posts that are discussing the same subject.
- Post often are permanently recorded on the Internet, since the posts are constantly being monitored and archived by some security or commercial organizations.

Newsgroups (3/4)

- comp.*
  - comp.games.development.design: Discuss computer games design for game developers
  - comp.lang.c++: Discuss the programming language C++
- news.*
  - news.admin.net-abuse.misc: Discuss abuses on the network systems
  - news.groups.questions: Discuss questions about newsgroups formation
- rec.*
  - rec.music.beatles.moderated: Discussion the Beatles, with moderator
  - rec.pets.cats: Discussion raising cats as pets
- sci.*
  - sci.physics.relativity: Discuss the theory of relativity
  - sci.engr.chem.: Discuss chemical engineering

Newsgroups (4/4)

- soc.*
  - soc.culture.japan: Discuss Japanese society and culture
  - soc.support.depression.family: Support group for family with depressed member
- talk.*
  - talk.politics.european-union: Discuss the politics of the EU
  - talk.religion.newage: Discuss new age religions
- misc.*
  - misc.education.medical: Discuss medical education
  - misc.consumers.house: Discuss consumer issues regarding home buying

Searching the Newsgroups (1/2)

- The newsgroups contain the following kinds of information that may be difficult to find elsewhere:

- Personal opinion and other informal information
- Very specific but unpopular topics (e.g. the solution to a rare bug in Microsoft Word, which may not have been formally documented)
- Very current topics about which web sites are not yet available

Searching the Newsgroups (2/2)

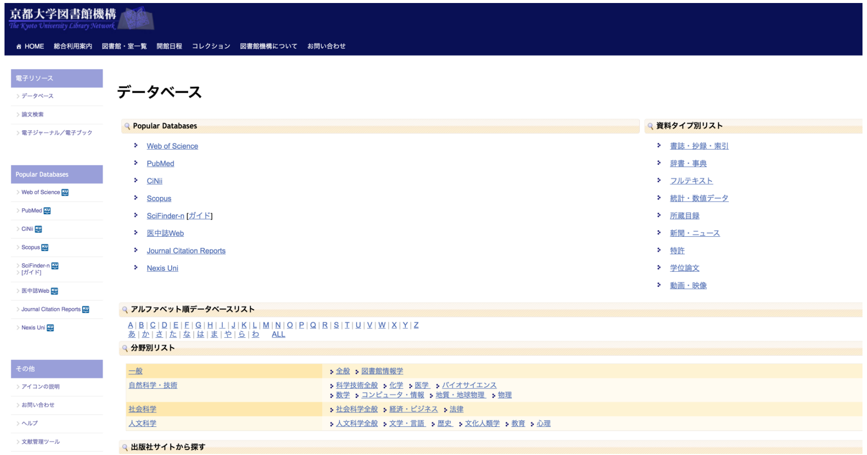
- Use Google Groups to search through various newsgroups: http://groups.google.com
  - When reading a post in a newsgroup, please pay special attention to the date of a post and the group in which the post belongs. These help you evaluate the usefulness of the post.
- Virtually any one can view and post in a newsgroup. This leads to the following problems:
  - Many spam (junk) posts (e.g. advertisements)
  - Privacy: others may capture email addresses from the posts
  - Many uninformative or inaccurate posts

Subject Directories

- A directory on the World Wide Web that specializes in linking to other web sites and categorizing those links.
- All linked pages are classified and reviewed by human. Some directories also provide evaluation by human expert to the quality of the linked web pages.
  - Yahoo: http://dir.yahoo.com/
  - Open Directory Project: http://www.dmoz.org
  - Infomine: http://www.infomine.com/
- Many people do not make enough use of the subject directories. Instead, they go directly to search engines.
- Keep in mind that subject directories often contain carefully chosen lists of quality Internet sites. They are sometimes more useful than a search engine.

Invisible Web / Deep Web

- There is a huge amount of information that is stored in databases accessible on the Web, but not available via search engines.
- It is likely to contain very current, dynamically changing information, including news, job listings, airline flights, etc.
- Please well use the Online Electronic Databases at Kyoto University Online Library



データベース

Popular Databases

- Web of Science
- PubMed
- CiNii
- Scopus
- SciFinder-n [ガイド]
- 医中誌Web
- Journal Citation Reports
- Nexis Uni

資料タイプ別リスト

- 書誌・抄録・索引
- 辞書・事典
- フルテキスト
- 統計・数値データ
- 所蔵目録
- 新聞・ニュース
- 特許
- 学位論文
- 動画・映像

アルファベット順データベースリスト

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z
あ | か | さ | た | な | は | ま | や | ら | わ ALL

分野別リスト

一般	全般 / 図書館情報学
自然科学・技術	科学技術全般 / 化学 / 医学 / バイオサイエンス / 数学 / コンピュータ・情報 / 地質・地球物理 / 物理
社会科学	社会科学全般 / 経済・ビジネス / 法律
人文科学	人文科学全般 / 文学・言語 / 歴史 / 文化人類学 / 教育 / 心理

出版社サイトから探す

WWW Search Strategies

- Think before you search!

- The general strategies of a search:
  - 1. Pre-search analysis
  - 2. Executing the search
  - 3. Looking for an overview
  - 4. Seeking expert advice

Pre-search Analysis

- Identify any societies, organizations that have the information you sought at their websites
  - E.g. To get a list of Departments at Kyodai, it is faster use visit the Kyodai's official website, instead of using a search engine.
- Identify any distinctive words, phrases, acronyms associated with the topic.
- Identify other words that are likely to appear in any web pages over the topic.
- Identify any synonyms, variations in spelling for the previously identified words or phrases, the OR search is useful here.
- Identify any irrelevant documents these search words/phrases may pick up. What are the words that distinguish these irrelevant documents? The NOT search is useful here.

Executing the Search

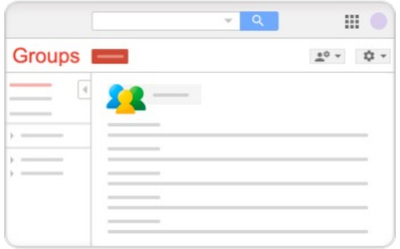
- Use your prepared search terms in a search engine.
- If your search returns too many matches:
  - Add more AND terms to pinpoint your area of interest.
  - Add more NOT terms to eliminate irrelevant matches.
  - Some search engines allow limiting matches by specifying date of publication, languages, web page URL, etc.
- If your search returns too little matches:
  - Try reducing the number of AND terms to broaden your area of focus
  - Try adding variants of your search terms with OR.
  - Use a broader term describing the subject matter in the search

Looking for an Overview

- Search in a subject directory using the broader subject term
- Search in a subject directory using your narrower keywords
  - Links from the subject category should point you to the main sites about the subject
- Lookup the subject matter in an encyclopedia (e.g. Wikipedia)

Seek Expert Advice

- Seek advice from relevant mailing lists, newsgroups, or other discussion groups.
- Sometimes you come across experts in those forums who can point you to find articles or resources for problem-shooting.
- Search Google Groups if you think the topic may have been previously discussed in newsgroups.



Task 7

- Discuss the merits and demerits of the PageRank Algorithm
  - 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