Web Searching
Needs for Search Engines

- We have lots of work waiting for us everyday
- Information is needed for many work
  - E-Commerce
    - Microsoft, Oracle web sites over 1 million pages
    - Multi-media databases
  - E-Governance
    - business, citizens: rules, regulations
  - E-Employee
    - rules, regulations, human resources
  - E-Business
    - product specifications, catalogs, contract details, availability

- **Information Overload**
  - Too much information available on the Internet as the World Wide Web becomes popular

- **Distributed and non-homogeneous data**
  - different people come together at the WWW without centralized rules and guidance
  - different structure of documents on the WWW
  - data is non-homogenous for different human association models and value schemes

- Most computer systems are not able to cope with the complexity
History of Search Engines

- Anonymous FTP sites
- Archie
  - for searching FTP sites
  - Collects the site listing of FTP archives into a database for users to query
- Gopher servers for text documents
- Veronica
  - for searching Gopher sites
- Spiders
  - Start on a popular site at a starter page and moves on to other pages following the hyperlinks
  - to collect and search URLs, titles and web headings
  - Compiled into a database that can be searched through query strings
How do Search Engines Work?
How do Search Engines Work?

- **Gathering**
  - visit web sites, bring document copies
  - Revisit sites at some interval of time
  - Most search engines claim that they cover the entire Web, but only a certain percentage of the Web
  - At early stage, search engines collect URLs, titles, and headings
  - Later, collects first hundred kilobytes of documents
  - Now, collects entire documents

- **Indexing**
  - collect key words, phrases in document and index against URLs
  - For each word and phrase, the database contains the URL address of the page as well as the location in the document

- **Searching**
  - search index of words/ phrases
  - Order the list of URLs according to relevance
How do Search Engines Work?

Web Page

Intelligent
Information Agents
by John Smith

Discovery and Management of Information using Agents in the Internet

Index

Keywords:
John Smith
Intelligent Agent

Phrases:
"John Smith"
"Information Agent"
"Discovery and Management"
Common Search Features

- **Boolean Searching**
  - `AND`
  - `OR`

- **NOT**
  - In case of polysemy (words carry different meanings)
  - E.g., `bank AND NOT river`
Common Search Features

- **Phrase Search**
  - “acid rain”

- **Proximity Search**
  - Searching for “the warming of the American continent”
  - Query: America NEAR warming
  - Words appear within N words (N = 10 or 25 or …)

- **Wild card Search**
  - * ?

- **Concept Search**

- **Natural Language Search**
  - Ask Jeeves
Common Display Features

• Relevance Ranking
  • How to arrange the output of the search engine?
    – Word frequency
    – Popularity of the site
    – Location of the words
    – Connectivity of the page and site
      • Page with many links pointing to them are taken as more important
    – Date of page being created
Directories and Databases

- **Searchable subject/topic directory** [Yahoo.com](http://www.yahoo.com)
  - A hierarchy of topics
    - entries are submitted to the site, verified and included in the directory.
    - Entire process is manual but good quality control

- **Virtual Libraries**
  - Maintained by professional such as librarians
  - Access to reference sources such as handbooks, dictionaries, and encyclopedias
    - [http://www.lii.org](http://www.lii.org)

- **Specialized databases**
  - Maintained by for-profit companies, charged
  - [Lexis-Nexis](http://www.lexisnexis.com), [Medline](http://www.medline.com), [Dialog](http://www.dialog.com)
Meta-Search Engines

Coordinate the services of several search-engines

Dogpile  www.dogpile.com
Metacrawler  www.metacrawler.com
Search Engine Evaluation

- Recall
  - Ratio of pages returned to total volume of pages on the world wide web
- Precision
  - Ratio of pages returned that are relevant
- Conflicting nature of the above two criteria
Search Engine Evaluation

• Directories like Yahoo are high in precision, poor in recall

• General sites such as Altavista are high in recall and poor in precision

• Design challenge in search engine is to have both high precision and recall
Search Engine Evaluation

• Search life cycle
  • Initial phase  *exploration phase* high recall is required
  • Later phase  *exploitation phase* high precision is required
Popular Search Engines
(Online Database)

• Yahoo!
  – Database of Web addresses in the manner of Yellow Pages
  – Offer 4 to 5 levels deep subject hierarchy (most others offer only 2 levels)

• Alta Vista
  – Largest copies of the Web at the earliest time
  – Scooter, a spider to roams the Web once every month to update the database
  – Similar subject directory as Yahoo! but smaller
  – “refine” search: suggests likely topics and keywords to tighten the search criteria
Popular Search Engines (continued)

• Lycos
  – “Lycos Top 5%”: manually listing on the basis of site popularity and content quality
  – Site selection is reliable and authoritative
  – Multimedia search, such as pictures, clip art, video clips, sound and music clips

• Excite
  – Concept-based approach instead of keyword-based approach for searching
  – Looks for Web pages that have related ideas and concepts instead of keyword matching
  – Search smaller portion of Web
  – Allows “query by example” using an entry in result

• HotBot
  – Powered by Inktomi

• Google

• Baidu
Flexible Search Strategy

- Determine specificity of search required
  - In exploration or exploitation phase?
- Choose a search engine that suits the specificity of search
  - meta-search engines for initial phase of exploration
- Build a search query
  - Boolean statements
Client-based spider (agents technology)

- machine learning techniques
  - content-based
  - collaborative
- **TueMosaic**
  - DeBra and Post, Edinhoven University of Technology (TUE)
  - users enter keywords, specify the depth and width of search for links contained in the current homepage displayed
  - Fish search algorithm, a modified Best First Search
    - Each URL corresponds to a fish
    - After the document is retrieved, the fish spawns a number of children (URLs)
    - These URLs are produced depending on whether they are relevant and how many URLs are embedded
    - The URLs will be removed if no relevant documents are found after following several links
• WebCrawler
  – Pinkerton
  – first appeared in April of 1994
  – purchased by American Online in January of 1995
  – WebCrawler extend the Fish Search algorithm
    • initiate the search using index
    • follow links in an intelligent order
      – evaluates the relevance of the link based on the similarity of anchor text to the user’s query
• TkWWW
  – Spetka
  – funded by Air Force Rome Laboratory
  – find logically related homepages and return a list of links
    • only one or two hops from the original homepages
    • run in the background to build the HTML indexes, compile WWW statistics, collect a portfolio of pictures, etc.
• **WebAnts**
  
  – Leavitt
  
  – investigates the distribution of information collection tasks to a number of cooperating processors
  
  – create cooperating explorers (ants)
    
    • share the searching results and the indexing loading without repeating each other’s effort
• RBSE (Repository Based Software Engineering)
  - David Eichmann
  - funded by NASA
  - first spider to index document by content
  - four searching algorithms are used,
    - breadth first search from a given URL
    - limited depth first search from a given URL
    - breadth first search from unvisited URLs in the database
    - limited depth first search from unvisited URLs in the database
• Lira System
  – Stanford
  – browse the Internet on users’ behalf
    - searches the Web by taking a bounded amount of time, selecting the best pages and receiving an evaluation from the user
• **Musag System**
  - Hebrew University
  - takes keywords from the users and searches the Web for relevant documents
  - system generates a kind of thesaurus that relates concepts that are semantically similar to each other
• Letizia
  – MIT
  – a user-interface agent for assisting Web browsing
    • does not require any keywords or rating from the user
    • infers users’ interests from browser behavior
    • perform depth-first search
• Personal WebWatcher
  – CMU
  – personal assistant that accompanies user from page to page and highlights interesting hyperlinks
  – generates a user profile based on the content analysis of the requested pages
    • without requesting keywords or ratings
• **Syskill & WeBert**
  
  – UCI
  
  
  
  – collects ratings of the explored Web pages from the user and learns a user profile from them
  
  – separate pages according to their topics and learns a separate profile for each topic
  
  – Applies a naive Bayesian classifier for learning and revising user profile by a set of positive and negative examples
• WAWA
  – Wincosin
  – let users input personal interests and preferences
  – stores them in a neural network, and uses theory revision to refine the obtained knowledge
• Others ….  

• Antagonomy  
  – NEC  
  – Learn user preferences based on both explicit feedback and implicit feedback for retrieving WWW-based newspaper articles  
  – Explicit feedback: user rating  
  – Implicit feedback: scrolling and enlarging operations  
  – User profile: user registered keywords
• CiteSeer
  – Tx, NEC, NMIACS
• **NewsWeeder**
  - CMU

• **Itsy Bitsy Spider**
  - University of Arizona
Hyperlink Analysis
Document Retrieval vs Web Information Retrieval

- **Document retrieval**
  - Find all documents relevant to a user query in a given collection of documents
  - Exclusively based on analysis on the words in the document
- **Web Retrieval**
  - Find all document (Web pages) relevant to a user query in the World Wide Web
  - Based on the analysis on the words in the Web pages and the hyperlink structure of the Web or markup language tags
Why is hyperlink useful?

- Web page authors use hyperlinks to provide valuable information content that is useful to readers
  - Navigational aids
    - Point to other links in the Web site
  - Access to documents that augment the content of the current page
    - Usually point to high quality page
- Hyperlink analysis improves the relevance of search result
  - Hyperlink analysis is adopted in ranking mechanism of most Internet search engines
Hyperlink Analysis

- **Assumption 1**
  - A hyperlink from page A to page B is a recommendation of page B by the author of page A

- **Assumption 2**
  - If page A and page B are connected by a hyperlink, they might be on the same topic

- **Hyperlink analysis are used in two tasks in Web information retrieval**
  - Crawling
  - Ranking
Crawling

- Crawling is the process of collecting Web pages
  - In typical document retrieval, the collection is given
  - In Web information retrieval, the search engine needs to find documents
- Crawling starts from a set of source Web pages
  - It follows the source page hyperlinks to find more Web pages
  - It repeats on each new set of pages and continues until no more new pages are discovered or until a predetermined number of pages have been collected
  - It uses the metaphor of a spider “crawling” along the Web
- Crawler decides in which order to collect hyperlinked pages that have not yet been crawled
  - Hyperlink analysis provides a means for judging the quality of pages that will be used in the priority of crawling
Ranking

- Ranking is the process of ordering the returned documents in descending order of relevance
- Ranking uses connectivity-based ranking in hyperlink analysis
- In typical document retrieval, it only uses the words in the documents
  - Vector space model by Salton
- In Web information retrieval, only using word occurrence analysis is not reliable.
  - Some Web page authors may add invisible text to manipulate the ranking algorithm due to commercial interests
- Hyperlink analysis uses the content of other pages to rank the current page
Connectivity-based Ranking

- Two classes of schemes
  - Query-independent schemes
    - Assign a score to a page independent of a given query
  - Query-dependent schemes
    - Assign a score to a page in the context of a given query

- Directed Graph representation
  - **Link graph**
    - Each Web page is modeled by a node
    - If page A contains a hyperlink to page B, there exists a directed edge (A,B)
    - Used for ranking
  - **undirected co-citation graph**
    - Nodes A and B are connected by an undirected edge if and only if there exists a third page C hyperlinking to both A and B
    - A and B are co-cited by C
    - Used for categorization
Query-independent Ranking

- Measure the intrinsic quality of a page
- Assuming the more hyperlinks pointing to a page, the better the page
  - However, it does not distinguish between the quality of a page pointed by a number of low-quality pages and the quality of a page pointed by the same number of high-quality pages
  - PageRank resolves this problem
- PageRank computes the score by weighting each hyperlink to the page proportionally to the quality of the page containing the hyperlink
  - To determine the quality, PageRank runs recursively with an arbitrary initial setting
PageRank

- PageRank $R(A)$ of a page $A$ is defined as
  \[
  P(i) = (1-d) + d \sum_{(j,i) \in E} \frac{P(j)}{O_j}
  \]
  - $d$ is a constant usually set between 0.8 and 0.9
  - $O_j$ is the number of edges leaving page $j$ (number of hyperlinks in page $B$)
- PageRank of a page $A$ depends on the PageRank of a page $B$ pointing to $A$
- PageRank is used by Google
PageRank

\[
\text{PR}(A) = (1-0.85) + 0.85 \times \frac{\text{PR}(D)}{\text{outdegree}(D)} = 0.15 + 0.85 \times \frac{1.0}{1} = 1
\]

\[
\text{PR}(B) = (1-0.85) + 0.85 \times \left( \frac{\text{PR}(A)}{\text{outdegree}(A)} + \frac{\text{PR}(C)}{\text{outdegree}(C)} \right) = 0.15 + 0.85 \left( \frac{1.0}{2} + \frac{1.0}{1} \right) = 1.425
\]

After 14 Iterations, the PageRank scores tend to converge...

<table>
<thead>
<tr>
<th>Iter</th>
<th>A (out=2)</th>
<th>B (out=2)</th>
<th>C (out=1)</th>
<th>D (out=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>0.575</td>
<td>1.0</td>
<td>1.425</td>
</tr>
<tr>
<td>2</td>
<td>1.36125</td>
<td>0.575</td>
<td>0.819375</td>
<td>1.244375</td>
</tr>
<tr>
<td>3</td>
<td>1.207719</td>
<td>0.728531</td>
<td>0.972906</td>
<td>1.090844</td>
</tr>
<tr>
<td>4</td>
<td>1.077217</td>
<td>0.66328</td>
<td>0.972906</td>
<td>1.286596</td>
</tr>
<tr>
<td>5</td>
<td>1.243607</td>
<td>0.607817</td>
<td>0.889712</td>
<td>1.258865</td>
</tr>
<tr>
<td>6</td>
<td>1.220035</td>
<td>0.678533</td>
<td>0.936855</td>
<td>1.164577</td>
</tr>
<tr>
<td>7</td>
<td>1.139891</td>
<td>0.668515</td>
<td>0.956891</td>
<td>1.234703</td>
</tr>
<tr>
<td>8</td>
<td>1.199498</td>
<td>0.634453</td>
<td>0.918572</td>
<td>1.247476</td>
</tr>
<tr>
<td>9</td>
<td>1.210355</td>
<td>0.659787</td>
<td>0.929429</td>
<td>1.20429</td>
</tr>
<tr>
<td>10</td>
<td>1.170365</td>
<td>0.664401</td>
<td>0.94481</td>
<td>1.220424</td>
</tr>
<tr>
<td>11</td>
<td>1.187361</td>
<td>0.647405</td>
<td>0.929775</td>
<td>1.235459</td>
</tr>
<tr>
<td>12</td>
<td>1.20014</td>
<td>0.654628</td>
<td>0.929775</td>
<td>1.215456</td>
</tr>
<tr>
<td>13</td>
<td>1.183138</td>
<td>0.66006</td>
<td>0.938277</td>
<td>1.218526</td>
</tr>
<tr>
<td>14</td>
<td>1.185747</td>
<td>0.652834</td>
<td>0.933359</td>
<td>1.22806</td>
</tr>
</tbody>
</table>
Query-dependent Ranking

Build a query-specific graph, called a neighborhood graph
Carriere and Kazman proposes the following approach to build a neighborhood graph
  – A start set of documents matching the query is fetched from a search engine (e.g. top 200 matches)
  – The start set is augmented by its neighborhood (the set of documents that either hyperlinked to or is hyperlinked to by documents in the start set)
  – Each document in both the start set and the neighborhood is modeled by a node
  – Hyperlink between pages on the same Web host can be omitted since the authors might be affiliated
Query-dependent Ranking

- Neighborhood graphs consist of thousands of nodes, relatively small comparing with the link graph in PageRank
- PageRank is a indegree-based ranking
  - it produce similar ranking in a neighborhood graph
- In query-dependent ranking, another approach divided pages into two classes in a neighborhood graph
  - Authorities – pages with good content on the topic
  - Hubs – pages with many hyperlinks to pages on the topic
- Kleinberg’s developed the hyperlink-induced topic search (HITS)
  - It determines the good hubs and authorities
HITS

- Given a user query,
  - HITS iteratively computes hub and authority scores for each node in the neighborhood graph
  - Ranks the nodes by hub and authority scores
1. Let $N$ be the set of nodes in the neighborhood graph
2. For every node $A$ in $N$, let $\text{Hub}[A]$ be its hub score and $\text{Aut}[A]$ be its authority score
3. Initialize $\text{Hub}[A]$ to 1 for all $A$ in $N$
4. While the vectors $\text{Hub}$ and $\text{Aut}$ have not converged
   1. For all $A$ in $N$, $\text{Aut}[A] = \sum_{(B,A) \in N} \text{Hub}[B]$
   2. For all $A$ in $N$, $\text{Hub}[A] = \sum_{(A,B) \in N} \text{Aut}[B]$
   3. Normalize the $\text{Hub}$ and $\text{Aut}$ vectors
- Elementary linear algebra shows that the $\text{Hub}$ and $\text{Aut}$ vectors will eventually converge
  - No bound on the number of iterations is known
  - In practice, the vectors converge quickly
Normalization:

\[
\text{Authority}(A) = \sum_{(B,A) \in G} \text{Hub}(B)
\]

\[
\text{Hub}(A) = \sum_{(A,B) \in G} \text{Authority}(B)
\]

\[
\sum_{\forall A} \text{Authority}(A) = 1
\]

\[
\sum_{\forall A} \text{Hub}(A) = 1
\]

<table>
<thead>
<tr>
<th>Iteration</th>
<th>A Authority(A)</th>
<th>A Hub(A)</th>
<th>B Authority(B)</th>
<th>B Hub(B)</th>
<th>C Authority(C)</th>
<th>C Hub(C)</th>
<th>D Authority(D)</th>
<th>D Hub(D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>0.166667</td>
<td>0.333333</td>
<td>0.333333</td>
<td>0.166667</td>
<td>0.333333</td>
<td>0.333333</td>
<td>0.333333</td>
<td>0.166667</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Applications of Hyperlink Analysis in Web Information Retrieval

- Search-by-example
- Mirrored Hosts
- Web Page Classification
- Geographical Scope
Search-by-example

- Looks for pages related to a given page
- HITS and co-citation graph perform well in this problem
  - Frequent co-citation indicates relatedness
Mirrored Hosts

- The path of a Web page is the part of the URL following the host
- E.g. in http://www.apple.com/ipod/index.html
  - www.apple.com is the host
  - /ipod/index.html is the path
- Two hosts, H₁ and H₂, are mirrors if and only if for every document on H₂, there is a highly similar document on H₂ with the same path, and vice versa
- Mirrors exhibits a very similar hyperlink structure both within the host and among the mirror host and other hosts
- Mirror Web hosts waste space in the index data structure and lead to duplicate results in Web search engines
- Hyperlink analysis with IP address analysis and URL pattern analysis can detect many near-mirrors
Web page Categorization

• Hyperlink analysis can compute statistics about groups of Web pages
  – E.g. average length and percentage that in a specific language
• PageRank-like random walks can sample Web pages in an almost uniform distribution
  – It can used to measure various properties of Web pages
Geographical Scope

- Some Web pages are of interest only for people in a given region
  - E.g. weather-forecasting page is interesting only to the region it covers
  - E.g. Internal Revenue Service page is of interest to US taxpayers
- A page’s hyperlink structure reflects its range of interest
  - Local pages are mostly hyperlinked to by pages from the same region
  - Hyperlinks to pages of nationwide interest are roughly uniform throughout the country
- Such analysis supports search engines to tailor query results to the region the user is in