Chapter 5

Locating Information

on the WWW
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
How a Search Engine Works

• The first step, **crawling**, visits every Web page that it can find

• How are the pages found?
  – The crawler has a **todo** list that is loaded with a set of pages to start
  – When a URL is found while crawling a page, it adds that URL to the **todo** list

• The main work of the crawler is to build an index
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.
How a Search Engine Works

• The index is a list of tokens (or words) that are associated with the page
• The token might be part of the page’s title
• There are other ways for a token to be associated with a page
• For each token, the crawler creates a list of the URLs associated with that token
How a Search Engine Works

• The second 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
• By creating the index ahead of time, search engines are able to answer user queries very quickly
Multiword Searches

- With a multiple-word query, the pages returned should be appropriate for all of the queried words
- **AND-query**
  - Each page returned *should* be associated with all the words
  - There is no index entry corresponding to a set
  - There is only a list for the individual words
Intersecting Queries

• For multiple words, the query processor fetches the index lists for each of the terms

• URLs that are in all of the lists are looked at and compared

• The query processor *intersects* the lists

• The URL lists are alphabetized to speed up the processing … it is easier to notice when the same URL is on multiple lists
Rules for Intersecting Alphabetized Lists

• To intersect several alphabetized lists:
  1. Put a marker (arrow) at the start of each token’s index list
  2. If all markers point to the same URL, save it, because all tokens are associated with the page
  3. Move the marker(s) to the next position for whichever URL is earliest in the alphabet
  4. Repeat Steps 2–3 until some marker reaches the end of the list
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 a quarter of a second
Fluency with Information Technology

About 3,640,000 results (0.25 seconds)
Descriptive Terms

• “Hits” on a page means the search term is “associated” with the page
• This does not mean the word is “on” the page
• Web page structure helps a lot to identify descriptive text
Descriptive Terms

• Descriptive text:
  – **Title**—The `<title>` encloses a short phrase describing the whole page
  – **Anchor text**—The highlighted link text, inside `<a . . . >` tags, describes the page it links to
  – **Meta**—A `<meta . . . >` tag in the head section can hold a description of the page
  – **Alt** attributes—The `<img . . . >` tag has an alt attribute that gives a textual description
  – **h1**—Text of top-level headers
Page Rank

• 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
• The higher the PageRank, the closer to the top of the list
Links to Other Pages

• 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 to Other Pages

- Links from pages with a high page ranking are also viewed as more important
- PageRank is computed by the crawler:
  - The crawler looks at page A
  - It notices the links to page B
  - It scores one for B
- Counting the number of links to a page is not sufficient
Links to Other Pages

• After the crawling is completed, the PageRank computation is completed
• The query processor puts together the hit list
• The URLs are sorted by their page ranking, highest to lowest, and returned in that order
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.
Advanced Searches

human *AND* powered *AND* flight

- The Logical Operator *AND*
  - Basic queries are **AND-queries**
  - All words given must be associated with the page for a hit
  - The word **AND** is a *logical operator*
  - **Logical operators** specify a logical relationship between the words it connects
Advanced Searches

human AND powered AND flight

• The Logical Operator **AND**
  – Search engines treat search words as three independent words
  – The words can appear *any*where on the page in *any* order

• Use quotes to mean the **exact phrase** must appear as given...this is more specific than an AND query
Complex Queries

marshmallow OR strawberry OR chocolate

• Another logical operator is OR
  – **OR-queries** hit on pages that are associated with at least of the words
Complex Queries

tigers AND NOT baseball

• Another logical operator is NOT
  – NOT queries specify words that are not to be associated with the page
  – AND is included because we want both requirements to be true
Combining Logical Operators

\[(\text{marshmallow OR strawberry}) \text{ AND sundae}\]

- The logical operators work like arithmetic.
- They can be combined and grouped using parentheses.
- Google uses a minus (−) as an abbreviation for NOT.
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

• The bottom of the advanced search box allows you to enter constraints
• These limit your results to specified languages, dates, sites or regions
• Constraints can be used to help pinpoint specific pages
Filtered Searches

- Site searches don’t necessarily use Page Ranking to order the hits
- Use of the advanced search filters can be used to get Page-Ranked hits
Web Searching

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

• How do you find the best search terms?
• It is a sequence of finding ever-more precise terms:
  I. Advanced Search
  II. General Topic
  III. Descriptive Terms
  IV. Refining (Adding Words)
  V. Avoiding Over Constraining
  VI. Removing Words
1. Selecting Search Terms

I. Use Advanced Search
   – Google’s Advanced Search gives control over the results returned
   – You can do complex queries, but Advanced Search provides control
1. Selecting Search Terms

II. 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

III. Choose Descriptive Words

– Be picky about the words we select
– Select precise terms
– Take care using terms that have many other meanings—they become less useful
1. Selecting Search Terms

IV. 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

V. 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
1. Selecting Search Terms

VI. 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
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
    • 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
2. Anatomy of a Hit

• What is displayed with each hit?
  – 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
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
3. Using the Hit List

• “Looking deeper” ranks the information:
  – **Title**
    • It’s the first source of information
    • When the title checks out, look at the . . .
  – **Snippet**
    • Search terms are shown in bold
    • The context around the occurrence is given
    • If the snippet checks out, look at the . . .
3. Using the Hit List

• “Looking deeper” ranks the information:
  – **URL**
    • The domain of the site hosting the page is given
    • The site name is the first check of how authoritative the information is
    • If the URL checks out, look at the . . .
  – **Page itself**
• 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
4. Once You Find a Likely Page

• 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, decide how important it is that the information is perfectly correct:
  • If the importance is high, cross-check the information with another site
  • Find corroborating information!
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
Wikipedia

- 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
    - 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 they are reliable
Primary Sources

• 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
  – They are three levels removed from direct knowledge of the event
Figure 5.6  A schematic diagram of how sources relate to an original information source.
Primary 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
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
Site Analysis… Good? Bad?

• *Possible* issues for bad sites:
  – Broken links
  – Failure to give contact information
  – Failure to have a non-Web identity
  – Simplistic design
  – 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
Summary

• In this chapter we have learned the following:
  – 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
Summary

• In this chapter we have learned the following:
  – 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