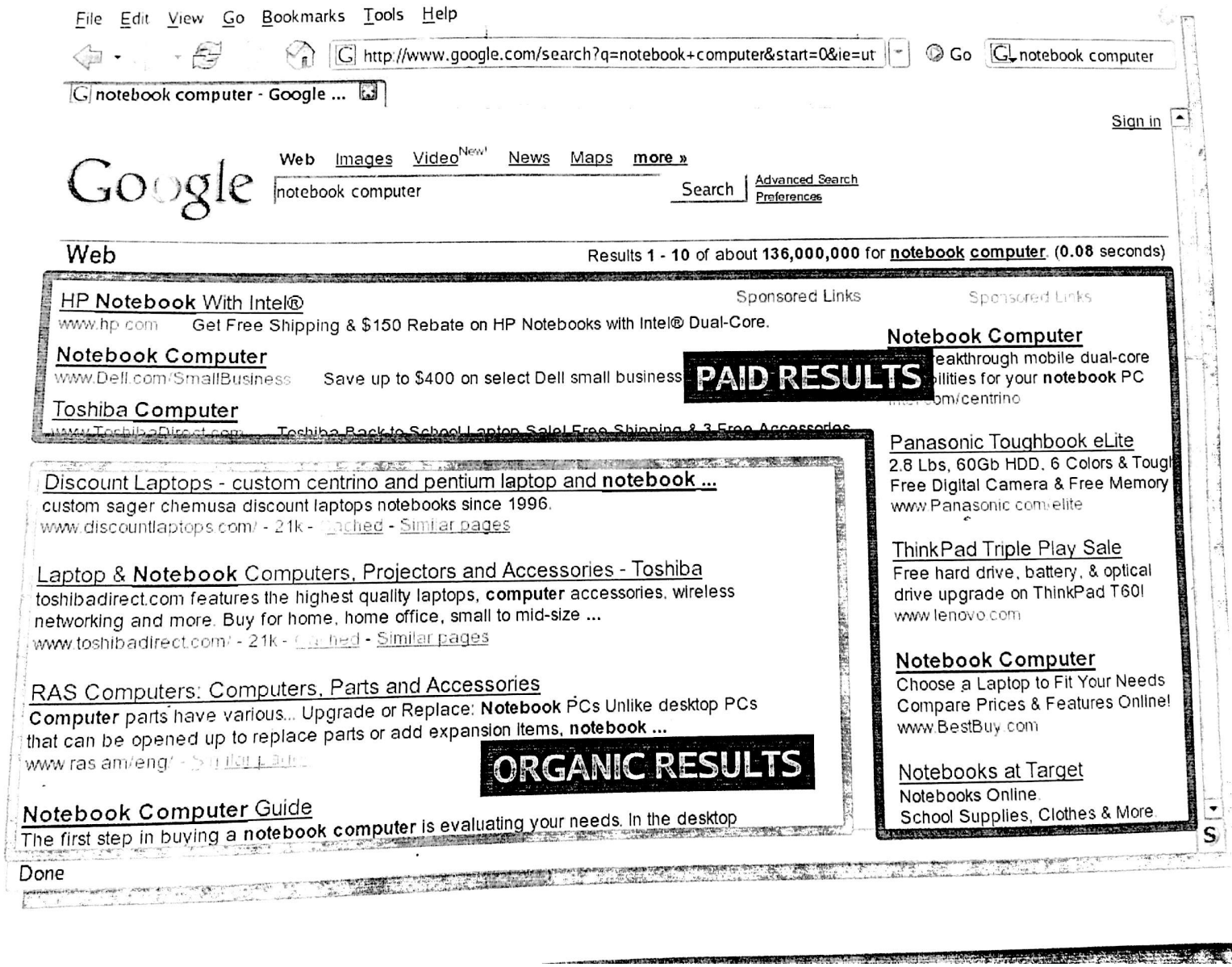
When searcher enters a query — paid results
— organic results.

How Search Engines Work

- Today we show how a search engine works
  - What happens when a searcher enters keywords
  - What was performed well in advance
  - Also explain (briefly) how paid results are chosen
- If we have time, we will also talk about the size of the Web

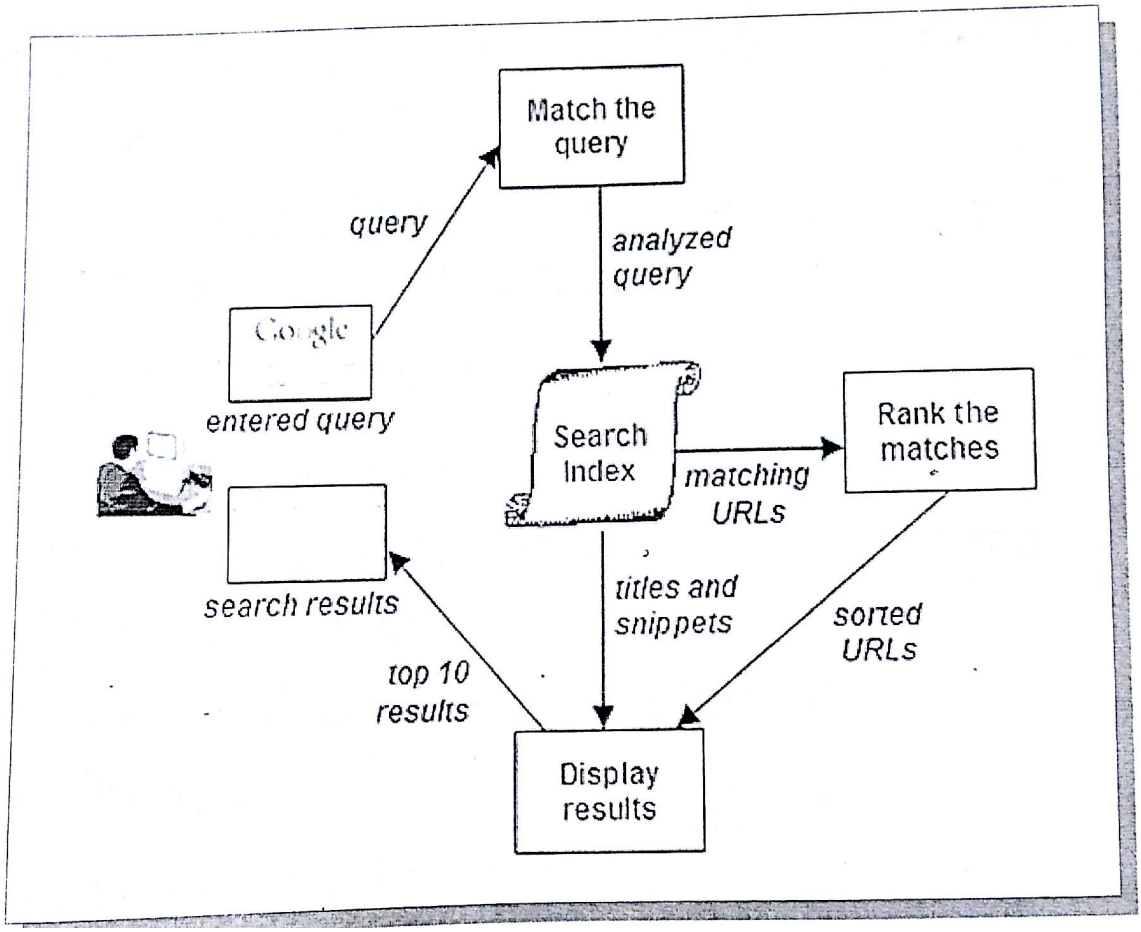
(If you *really* want to know how web search engines work, take my CSE345 WWW Search Engines course in the spring!)



Building an index

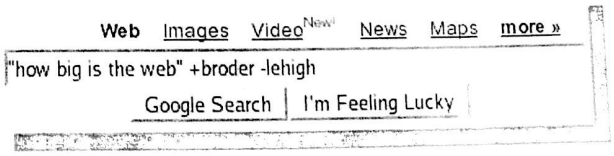
- A search engine does not examine every page on the web when a user puts in a query
- The engine first builds an index
  - Custom database of all the words on all pages
  - Search engine also stores other information

Overview of organic search



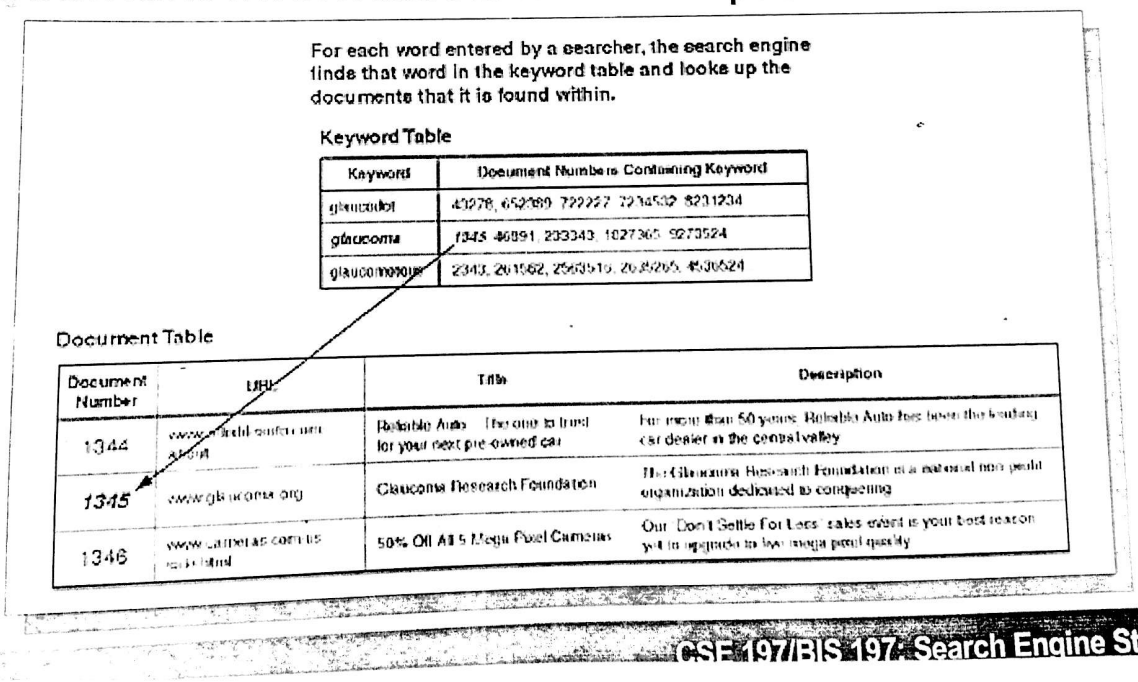
Matching the Search Query

- The *search query* is everything that the user types to get results
  - It is made up of one or more *search terms*, plus optional special characters
- Analyzing the Query
  - Expanding the query
    - Word variants: plural/singular, various verb forms
    - Spelling correction
  - Phrases, anti-phrases, and stop words
  - Word order
  - Search operators



Matching the Search Query

- Organic query matches
  - Find pages with each of the remaining query terms
  - Document IDs are listed in a term index
  - Document information is in a separate doc index



For each word entered by a searcher, the search engine finds that word in the keyword table and looks up the documents that it is found within.

Keyword Table

Keyword	Document Numbers Containing Keyword
glaucubot	40978, 652369, 722937, 7214502, 8211234
glaucoma	*1345*, 46891, 233343, 1027365, 9273524
glaucomatous	2343, 261562, 2563519, 2685265, 4530524

Document Table

Document Number	URL	Title	Description
1344	www.[illegible]	Reliable Auto - The one to trust for your next pre-owned car	For more than 50 years, Reliable Auto has been the leading car dealer in the central valley
1345	www.glaucoma.org	Glaucoma Research Foundation	The Glaucoma Research Foundation is a national non-profit organization dedicated to conquering
1346	www.cameras.com/[illegible]	50% Off All 5 Mega Pixel Cameras	Our Don't Settle For Less sales event is your best reason yet to upgrade to low mega pixel quality

Matching the Search Query

- Paid placement matches
  - Similar to organic match, but using a separate database of ads
  - Uses similar processing to select which query terms to use
  - Advertisers choose which queries can match
    - Might require exact match, or allow broad matching
  - Simpler/faster because there are fewer ads to search through

Ranking Organic Matches

- This is a complex, active research area
  - Goal is to sort matching results from 'best' to 'worst'
  - Many factors contribute to different rankings in the various engines
  - Ranking functions are under continuous change
- Primary factors
  - Text analysis: keyword density and prominence
  - Link analysis: page and site authority estimates
  - Anchor text: terms used to describe page by others
  - Traffic analysis: which results get clicked on

Search and Information Retrieval

- Search on the Web[1] is a daily activity for many people throughout the world
- Search and communication are most popular uses of the computer
- Applications involving search are everywhere
- The field of computer science that is most involved with R&D for search is *information retrieval (IR)*

[1] or is it web?

Information Retrieval

SOSAR

- *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."* (Salton, 1968)
- General definition that can be applied to many types of information and search applications
- Primary focus of IR since the 50s has been on *text* and *documents*

What is a document? →

Big Issues in IR

- Relevance
  - *Retrieval models* define a view of relevance
  - *Ranking algorithms* used in search engines are based on retrieval models
  - Most models describe statistical properties of text rather than linguistic
    - i.e. counting simple text features such as words instead of parsing and analyzing the sentences
    - Statistical approach to text processing started with Luhn in the 50s
    - Linguistic features can be part of a statistical model

Big Issues in IR

- *Evaluation*
  - Experimental procedures and measures for comparing system output with user expectations
    - Originated in Cranfield experiments in the 60s
  - IR evaluation methods now used in many fields
  - Typically use *test collection* of documents, queries, and relevance judgments
    - Most commonly used are TREC collections
  - *Recall* and *precision* are two examples of effectiveness measures

Big Issues in IR

- Users and Information Needs
  - Search evaluation is user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as *query expansion, query suggestion, relevance feedback* improve ranking

IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections
- Web search engines are best-known examples, but many others
  - *Open source* search engines are important for research and development
    - e.g., Lucene, Lemur/Indri, *Galago*
- Big issues include main IR issues but also some others

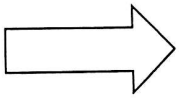
IR and Search Engines

Information Retrieval

Relevance
- *-Effective ranking*

Evaluation
- *-Testing and measuring*

Information needs
- *-User interaction*



Search Engines

Performance
- *-Efficient search and indexing*

Incorporating new data
- *-Coverage and freshness*

Scalability
- *-Growing with data and users*

Adaptability
- *-Tuning for applications*

Specific problems
- *-e.g. Spam*

Search Engine Issues

- Performance
  - Measuring and improving the efficiency of search
    - e.g., reducing *response time*, increasing *query throughput*, increasing *indexing speed*
  - *Indexes* are data structures designed to improve search efficiency
    - designing and implementing them are major issues for search engines

Search Engine Issues

- Dynamic data
  - The "collection" for most real applications is constantly changing in terms of updates, additions, deletions
    - e.g., web pages
  - Acquiring or "crawling" the documents is a major task
    - Typical measures are *coverage* (how much has been indexed) and *freshness* (how recently was it indexed)
  - Updating the indexes while processing queries is also a design issue

Search Engine Issues

- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithm, indexing strategy, interface for different applications