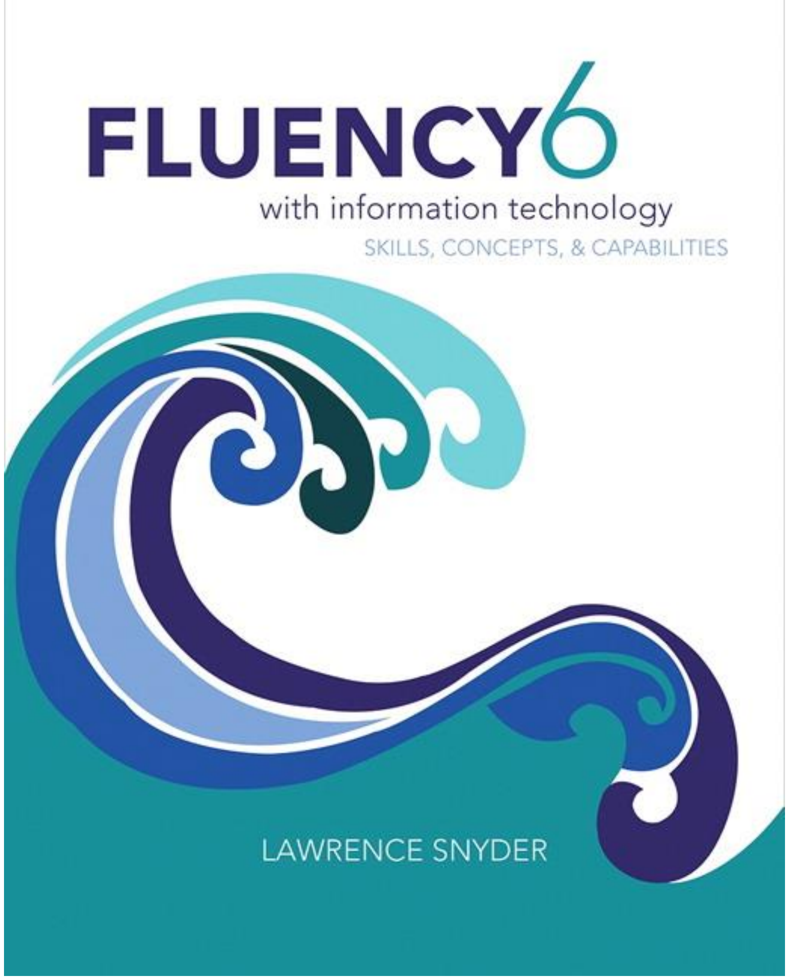


Chapter 5

Locating Information on the WWW

Table of Contents

- Part 1: Becoming Skilled at Computing
  - Chapter 1: Defining Information Technology
  - Chapter 2: Exploring the Human-Computer Interface
  - Chapter 3: The Basics of Networking
  - Chapter 4: A Hypertext Markup Language Primer
  - Chapter 5: Locating Information on the WWW
  - Chapter 6: An Introduction to Debugging
- Part 2: Algorithms and Digitizing Information
- Part 3: Data and Information
- Part 4: Problem Solving

Learning Objectives

- Explain how a Web search engine works
- Find information by using a search engine
- Use logical operators, filtering to express complex queries
- Recognize and find authoritative information sources
- Decide whether Web information is truth or fiction

Web Search Fundamentals

- A search engine is a collection of computer programs that help us find information on the Web
- No one organizes the information posted on the Web
- Search Engines must look around to find out what’s out there and then organize what is found
  - Need to perform indexing most words in the WWW in the world

How a Search Engine Works [1/2]

- The 1st step, crawling, visits every Web page that it can find
- The main work of the crawler is to build an index which is a list of tokens (or words) that are associated with the page

  (token1, a list of URLs)

  (token2, a list of URLs)

  …..

  (token_n, a list of URLs)

- Crawling Process
  - The crawler has a to-do list that is loaded with a set of pages to start
  - When a new URL is noticed while crawling a page, it adds that URL to the to-do list
  - For each token, the crawler creates a list of the URLs associated with that token by inspecting all pages of URLs in the to-do list

How a Search Engine Works [2/2]

- The 2nd step is query processing
- The user presents tokens (aka search terms) to the query processor
- The search engine then looks up the word in the index and returns a hit list which is a set of URLs

  (token1, a list of URLs)

  (token2, a list of URLs)

  …..

  (token_n, a list of URLs)

- By creating the index ahead of time, search engines are able to answer user queries very quickly

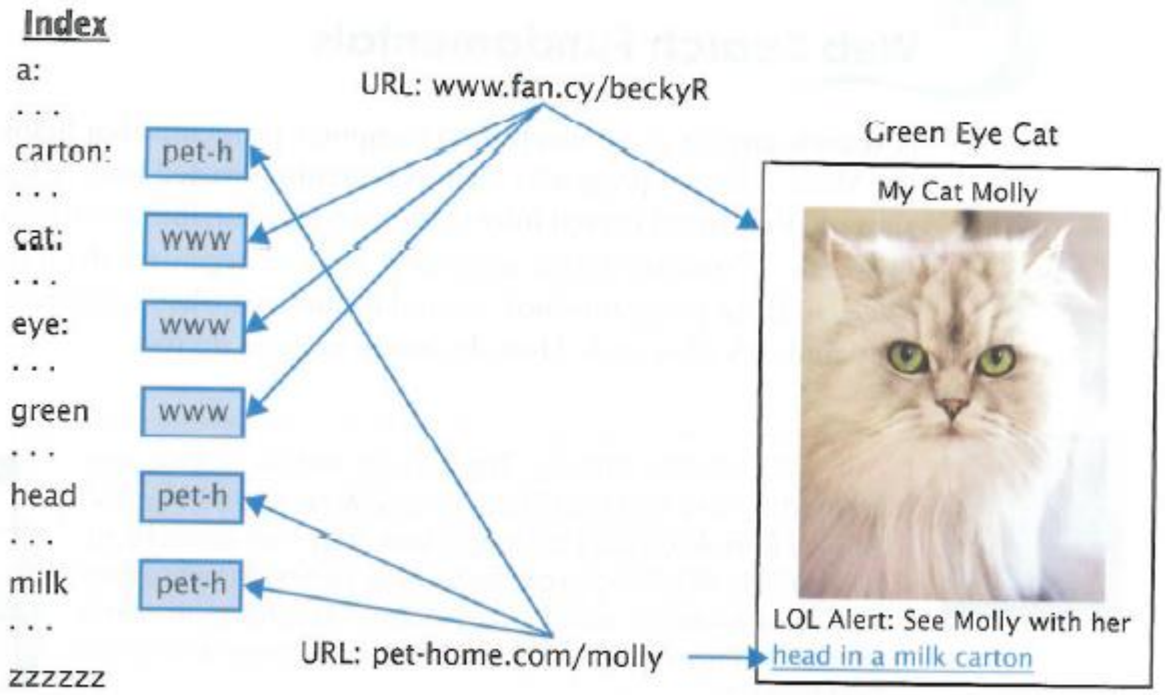
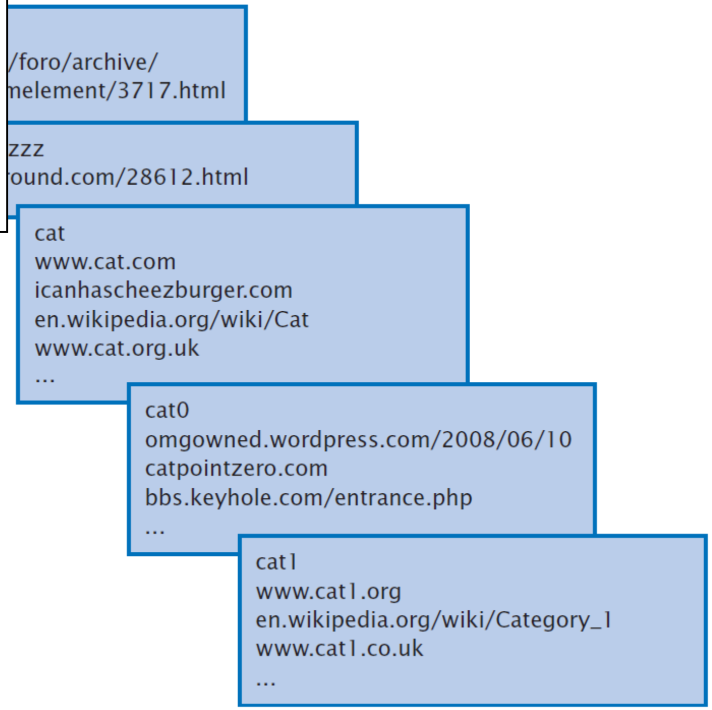


Figure 5.1 Crawling over the Green Eye Cat page: The crawler adds the page's URL to the lists for each word in its title; for words in the anchor text, the link URL is added to their lists.



/foro/archive/
melement/3717.html

zzz
ound.com/28612.html

cat
www.cat.com
icanhascheezburger.com
en.wikipedia.org/wiki/Cat
www.cat.org.uk
...

cat0
omgowned.wordpress.com/2008/06/10
catpointzero.com
bbs.keyhole.com/entrance.php
...

cat1
www.cat1.org
en.wikipedia.org/wiki/Category_1
www.cat1.co.uk
...

Sample index entry lists for tokens around "cat" produced by a Web crawler. Some lists are tiny (caszzzzzzz has one entry) and others are very long—there are more than 2.05 million URLs following cat.

보조자료

Large Scale Web Search

- Google server cluster
  - "less than $1,000" server for Error isolation, Easy to repair, Easy to scale
  - 450,000 servers (NYT estimate, Oct, 2006)
  - 900,000 servers (2011)
  - Maybe more than 1 million servers now (2015)!

- Google's search index
  - Indexing most words in the WWW in the world
  - 100 million Giga bytes = 10^{17} bytes
  - Index Structure

```
potato: (url_ZZ; 3, 101, 178, 2009); (url_pq; 1; 809); …
quake:  (url_ds; 1; 16);  (url_lk; 4; 3, 11, 12, 678); …
```

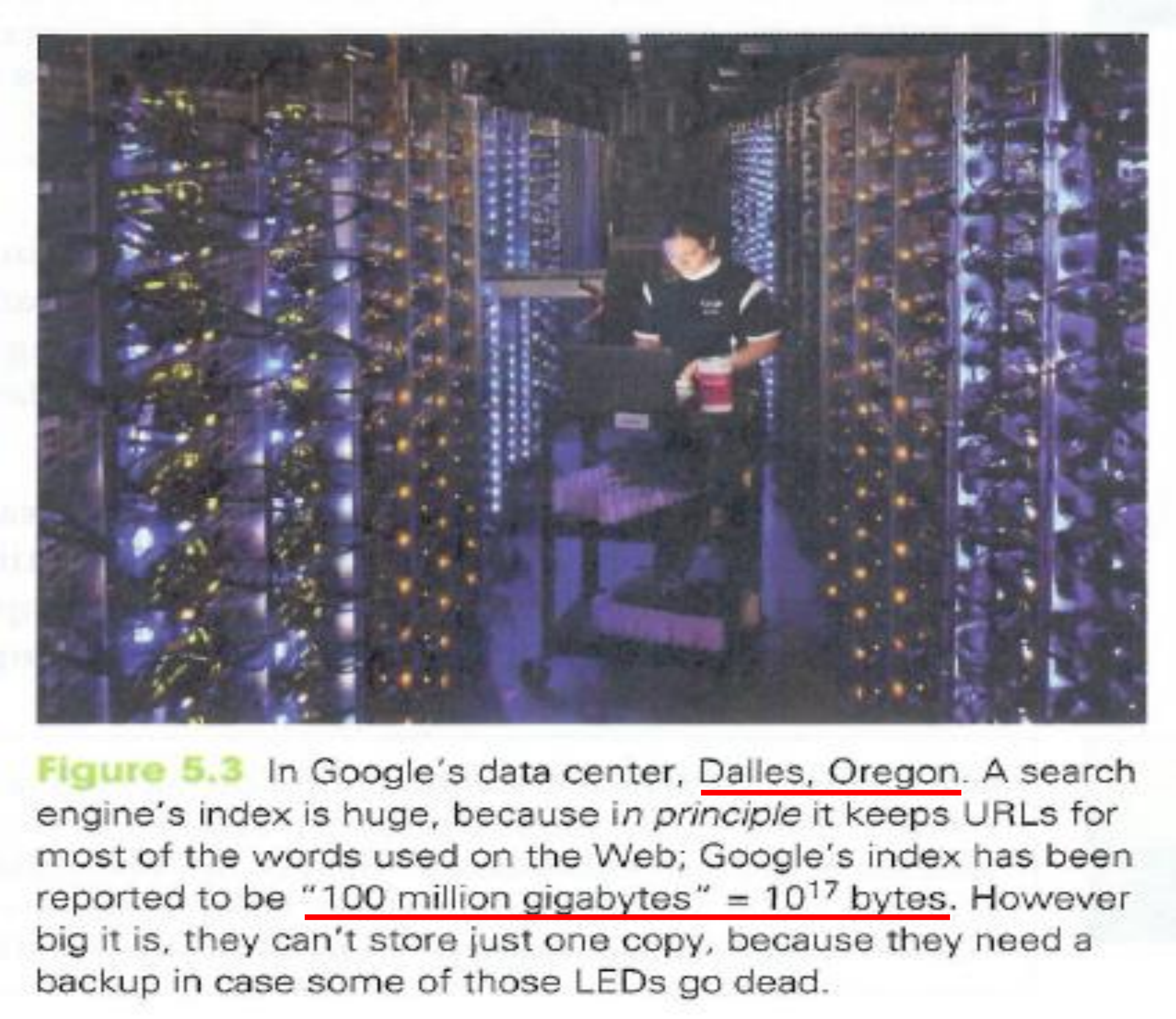
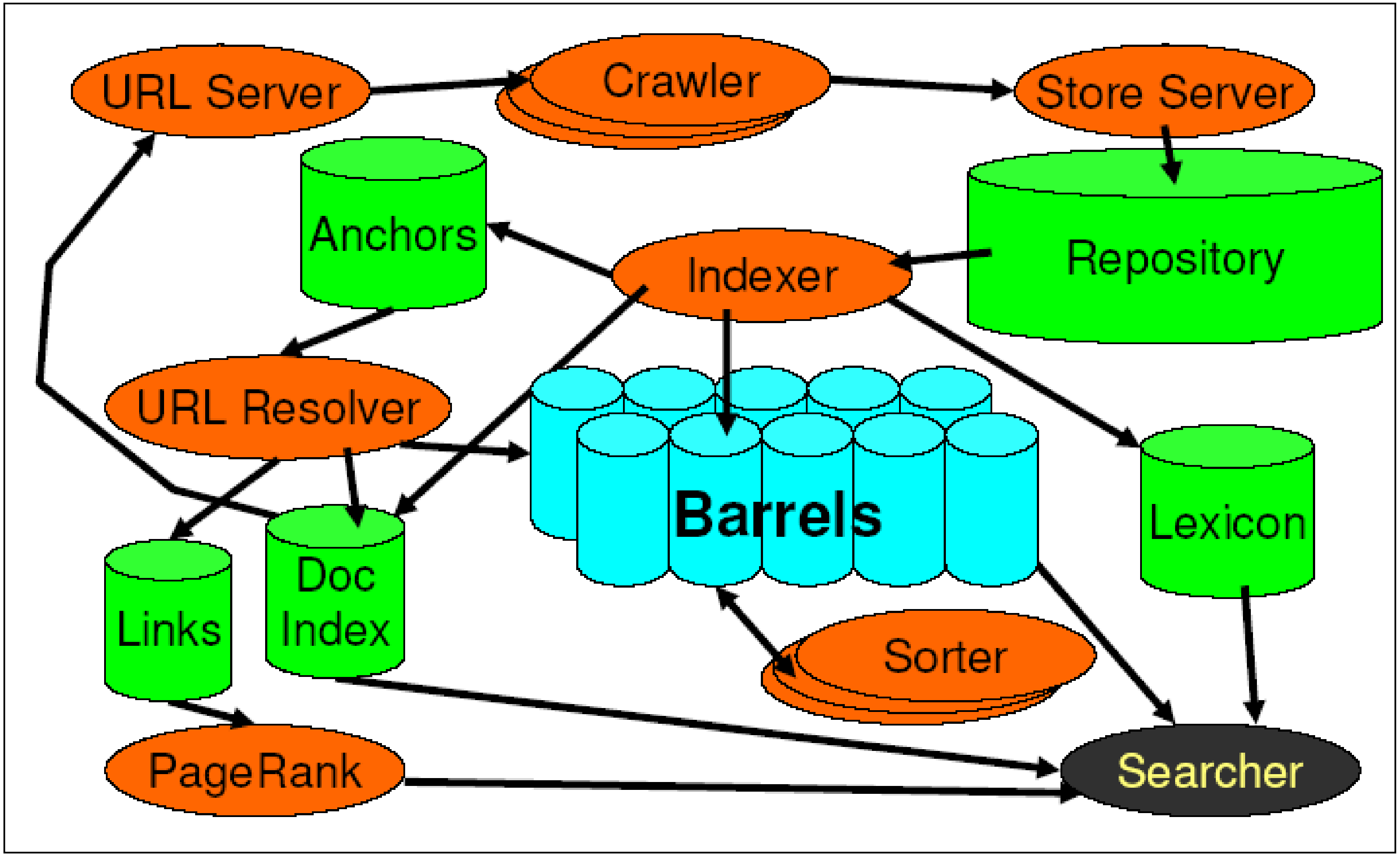


Figure 5.3 In Google's data center, Dalles, Oregon. A search engine's index is huge, because i*n principle* it keeps URLs for most of the words used on the Web; Google's index has been reported to be "100 million gigabytes" = 10^{17} bytes. However big it is, they can't store just one copy, because they need a backup in case some of those LEDs go dead.

보조자료

Large Scale Web Search

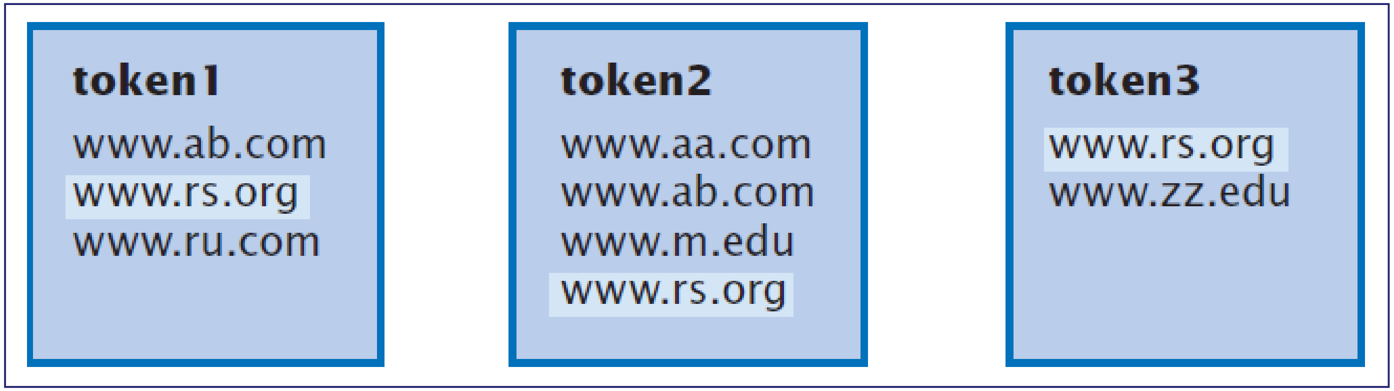
**** Google Search Engine Architecture**



** Google Map-Reduce Framework! ➔ Big Data Processing

Multiword Searches

- Example: human powered flight
- With a multiple-word query, the pages returned should be appropriate for all of the queried words
- AND-query of token1, token2, token3
  - Each page returned should be associated with all the words
  - There is no index entry corresponding to a set, only a list for the individual words



token1
www.ab.com
www.rs.org
www.ru.com

token2
www.aa.com
www.ab.com
www.m.edu
www.rs.org

token3
www.rs.org
www.zz.edu

Intersecting Queries

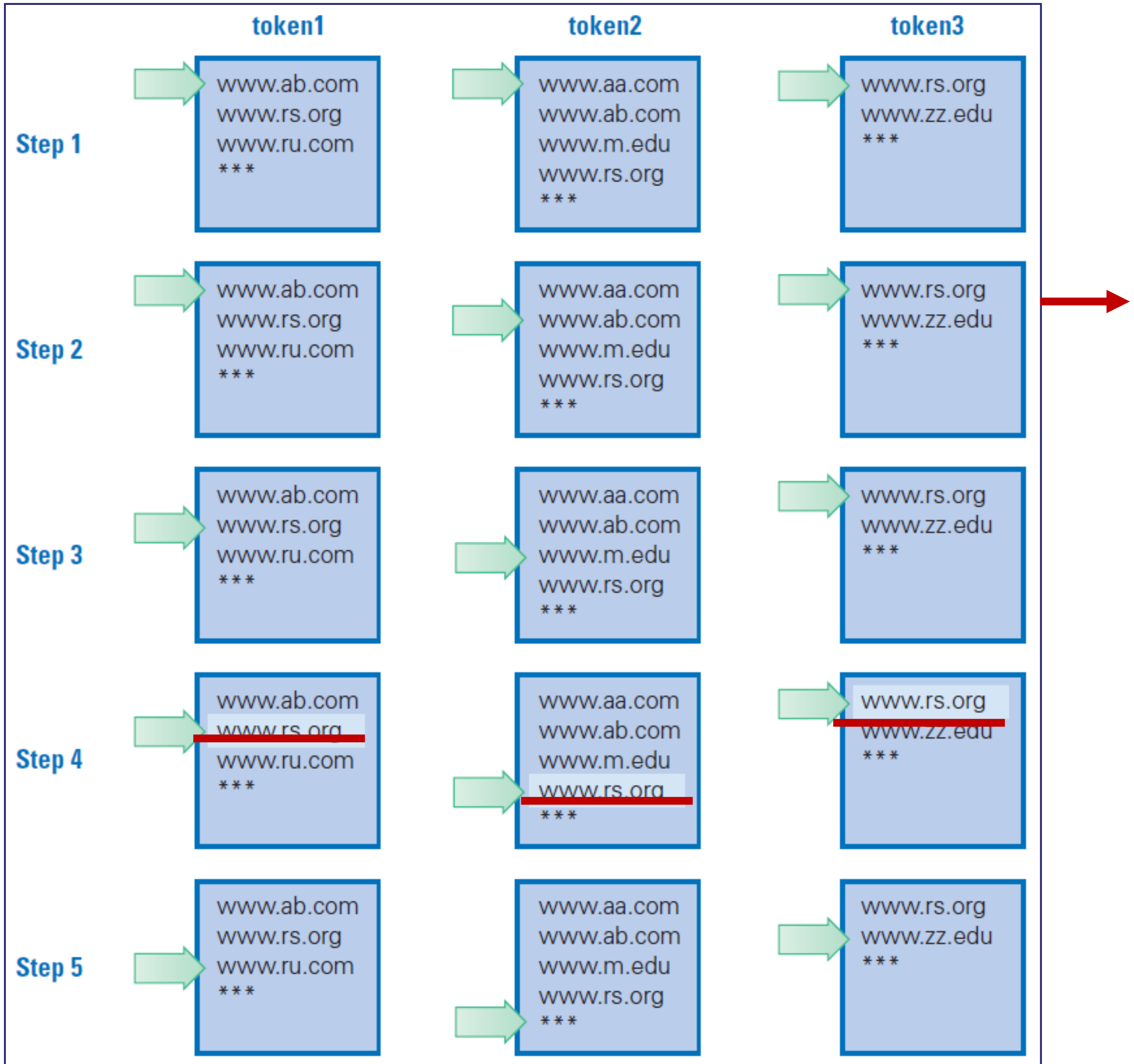
- For multiple words, the query processor fetches the index lists for each of the terms
- The query processor intersects the lists
- The URL lists are alphabetized (sorted) to speed up the processing … it is easier to notice when the same URL is on multiple lists
- To intersect several alphabetized lists:

  (step1) Put a marker (arrow) at the start of each token’s index list

  (step2) If all markers point to the same URL, save it, because all tokens are associated with the page

  (step3) Move the marker(s) to the next position for whichever URL is earliest in the alphabet

  (step 4) Repeat Steps 2–3 until some marker reaches the end of the list

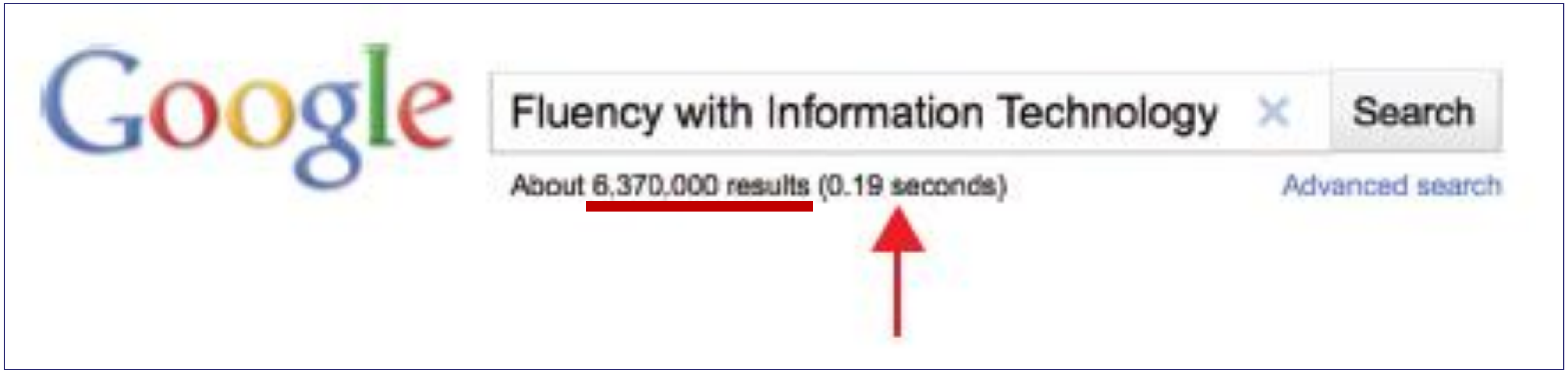


	token1	token2	token3
Step 1	www.ab.com www.rs.org www.ru.com ***	www.aa.com www.ab.com www.m.edu www.rs.org ***	www.rs.org www.zz.edu ***
Step 2	www.ab.com www.rs.org www.ru.com ***	www.aa.com www.ab.com www.m.edu www.rs.org ***	www.rs.org www.zz.edu ***
Step 3	www.ab.com www.rs.org www.ru.com ***	www.aa.com www.ab.com www.m.edu www.rs.org ***	www.rs.org www.zz.edu ***
Step 4	www.ab.com www.rs.org www.ru.com ***	www.aa.com www.ab.com www.m.edu www.rs.org ***	www.rs.org www.zz.edu ***
Step 5	www.ab.com www.rs.org www.ru.com ***	www.aa.com www.ab.com www.m.edu www.rs.org ***	www.rs.org www.zz.edu ***

Figure 5.2 Illustrating the Intersecting Alphabetized Lists rules: In each step (row of boxes) one or more arrows advance; notice Step 3 where the arrow has advanced in two lists in the same step because the earliest URL is on both lists.

Power of an Indexed Search

- The computer:
  - takes the time to crawl the data (Web pages)
  - build an index first
  - find the index entries for each word
  - intersect the lists to find the information for an AND-query
- Search engines can look at billions of Web pages and return an answer in less than a fifth of a second (< 0.2 sec)



Fluency with Information Technology

About 6,370,000 results (0.19 seconds)

Descriptive Terms

- “Hits” on a page means the search term is “associated” with the page
- This does not mean that the word is “on” the page
- Various HTML Tags and attributes help a lot to identify descriptive text
  - Title tag: The <title> encloses a short phrase describing the whole page
  - Anchor text tag: The highlighted link text, inside <a . . . > tags, describes the page it links to
  - Meta tag: A <meta name=“description” content=‘travel photos: volcanoes of Hawaii > tag in the head section can hold a several sentence description of the page
  - Top-level heading tag: <h1> often give a general description of a section
  - Alt attributes: The <img src=“volcano.jpg” alt=“Lava on Hawaii’s Kilauea Volcano> tag has an alt attribute that gives a textual description

Page Rank

- The hit list may include thousands of web pages
- Why, when the hit list is returned, the page you’re looking for is often first on the hit list or in the top 10?
- The order in which hits are returned to a query is determined by a number called the PageRank (by Larry Page and Sergey Brin)
- The higher the PageRank, the closer to the top of the list
- The influence factors for PageRank
  - Number of Incoming Links
  - Number of Outgoing Links
  - Number of direct typing of URL
  - Timeliness

Links to Other Pages [1/2]

- Google pioneered page ranking as a way to determine which pages are likely to be most important
- PageRanking works like a voting system:
  - If page A links to page B, A’s link adds to B’s importance
- Pages that are linked-to by many pages have a higher page ranking and are assumed to be more important!
- Links from pages with a high page ranking are also viewed as more important than links from pages with a low page ranking

Links to Other Pages [2/2]

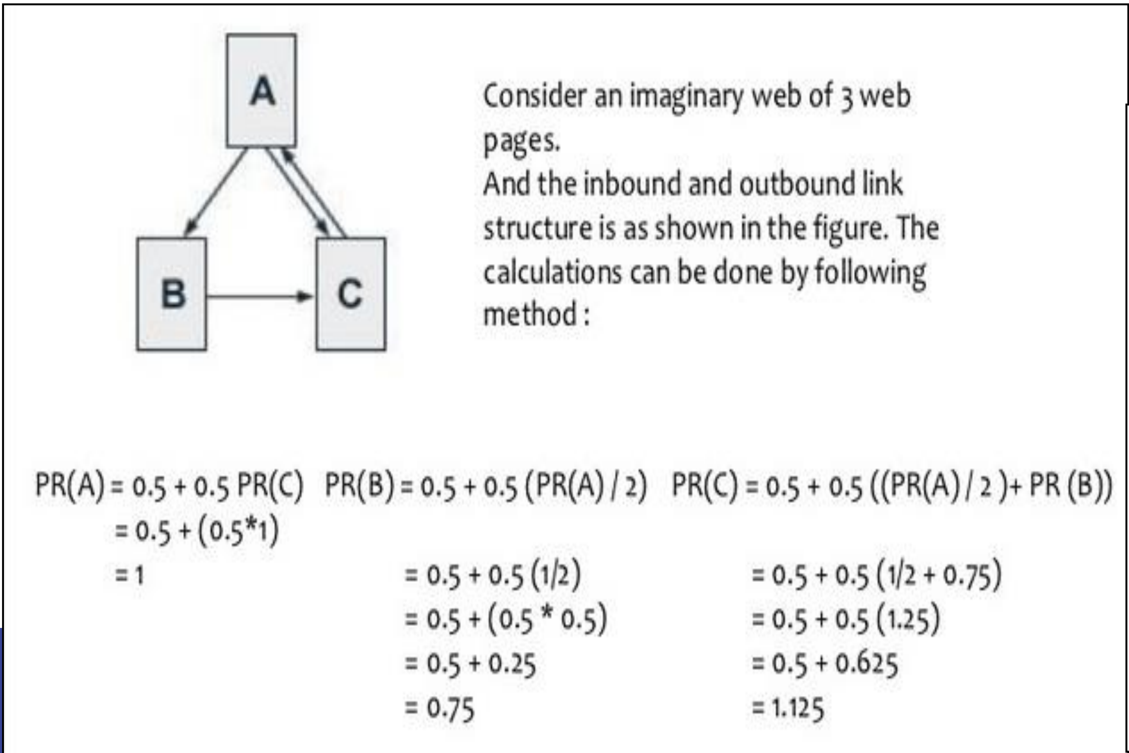
- PageRank is computed by the crawler:
    - The crawler looks at page A
    - It notices the links to page B
    - scores one for B
- Counting the number of links to a page is not sufficient: the current pagerank score is distributed to outgoing links
- So, after the crawling is completed, the PageRank computation is finally completed
- The query processor puts together the hit list
- The URLs in the hit list are sorted by their page ranking, highest to lowest, and returned to the user in that order

PageRank 공식: PR (A) = (1-d) + d (PR (T1) / C (T1) + ... + PR (Tn) / C (Tn))

참고자료

- d: damping factor (링크를 타고 페이지들을 이동할 확률)
- 1-d: 랜덤하게 이동할 확률
- C(T1): node T1에서 나가는 edge의 카운트
- 예시에서 d는 0.5를 보통은 d를 0.85정도로 사용합니다.

보통 PageRank 초기값은 1을 많이 사용하고,
특정 ε를 주어서 무한정 반복되는 상황을 방지

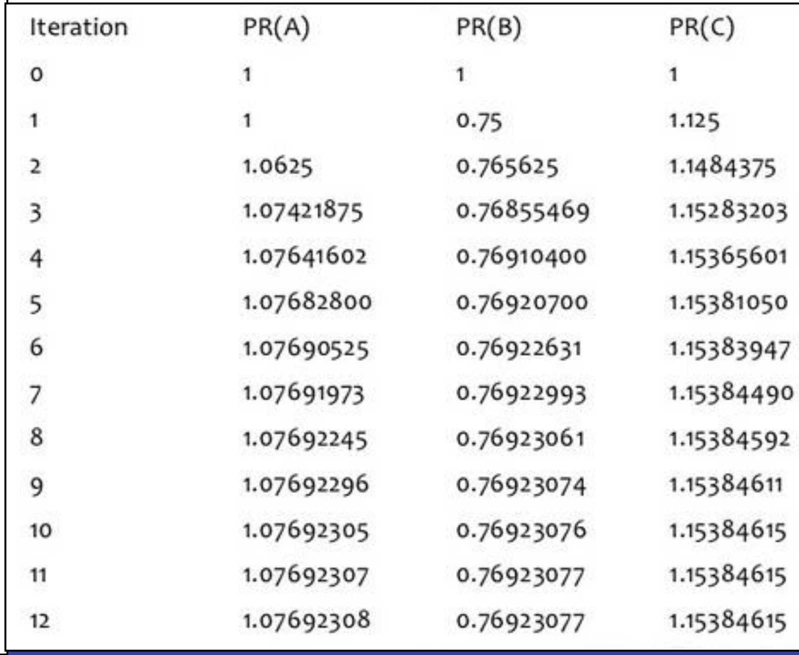


Consider an imaginary web of 3 web pages.
And the inbound and outbound link structure is as shown in the figure. The calculations can be done by following method :

$$PR(A) = 0.5 + 0.5\,PR(C) = 0.5 + (0.5*1) = 1$$

$$PR(B) = 0.5 + 0.5\,(PR(A)/2) = 0.5 + 0.5\,(1/2) = 0.5 + (0.5*0.5) = 0.5 + 0.25 = 0.75$$

$$PR(C) = 0.5 + 0.5\,((PR(A)/2) + PR(B)) = 0.5 + 0.5\,(1/2 + 0.75) = 0.5 + 0.5\,(1.25) = 0.5 + 0.625 = 1.125$$

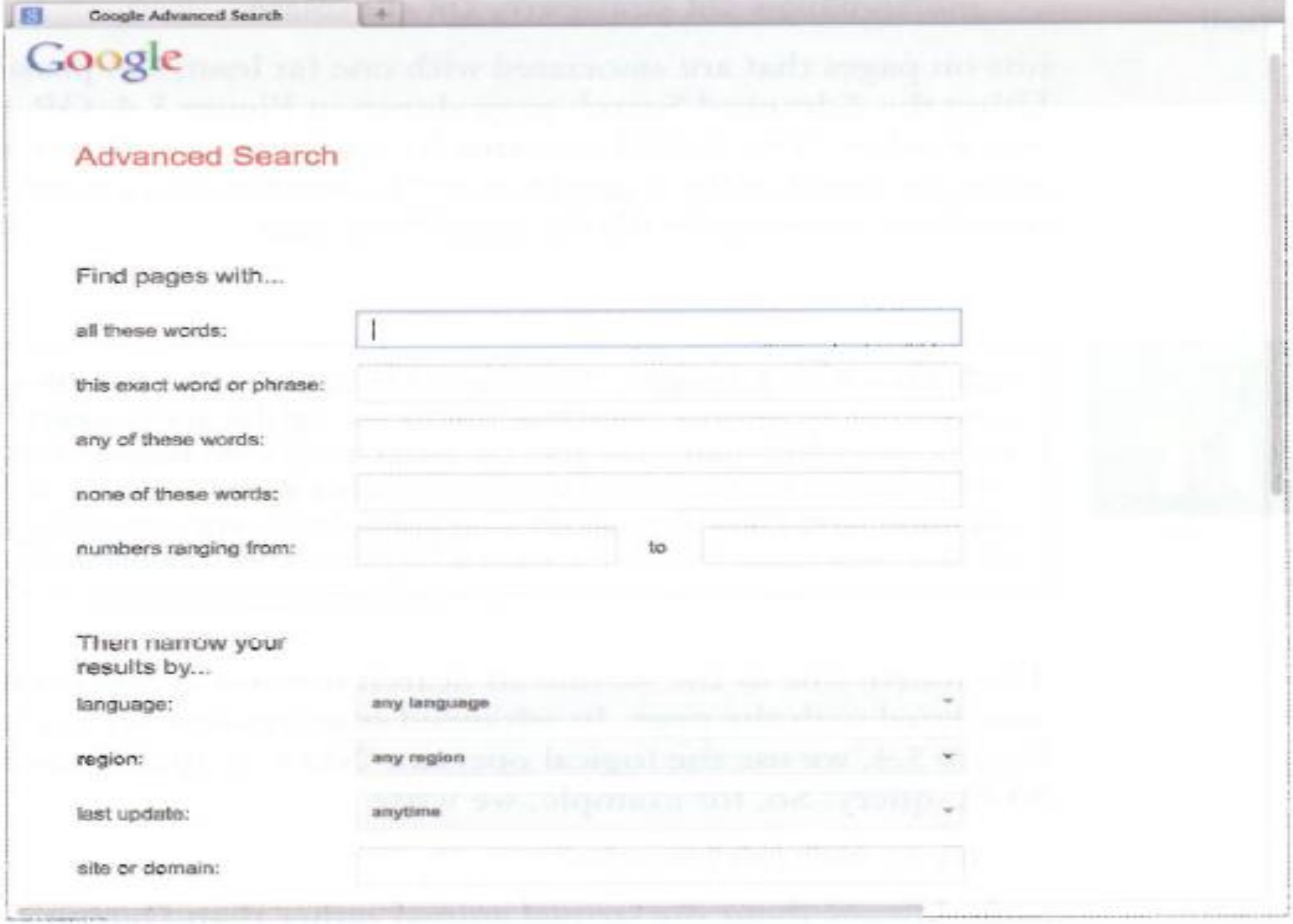


Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

Advanced Searches [1/2]

Logical operators (AND, OR, NOT) specify a logical relationship between the words it connects

- The Logical Operator AND
- Ex: human AND powered AND flight
  - Basic queries are AND-queries (Ex: human powered flight)
  - All words given must be associated with the page for a hit
  - Search engines treat search words as 3 independent words
  - The words can appear anywhere on the page in any order
  - Use of quotes mean the exact phrase must appear as given…this is more than an AND query Ex: “human powered flight”



Google Advanced Search

Google

Advanced Search

Find pages with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from: to

Then narrow your results by...

language: any language

region: any region

last update: anytime

site or domain:

Figure 5.4 Google's Advanced Search window. Notice that text panes are provided for AND-words, quote phrases, OR-words, and NOT-words; the combined query is in the text window surrounded by blue.

Advaned Searches [2/2]

- Another logical operator is OR
  - Ex: marshmallow OR strawberry OR chocolate
  - OR-queries hit on pages that are associated with at least one of the words

- Another logical operator is NOT
  - Ex: tigers AND NOT baseball
  - NOT queries specify words that are not to be associated with the page
  - AND is included because we want both requirements to be true
  - Google uses a minus (–) as an abbreviation for NOT

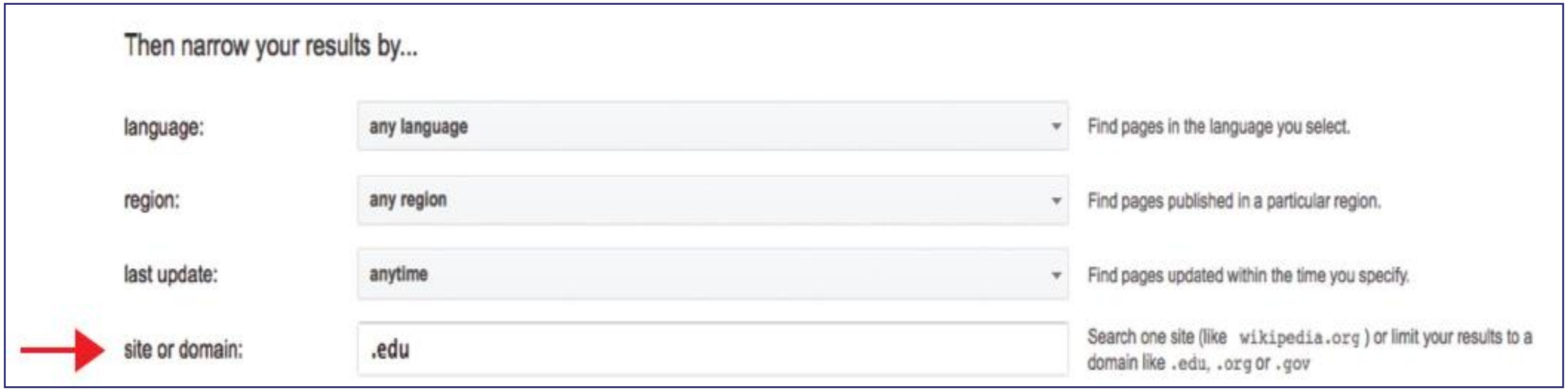
- The logical operators work like arithmetic
  - can be combined and grouped using parentheses
  - EX: (marshmallow OR strawberry) AND sundae

Restricting Global Search

- Many sites offer the opportunity to perform a site search
  - A site search means looking only on the current site
- The site search is usually offered on the homepage with a search window and a Go button

Filtered Searches

- Constraints in the filter can be used to help pinpoint specific pages
- Site searches don't necessarily use PageRanking to order the hits



Then narrow your results by...

language:	any language	Find pages in the language you select.
region:	any region	Find pages published in a particular region.
last update:	anytime	Find pages updated within the time you specify.
site or domain:	.edu	Search one site (like wikipedia.org) or limit your results to a domain like .edu, .org or .gov

Research Activity using Web Searching

- In general we are good at web surfing.. Everyday 1hr? 5hrs?

- What if your are preparing a report in which the information must be correct?

- Research on the web requires serious consideration and strategies:
  - Selecting Search Terms: choosing good words to include in a query
  - The Anatomy of a Hit: how to use the information returned
  - Using the Hit List: skimming to find what you want quickly
  - Once You Find a Likely Page: locating the desired data on the page

Skim: (기름기를) 거둬내다, (스치듯) 대충지나가다

1. Selecting Search Terms [1/4]

- How do you find the best search terms?
- It is a sequence of finding ever-more precise terms:
  - Using the Advanced Search tool in the web search engine
  - Begin with General Topic
  - Choose Descriptive Terms
  - Refine by Adding Words
  - Avoid Over Constraining
  - Remove Specific Words

1. Selecting Search Terms [2/4]

- Use Advanced Search
  - Google's Advanced Search gives control over the results returned
  - You can do complex queries, but Advanced Search provides control

- Begin with the General Topic
  - Words have multiple meanings
  - Giving the topic can eliminate most of those conflicting words
  - Many hits can be eliminated
  - Start with the topic word(s)

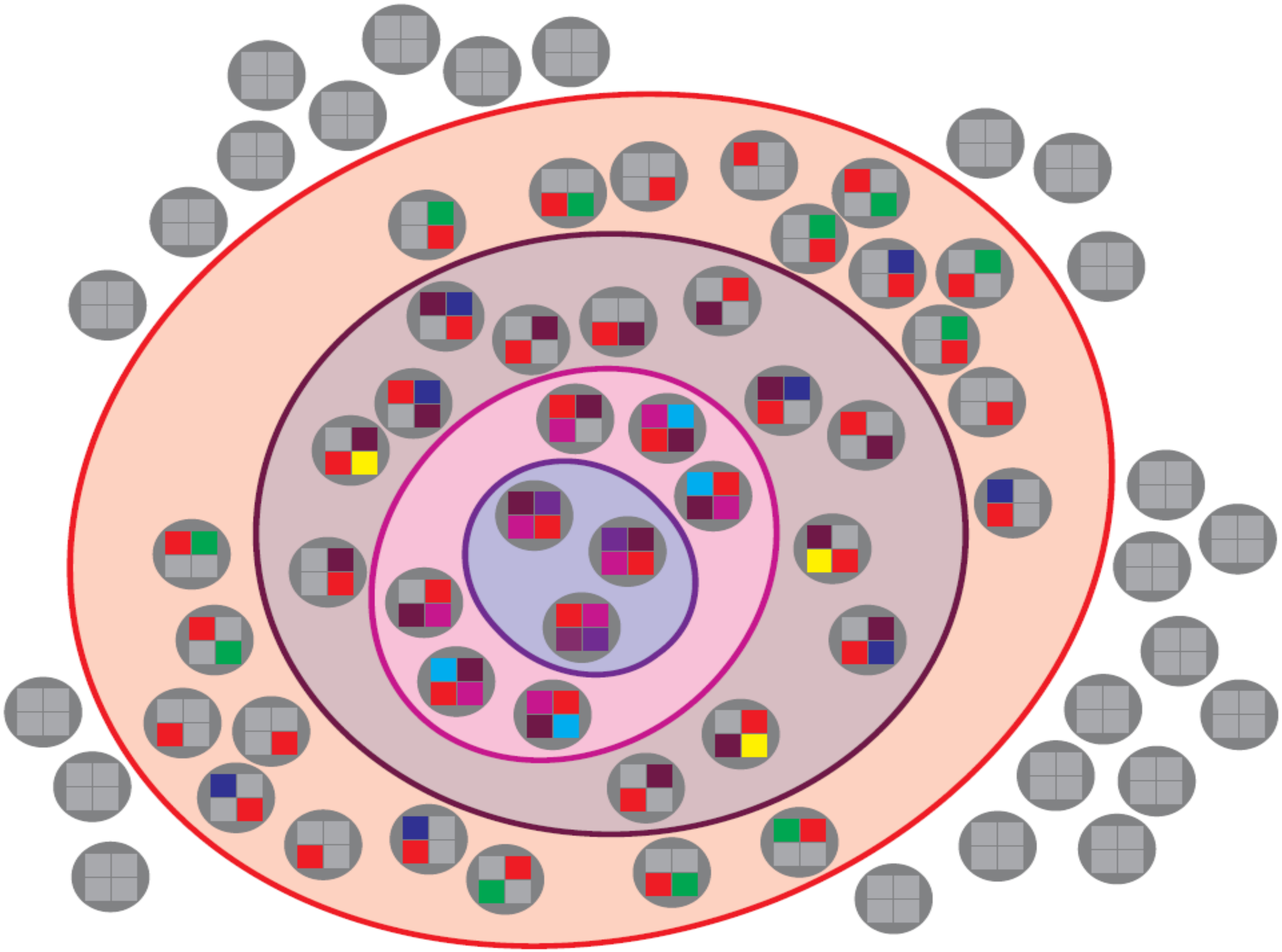
1. Selecting Search Terms [3/4]

- Choose Descriptive Words
  - Be picky about the words we select
  - Select precise terms if possible
  - Take care using terms having many other meanings - less useful!
- Refine by Adding Words
  - Begin with a “first guess” search
  - Check the results
  - Check the initial hit list to see what you’ve found
  - The initial list often suggests additional terms to search
  - This may require several rounds of adding a word to the query each time

1. Selecting Search Terms [4/4]

- Avoid Over-Constraining
  - – Adding more words one at a time works best because interesting pages might be accidentally overlooked
  - – Care is required in this process:
    - •Add a word only if you are sure the pages you want will have it

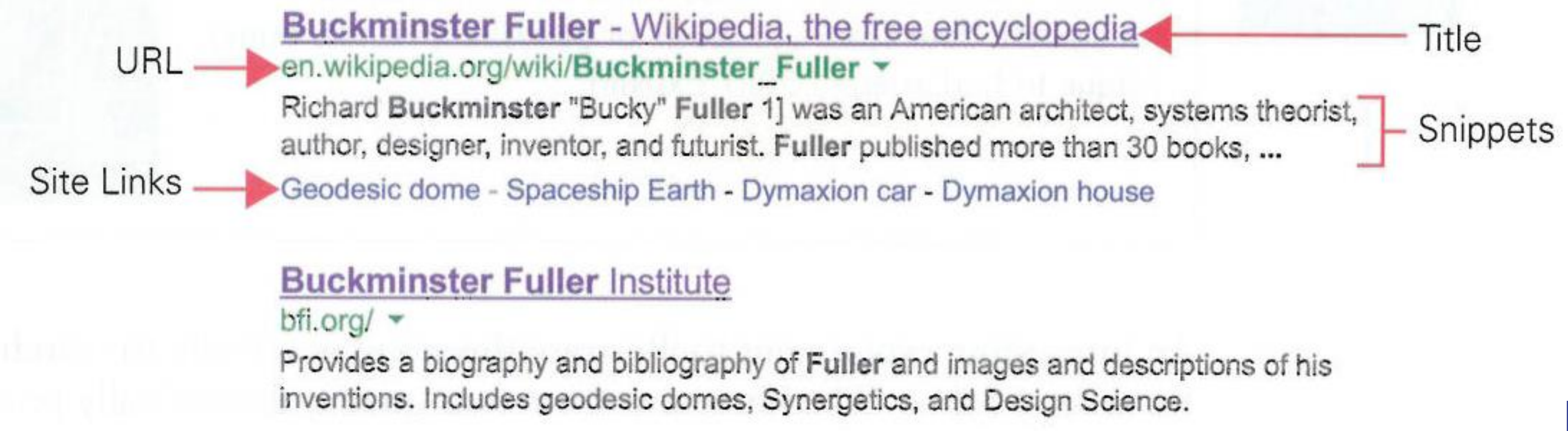
- Remove Specific Words (Minus or NOT)
  - – It is useful to consider eliminating pages with certain words
  - – It is the opposite of adding words that may/will appear on a page
  - – The minus sign is a good way to eliminate wrong interpretations of words



A diagram illustrating how a Web search's hits are narrowed by adding additional terms; each circle represents a Web site, the four boxes represent words descriptive of the Web page; the four searches based on the red words, burgundy words, fuchsia words, and purple words collect the pages into the area of that color.

2. Anatomy of a Hit

- What is displayed with each hit?
  - Title: This is the text between the page's <title> and </title> tags
  - Snippet [작은조각, 자투리]
    - This information is a preview of what might be on the page
    - Usually a short phrase from the page containing one or more of the searched words
  - URL: The URL that is linked from the title line
  - Site Links
    - These are useful links from the site, which are basically shortcuts
    - These links are found algorithmically
    - They are not "sponsored" links – no one is paying for them



Buckminster Fuller - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Buckminster_Fuller ▾
Richard **Buckminster** "Bucky" **Fuller** 1] was an American architect, systems theorist, author, designer, inventor, and futurist. **Fuller** published more than 30 books, ...
Geodesic dome - Spaceship Earth - Dymaxion car - Dymaxion house

Buckminster Fuller Institute
bfi.org/ ▾
Provides a biography and bibliography of **Fuller** and images and descriptions of his inventions. Includes geodesic domes, Synergetics, and Design Science.

Figure 5.5 The first two hits from a Google search for **buckminster fuller.**

3. Using the Hit List

- Checking the hits is a process of filtering
  - Skim the top level of information, and look deeper if it looks promising
  - If not, continuing skimming
- “Looking deeper” ranks the information:
  - Title: It’s the first source of information
  - Snippet: Search terms are shown in bold and some context is given
  - URL: The site name is the first check of how authoritative the information is
- At this point, there is some likelihood that the page includes information you need

4. Once You Find a Likely Page

- A page has been found!:
  - A suggestive title / A promising snippet / A reliable domain
- How close are we to what is wanted?
  - – First, "roll through the page" checking out its main features
  - – Next, look for a date to see how current it is
  - – Finally, find the location of one of the keywords and decide whether you go further
- The goal is still to determine whether you want to stay on the page.
  - –If you've found what you want, you're done. If not, return to the hit list
- If the site is good, need to decide whether the information is perfectly correct:
  - If the importance is high, cross-check the information with another site
  - Find corroborating information!   (corroborate: 입증하다, 확증하다)

Authoritative Information

- Don’t Believe Everything You Read
  - No one in charge of the WWW, so no one checks to see that the information is correct
  - Some information is inaccurate
  - Most information on the Web should be considered suspect
  - Anyone can post a Web page and make random statements or claims… true, partly true, and completely false

Wikipedia

- Wikipedia is an open source document created by knowledgeable and community-minded Internet users
- Anyone can contribute to Wikipedia
- Wikipedia covers an enormous number of topics
- The information contained might not be included in a printed encyclopedia
- Its coverage and timeliness make it a valuable resource
- The fact that many people contribute to Wikipedia is both a weakness and a strength
  - Anyone can add anything to or edit an entry without control
  - Use Wikipedia as a starting point for research

Using the Web for Research

- Applying good research practices will ensure highly reliable information:
  - Question the information from the web
    - Does it make sense? Is it believable? Is it consistent? Does the information fit with everything you already know?
  - Never rely on a single source; always use multiple sources
  - Assess the site’s authoritativeness
  - Vary the kinds of resources you use, including off-line resources.

What Is Authoritative?

- Authoritative means that we are looking for what experts say
  - We assume that experts are well informed on the topic
  - We assume that what experts say is true
- There is no way to verify that what “they” say is true, so we accept that it’s the best available information
- Respected Sources
  - Authoritative information can be gotten from respected organizations
  - Thousands of professional organizations host trustworthy Web sites
  - Individuals are also good sources of information as long as we have some reason to believe them

Sources of Information

- A primary source is a person who has direct knowledge of the information
- People who interview primary sources are secondary sources
  - They are not as reliable as primary sources
- People who watch journalists on TV or read newspaper reports are tertiary sources
- A secondary source does not mean that the information is wrong, only that the possibility exists
- There is no guarantee that even primary sources tell the truth

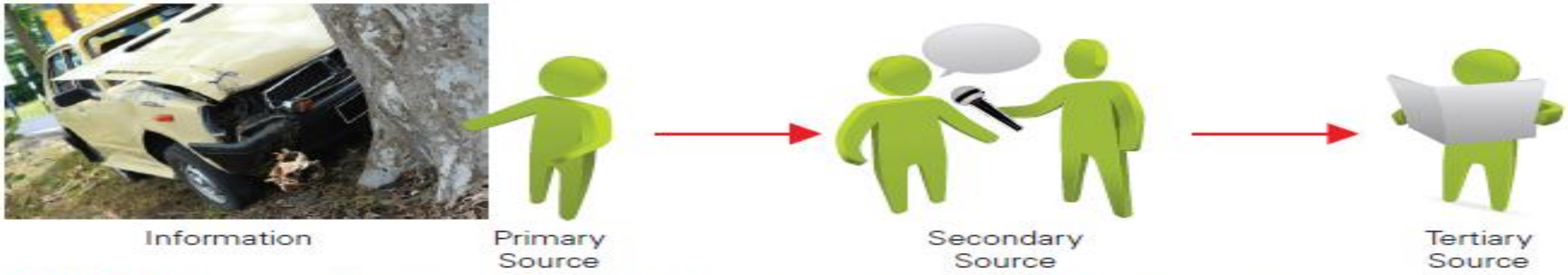
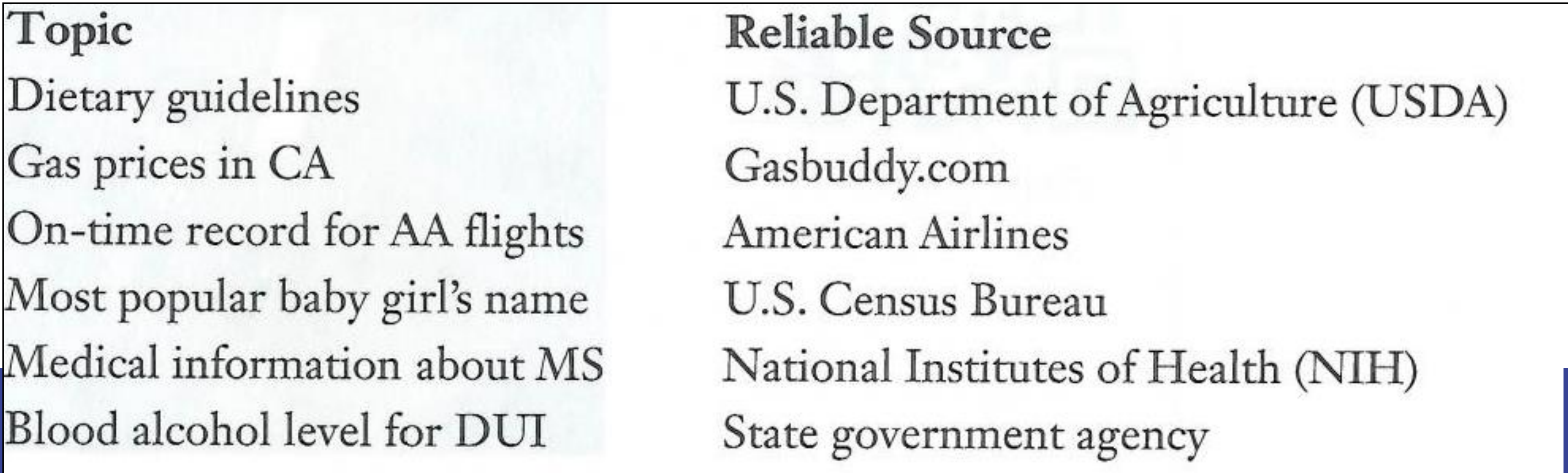


Figure 5.6 A schematic diagram of how sources relate to an original information source.

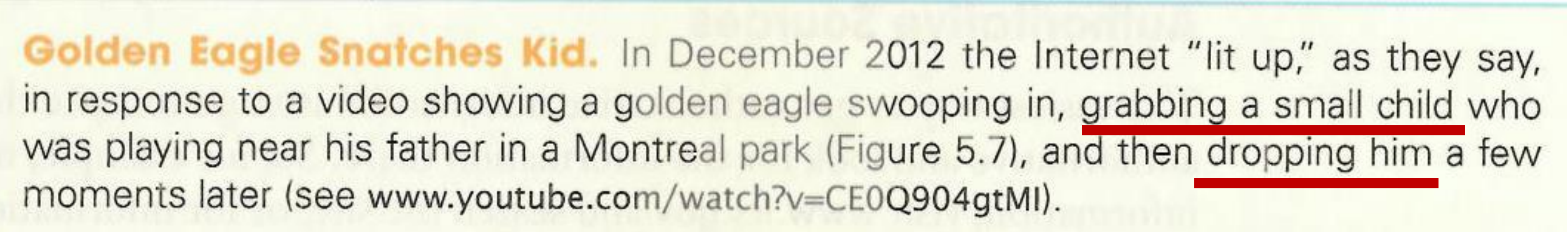
Authoritative Sources

- The easiest way to get authoritative information is to go to a site that you know to be authoritative
- Many agencies and organizations publish information you can depend on
- By going directly to an authoritative source, you are ensured that the information is reliable
- Reliable TAX information ➔ www.irs.gov
- Environment Protection Agency (EPA) ➔ www.epa.gov



Topic	Reliable Source
Dietary guidelines	U.S. Department of Agriculture (USDA)
Gas prices in CA	Gasbuddy.com
On-time record for AA flights	American Airlines
Most popular baby girl's name	U.S. Census Bureau
Medical information about MS	National Institutes of Health (NIH)
Blood alcohol level for DUI	State government agency

An Episode!



Golden Eagle Snatches Kid. In December 2012 the Internet "lit up," as they say, in response to a video showing a golden eagle swooping in, grabbing a small child who was playing near his father in a Montreal park (Figure 5.7), and then dropping him a few moments later (see www.youtube.com/watch?v=CE0Q904gtMI).

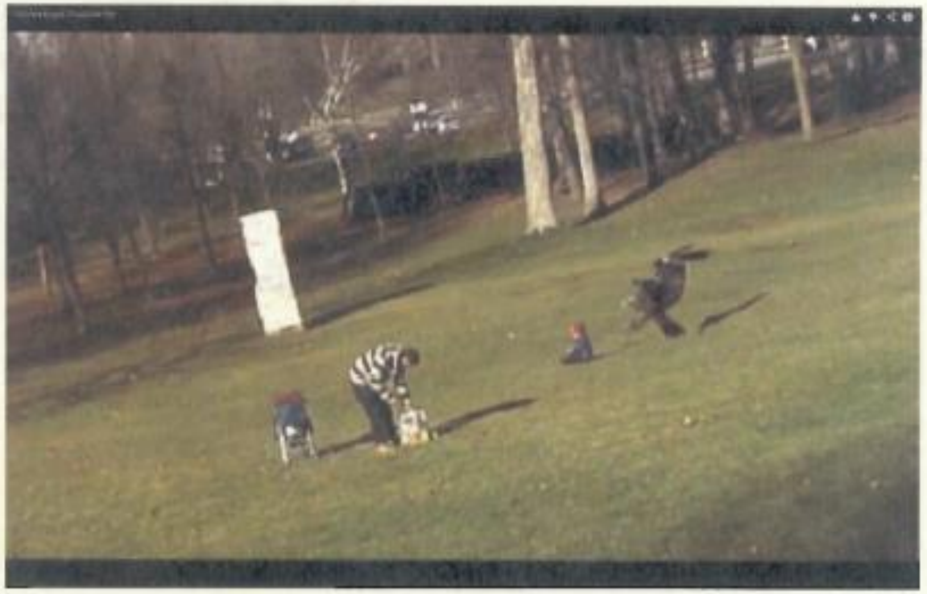
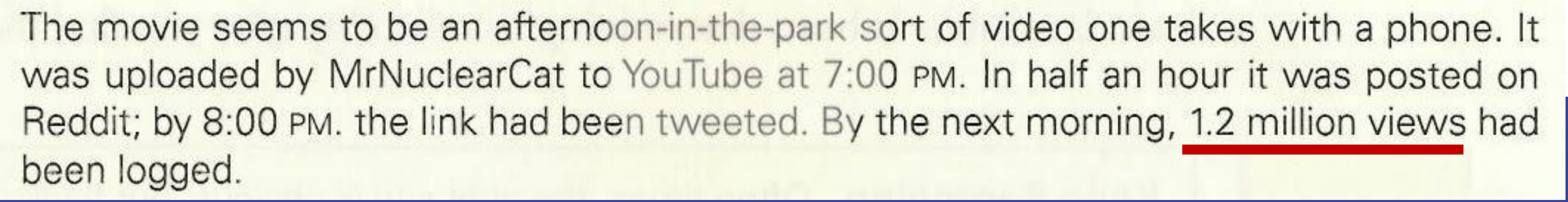


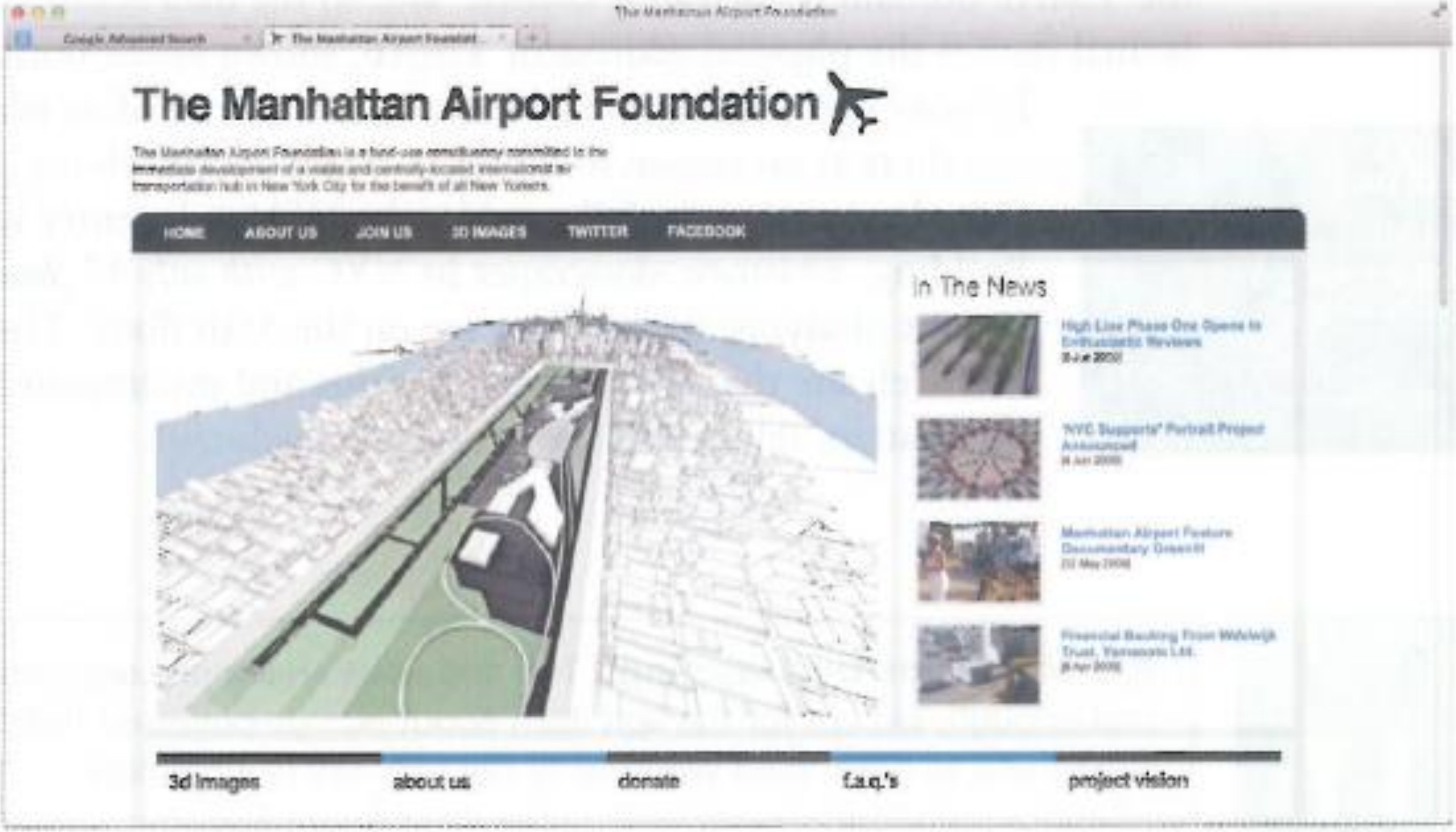
Figure 5.7 Frame from the video *Golden Eagle Snatches Kid.*



The movie seems to be an afternoon-in-the-park sort of video one takes with a phone. It was uploaded by MrNuclearCat to YouTube at 7:00 PM. In half an hour it was posted on Reddit; by 8:00 PM. the link had been tweeted. By the next morning, 1.2 million views had been logged.

Site Analysis… Good? Bad?

- Possible issues for bad sites (Bogus Site?): (useful red flags)
  - Broken links
  - Failure to give contact information
  - Failure to have a non-Web identity
  - Simplistic design (극단적으로 단순한, 허접스러운)
  - No recent updates or blog entries
  - Spelling mistakes
- Legitimate sites may have these issues, too!
- To decide if a site is legitimate, do a little direct research on the information found there
- Bogus Site Samples! (bogus: 가짜의, 엉터리의)
  - The Pacific Northwest Tree Octopus site: zatatopi.net/treeoctopus
  - The Manhattan Airport Foundation site: manhattanairport.org



The Manhattan Airport Foundation

The Manhattan Airport Foundation is a fund-use constituency committed to the immediate development of a viable and centrally-located international air transportation hub in New York City for the benefit of all New Yorkers.

In The News

High Line Phase One Opens to Enthusiastic Reviews
[illegible]

"NYC Supports" Portrait Project Announced
[illegible] Jun 2009

Manhattan Airport Feature Documentary Greenlit
22 May 2009

Financial Backing From Wolvlwijk Trust, Yamamoto Ltd.
8 Apr 2009

Figure 5.8 The Manhattan Airport Foundation home page.

Summary

- We need software and our own intelligence to search the Internet effectively
- Search engines are composed of a crawler and a query processor
- We create queries using the logical operators AND, OR, and NOT, and specific terms to pinpoint the information we seek
- Filtering and “subtracting” search terms removes extraneous hits
- Once we’ve found information, we must judge whether it is correct by investigating the organization that publishes the page, and examining the facts claimed on the page
- We must cross-check the information with other sources, especially when the information is important