CalPolyPomona

Introduction

Lecture 1

1

- **Introduction**
- History
- Core Issues

2

Search and Information Retrieval

- Search on the Web 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)*

Information Retrieval

- *"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

- Introduction
- **History**
- Core Issues

Brief Search Engine History

- The idea of using computers to search for information dates back to as early as 1945 (Vannevar Bush – "As We May Think")
- Long history of (non-web) search engines in digital libraries, creating the academic field of Information Retrieval (IR).
- Early Web search engines included JumpStation (1993), Lycos (1994), AltaVista (1995), Yahoo (1995)
- JumpStation (developed at the University of Stirling in Scotland), was the first WWW resource-discovery tool to combine the three essential features of a modern web search engine (crawling, indexing, and searching)
- In 1996, graduate students at Stanford, Lawrence "Larry" Page, and Sergey Brin began working on a crawler. They incorporated as Google Inc. in 1998, and by June 2000 Google had grown their index to over 1 billion URLs (now estimated at 30 trillion).

History of Web Search Engines

Year	Engine	Current status
1993	Aliweb	Inactive
	W3Catalog	Inactive
	JumpStation	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	AltaVista	Inactive (URL redirected to Yahoo!)
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo!	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active (rebranded ask.com)
1997	Northern Light	Inactive
	Yandex	Active
1998	Google	Active
	MSN Search	Active as Bing
1999	AlltheWeb	Inactive (URL redirected to Yahoo!)
	GenieKnows	Active, rebranded Yellowee.com
	Naver	Active
	Teoma	Active
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Acquired by Dassault Systèmes

Prior to Web Search

In the days before major web search engines, there was **no capacity to search the entire WWW**. Users would learn about websites by following a link from an email, a message board, or other site.

By 1991 sites dedicated to organized lists of websites started appearing, often created and curated by the Internet Service Providers who wanted to provide added value to their growing clientele.

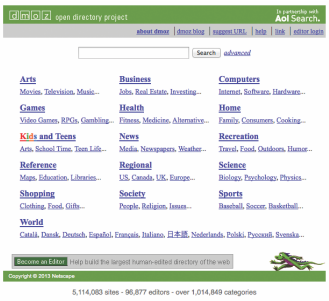
These **web directories** categorized websites into a hierarchy and still exist today.

Prior to Web Search

To be added to a web directory, one would have to submit a request, often by email.

In curated directories the webmasters would then decide whether or not to list you, and if so, where. Many sites took it upon themselves to censor which sites would be listed.

The Open Directory Project (**dmoz.org**) has a more open philosophy.



Impact of Web Search

The impact of search engines is so pronounced that *The Oxford English Dictionary* now defines the verb **google** as

> *Search for information about (someone or something) on the Internet using the search engine Google.**

This shift in the way we retrieve, perceive, and absorb information is of special importance to the web developer since search engines are the medium through which most users will find our websites.

*Note: this should actually say “search…on the WWW...”

- Introduction
- History
- **Core Issues**

Document Search

- Primary focus of IR since the 50s has been on *text* and *documents*
- Example documents:
  - web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, etc.
- Common properties
  - Significant text content
  - Some structure (e.g., title, author, date for papers; subject, sender, destination for email)

Documents vs. Database Records

- Database records (or *tuples* in relational databases) are typically made up of well-defined fields (or *attributes*)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is more difficult

Documents vs. Records

- Example bank database query
  - *Find records with balance > $50,000 in branches located in Amherst, MA.*
  - Matches easily found by comparison with field values of records
- Example search engine query
  - *bank scandals in western mass*
  - This text must be compared to the text of entire news stories

Comparing Text

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval
- Exact matching of words is not enough
  - Many different ways to write the same thing in a "natural language" like English
  - e.g., does a news story containing the text *"bank director in Amherst steals funds"* match the query?
  - Some stories will be better matches than others

Database Records vs Documents (IR) - Summary

Relational Database Management Systems (RDBMS):

- Semantics of each object are well defined
- Complex query languages (e.g., SQL)
- Exact retrieval for what you ask
- Emphasis on efficiency

Information Retrieval (IR):

- Semantics of object are subjective, not well defined
- Usually simple query languages (e.g., natural language query)
- You should get what you want, even if the query is bad
- Effectiveness is the primary issue, although efficiency is important

Dimensions of IR

- IR is more than just text, and more than just web search
  - although these are central
- People doing IR work with different media, different types of search applications, and different tasks

Other Media

- New applications increasingly involve new media
  - e.g., video, photos, music, speech
- Like text, multimedia content is difficult to describe and compare
  - text may be used to represent them (e.g. tags)
- IR approaches to search and evaluation are appropriate

IR Tasks

- Ad-hoc search
  - Find relevant documents for an arbitrary text query
- Filtering
  - Identify relevant user profiles for a new document
- Classification
  - Identify relevant labels for documents
- Question answering
  - Give a specific answer to a question

Dimensions of IR - Summary

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

Core Issues in IR

- Relevance
  - What is it?
  - Simple (and simplistic) definition: "A relevant document contains the information that a person was looking for when they submitted a query to the search engine"
  - *Vocabulary mismatch* problem
  - Many factors influence a person's decision about what is relevant:
    - e.g., task, context, novelty, style
  - *Topical relevance* (same topic) vs. *user relevance* (everything else)

Core Issues in IR

- Relevance
  - *Retrieval models* define a view of relevance
    - formal representation of the process of matching a query and a document.
  - *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 in the 50s
    - Linguistic features can be part of a statistical model

Core 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
    - proportion of retrieved documents that are relevant
    - Recall is the proportion of relevant documents that are retrieved

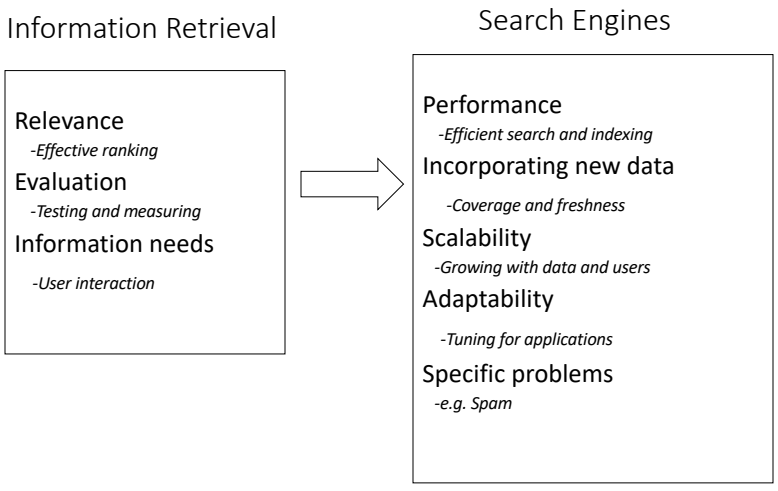
Core Issues in IR

- 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
- Core issues include main IR issues but also some others

IR and Search Engines



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

Reading

- Chapter 1