Lecture 38: Search engines and web crawler: Part I

On completion, the student will be able to:

1. Understand the purpose of a search engine and how it works.
2. Classify the different types of search engines.
3. Identify and explain the working of crawler-based and human-powered search engines.
4. Identify the design challenges of search engines.
5. Understand how keywords are picked up to build search indices.
Search Engines

Introduction

• What is a search engine?
  ➢ It is a software program that helps in locating information stored on a computer system, typically on the world wide web.

• Two types of search engines:
  ➢ Crawler-based
  ➢ Human-powered
Crawler-based Search Engines

• These search engines create their listings automatically.
  ➢ Examples: Google, Yahoo.
  ➢ They crawl or spider the web to create a directory of information.
  ➢ People can search through the directory created in the above process.

• When changes are made to a page
  ➢ Such search engines will find these changes eventually.

Human-powered Directories

• These depend on humans for the creation of the directory.
  ➢ Example: OpenDirectory (http://dmoz.org)
  ➢ One submits a short description for the web site to be listed to the directory.
  ➢ When searching, only the descriptions submitted are looked for.
  ➢ Alternatively, editors can write reviews for some web sites.

• When changes are made to a page
  ➢ Has no effect on the listing.
Hybrid Search Engines

- Can accept both types of results.
  - Based on web crawlers.
  - Based on human-powered listings.
- Such hybrid engines can assign priorities.
  - MSN search gives priority to human-powered listings (LookSmart).
  - MSN search also presents crawler-based search results for complex queries (Inktomi, SubmitExpress).
Components of Crawler-based Engines

• They have three major elements:
  ➢ Crawler or spider
    ▪ Visits a web page, retrieves it, and follows the hyperlinks to other pages within the site.
    ▪ Visits the site regularly (say, once every month) and look for changes.
  ➢ Index or catalog
    ▪ It is like a huge book containing a copy of every web page that the crawler finds.
    ▪ Updated when a page changes.
    ▪ Until a page is indexed, it is not available for search.
Search engine software

- This program searches through the millions of entries in the index to find matches to a search.
- Can also rank the matches based on relevance.

• All crawler-based engines have the above basic components, but differ in the ways these are implemented and tuned.
History of Search Engines

Introduction

• In the early days of Internet
  ➢ Anonymous FTP sites were very common and heavily used.
  ➢ Do an anonymous ftp, and download the files needed.
  ➢ But how to know where to go?
• The first search engine:
  ➢ Archie
    ▪ Created in 1990. Downloaded directory listings of all files on anonymous FTP sites, and created searchable database.
• Subsequent development:
  ➢ Veronica and Jughead
    ▪ Used to search files stored in Gopher index systems.

• Web search engines:
  ➢ Wandex
    ▪ Developed in 1993.
  ➢ WebCrawler
    ▪ Developed in 1994.
    ▪ The first full-text crawler-based search engine.

➢ Lycos
  ▪ Developed in 1994 at Carnegie Mellon University.

➢ Infoseek, Excite, Altavista, …
  ▪ Emerged subsequently.

➢ Google
  ▪ Important concepts of “link popularity” and “page rank” were introduced.
  ▪ What is the basic concept?
- Keep a count of the number of other websites and webpages that point to a given page.
- Assumption: good pages are pointed to by more than not so good pages.
- This simple measure allows Google to rank the results by number of links.
- In general, search engines like Google use a number of other criteria to determine relevance.

Yahoo!
- Prior to 2004, Yahoo! used Google to provide users with search results.
- Launched its own search engine in 2004.
- Used technologies used in Inktomi and AltaVista, which Yahoo! acquired.

MSN Search
- Most recent search engine, owned by Microsoft.
- Increasing in popularity.
- Windows live search --- a new search platform.
Challenges Faced

• Problem of size
  ➢ The web is growling at a pace much faster than any present-day search engine can possibly index.

• Problem of consistency
  ➢ Many webpages get updated frequently.
  ➢ Requires search engines to visit these pages periodically; thus adding to the work.

• Other problems:
  ➢ The allowed queries are typically limited to search for keywords.
    ▪ Generates many spurious results.
    ▪ Better results may be obtained by limiting matches within a paragraph or phrase, rather than matching random words distributed across the entire page.
  ➢ Dynamically generated sites may be slow to index.
Some search engines do not rank by relevance, but by the amount of money the matching websites have paid them.
- May cause search results to become polluted.

Storage Costs / Time Taken

- Consider a scenario:
  - 10 billion pages of 10KB each.
  - Requires 100TB of storage for index.
- A public search engine requires much more resources.
  - To provide high availability.
  - To calculate query results.
• **Time required:**
  - Suppose we have 100 machines.
  - Crawling 10B pages with 100 machines crawling at 100 pages/second.
    - Requires 11.6 days on a very high speed connection.

• **What most search engines do?**
  - Crawl a small fraction of the web (15-20%) at around the above frequency.
  - Crawl dynamic websites (news, blog) at a much higher frequency.

**Some Basic Concepts**
Keyword Searching

• This is the most basic feature of a search engine.
  - Most search engines do their text query and retrieval using keywords.

• How to associate keywords with documents?
  - Author of the web document specifies them (through meta tags).
  - The search engine determines them.

- How can the search engine determine keywords?
  - Tries to pull out and index words that appear to be significant.
  - The title of the page can give useful information about the document.
  - Words that are mentioned towards the beginning of the document are given more weightage.
  - Words repeated several times in the document are also given more weightage.
Search Refinement

- In addition to basic keyword search, most search engines allow advanced or refined search.
  - Additional options are provided.
  - Typical options provided:
    - Ability to search on more than one word.
    - To give more weight to one search term as compared to others.
    - Exclude certain specified words or phrases.
    - To specify Boolean operators to refine the search: AND, OR, NOT.

- Search for exact phrases, typically enclosed within quotation marks.
- Capitalization, which is useful when searching for some people’s names.
Ranking Pages by Relevance

• How relevancy is measured?
  ➢ A common measure is the search term frequency.
  ➢ Some search engines consider the search term frequency as well as where they are positioned.
    • Words appearing earlier in the document is given more importance.
  ➢ Which documents are most frequently linked to other documents of the web.

• Relevance ranking is very important for the users, because of the sheer volume of information available on the web.

Information on “meta” Tags

• Basic idea:
  ➢ Some search engines index web documents using the meta tags in the HTML document.
  ➢ The web page author can have some control over which keywords are used to index the document.
  ➢ For example, if a web page for Asian Test Symposium 2005 is being created, some of the specified phrases may be:
    • ATS 2005, ATS’05, Asian Test Symposium
**How to Specify Meta Tags?**

```
<HEAD>
    <TITLE>Asian Test Symposium 2005</TITLE>
    <META name="description" content="web site of the symposium">
    <META name="keywords" content="ATS 2005, ATS’05, Asian Test Symposium">
</HEAD>
```

- To prevent a search engine from indexing the web page, the following line can be added:

```
<META name="ROBOTS" content="NOINDEX">
```

**Some observations:**

- Meta tagging was supposed to be a useful feature.
- However, they are less useful nowadays because of the high rate of spam indexing.
  - Web authors use false and misleading keywords in the meta tag.
- Google does not look at the keyword meta tags.
- The description keyword is considered to be more useful.
  - Proper attention to be given here.
End of Lecture 38

SOLUTIONS TO QUIZ QUESTIONS ON LECTURE 37
1. How does Internet telephony basically work?

The speaker speaks into the microphone connected to the PC.
Packets get generated only during the talk spurts
- Application-layer header is added to each chunk.
- The data chunk and the header is encapsulated into a UDP packet.
- The UDP packets are transmitted.

2. What are the possible reasons for packet loss?

There are two main reasons: normal packet loss, and excessive packet delay.

3. How are jitters typically handled?

Three common approaches: use sequence number with each packet, timestamps in the packet header, and delayed playout.
4. What are the typical steps in a SIP session?

There are three parts to a SIP session: (1) establishing a session, (2) communication, (3) terminating the session.

5. What is the difference between RTP and RTCP?

RTP is the Internet standard for sending real-time data over the network.

RTCP is the control part of RTP, and provides data delivery monitoring, source identification, etc.
6. What are the typical information that are sent as part of RTCP status messages?

They are: time stamps, fraction of packets lost, total number of packets sent, senderID of status message.

7. How is entity broadcast typically handled in an asymmetric environment?

Individual entity sends control/status information to the source (i.e. satellite).
Source (satellite) broadcast information to all entity receivers (Reflection).
Quiz Questions on Lecture 38

1. What is the basic purpose of search engine?
2. What is the difference between a crawler-based and a human-powered search engine?
3. In a crawler-based search engine, what is a crawler?
4. In a crawler-based search engine, what is the index?
5. What was the purpose of the software tool Archie?
6. Do the search engines crawl through the entire web? Explain.

Quiz Questions on Lecture 38

7. How can a search engine automatically generate the keywords from a document?
8. How is the page relevance typically measured?
9. What is a meta tag? How is it specified in a HTML file?
10. How can you prevent a crawler from using the keywords meta tag information for indexing?
11. Why do some very popular search engines do not index using keywords meta tags?