Chapter 19: Web Search Basics

Chapter 19: Web Search Basics (Part 1)

Main Points and Topics:

1. **Introduction to Web Search Engines**:
   - Early attempts to manage the vast information on the web.
   - The rise of search engines like Altavista, Yahoo, and Infoseek that used different methods for retrieving information, including taxonomies and full-text search.
2. **The Nature of the Web**:
   - The Web is decentralized and lacks coordination, making search engines face challenges that traditional document collections did not.
   - The **client-server architecture** using **HTTP (Hypertext Transfer Protocol)** allows seamless content retrieval using URLs, which represent the web page addresses.
3. **Growth of Web Usage**:
   - The World Wide Web grew significantly in the 1990s, democratizing information sharing.
   - Tools like web browsers (Netscape, Internet Explorer) simplified content creation by hiding HTML's complexity, promoting **HTML-based publishing**.
   - **Client-server** model where browsers fetch content from servers using requests and responses via URLs and HTTP.
4. **Search Engine Architecture**:
   - Web search engines leverage **inverted indexes** and ranking algorithms for fast, efficient searches.
   - Search engines also faced challenges such as processing and indexing massive web content efficiently.

Detailed Explanations of Various Topics:

1. **HTTP and URL**:
   - **HTTP (Hypertext Transfer Protocol)** is a foundation for the web, enabling the transmission of various types of data (text, images, multimedia) between servers and clients (browsers). It is lightweight and flexible, which contributed to the rapid growth of the Web.
   - **URL (Uniform Resource Locator)** is a unique identifier for a web resource, essentially an address used to access content on the internet. A typical URL includes the protocol (HTTP), domain (such as www.stanford.edu), and path (a specific location within a domain, e.g., `/home/atoz/contact.html`).

2. **HTML and Browsers**:
   - **HTML (Hypertext Markup Language)** is the primary language for structuring content on the web. It defines the structure and format of web pages using elements and tags (e.g., `<p>` for paragraphs, `<a>` for links).
   - **Browsers**: Web browsers interpret HTML and display web content to users. Early browsers like Netscape and Internet Explorer were instrumental in popularizing the Web by making it easy for users to view and create content, without needing deep technical knowledge.

3. **Web Indexing and Search**:
   - Search engines like **Altavista** and **Yahoo** pioneered web search using different approaches. Altavista introduced full-text search with inverted indexes, allowing users to search keywords across vast datasets.
   - Inverted indexes store a mapping of content keywords to their locations in documents, making search efficient.

4. **Challenges in Web Search**:
   - **Scaling Issues**: Search engines in the 1990s had to deal with millions of web pages. This required significant computational power to continuously crawl, index, and rank pages.
   - **Relevance and Ranking**: Simply indexing millions of web pages wasn't enough; search engines had to ensure that the results returned were relevant. Ranking algorithms evolved to ensure that high-quality and authoritative pages appeared first, with later developments focusing on metrics like **authoritativeness** and **link analysis**.

This part of the chapter focuses on the early development of web search engines, the fundamental technologies like HTTP and URLs, and the challenges posed by the vast, decentralized nature of the web. It provides context for the later development of more sophisticated search and ranking algorithms, as well as anti-spam measures.

For Part 2, we can look into the **characteristics of the web** and **ranking systems** further, including the emergence of **spam** and **search engine economics**.

Chapter 19: Web Search Basics (Part 2)

Main Points and Topics:

1. **Characteristics of the Web**:

    - **Decentralization**: No central control over web content creation, resulting in diverse languages, styles, and quality of web pages.
    - **Heterogeneity of Content**: Massive variation in grammar, structure, and content quality. Some pages are professionally written, while others have poor grammar or inconsistent formatting.
    - **Dynamic vs. Static Pages**: Web pages can be static (unchanging content) or dynamic (content generated in response to user queries, often from databases).

2. **The Web Graph**:

    - **Hyperlinks and Graph Structure**: The Web can be modeled as a **graph**, where each web page is a **node**, and each hyperlink between pages is a **directed edge**.
    - **In-Links and Out-Links**: Pages can have **in-links** (links pointing to them) and **out-links** (links from them to other pages). These links form the basis of the **web graph**, which is not strongly connected (some pages are not reachable from others).
    - **Power Law Distribution**: The distribution of in-links (or out-links) does not follow a random pattern but a **power law**, meaning a few pages have many in-links, while most have very few. This is similar to **Zipf's law**, which describes the frequency distribution of words in a language.
    - **Bowtie Structure**: The web graph forms a **bowtie** shape with three main regions:
        - **IN**: Pages that link to the **strongly connected component (SCC)**.
        - **SCC**: Pages that can be reached from any other page in the SCC and vice versa.
        - **OUT**: Pages that are reachable from SCC but do not link back to it.
        - **Tendrils and Tubes**: Isolated pages or paths that are loosely connected to the main web graph.

3. **Spam in Web Search**:

    - **Motivation for Spam**: Web pages with commercial motives often attempt to manipulate search engines by artificially boosting their rankings, a practice known as **spam**.
    - **First Generation of Spam**: Early web spam involved techniques like keyword stuffing (repeating keywords many times) and hiding spammy content by making it invisible (e.g., white text on a white background).
    - **Cloaking**: One sophisticated spam technique involves **cloaking**, where a web server shows one version of a web page to search engine crawlers and another to users. This deceives the search engine into indexing misleading content.
    - **Doorway Pages**: Pages created solely to rank for specific search terms. Once a user clicks on these pages, they are redirected to commercial or irrelevant content.

Detailed Explanations of Various Topics:

1. **Heterogeneity of Web Content**:

    - The **decentralized nature of the web** allows anyone to create content, regardless of expertise. This means there's a wide variation in content quality—ranging from well-structured corporate pages to poorly written amateur blogs.
    - **Language and Style Variability**: The web is multilingual, which complicates the task of information retrieval systems. They need to handle different languages, writing systems, and stylistic variations, making tasks like **stemming** (reducing words to their base forms) much harder than in traditional information systems.

2. **Static vs. Dynamic Web Pages**:

    - **Static Pages**: These do not change unless manually updated, e.g., a professor's personal homepage that is edited periodically.
    - **Dynamic Pages**: These are generated on the fly, often in response to user queries or database requests. Examples include airline booking systems, where flight information is updated in real time. Search engines must crawl both types of pages but face difficulties with dynamic content due to its constant changes.

3. **The Web Graph**:

    - **Graph Representation**: The web is modeled as a **graph** with nodes representing web pages and directed edges representing hyperlinks between pages.

- **In-Links and Out-Links**: The number of in-links and out-links are key metrics in determining the importance of a web page. Pages with many in-links are often more authoritative.
- **Power Law Distribution**: Unlike random graphs, where the probability of links between nodes follows a Poisson distribution, the web follows a **power law**. A few pages (e.g., Google, Wikipedia) have a huge number of in-links, while most pages have very few.
- **Bowtie Structure of the Web**: Studies of the web graph have shown a **bowtie structure**:
  - **SCC (Strongly Connected Component)**: The core of the web where most major sites are found. Pages in the SCC can link to each other.
  - **IN and OUT**: Pages in **IN** link into the SCC, but can't be reached from it. Pages in **OUT** are reachable from the SCC but don't link back.
  - **Tendrils**: Pages outside the bowtie structure that can't reach or be reached by the main SCC.

4. **Spam Techniques**:
   - **Keyword Stuffing**: One of the earliest forms of web spam, where pages would repeat keywords excessively to rank higher on search results. This was eventually detected by search engines and penalized.
   - **Invisible Text**: Spammers would hide keywords in the page by matching the text color to the background. While humans couldn't see the text, search engines could still index it.
   - **Cloaking**: Spammers show different content to users and search engine crawlers. For example, a page may show relevant content to the search engine to rank high, but when users visit, they see irrelevant or commercial content.
   - **Doorway Pages**: Pages designed to rank for specific queries, but redirect users to other, often commercial or unrelated pages once clicked.

This section covers the more technical and structural aspects of the web, including the modeling of web pages as a graph and the emergence of spam tactics designed to exploit search engine ranking systems. These concepts highlight the complexities and challenges that search engines face in delivering relevant and trustworthy results.

For Part 3, we will cover **advertising models**, **the search user experience**, and **indexing challenges**, including **duplicate content detection**.

Chapter 19: Web Search Basics (Part 3)

Main Points and Topics:

1. **Advertising as the Economic Model**:
   - Early web advertising relied on banner ads placed on popular sites like Yahoo, MSN, and CNN.
   - **Cost per mil (CPM)** and **Cost per click (CPC)**: Two major pricing models emerged, with CPM based on impressions (views) and CPC based on user clicks.
   - **Sponsored Search**: The model where advertisers bid on keywords to have their web pages shown in search results, leading to the rise of search engine advertising (such as Google Ads).
2. **The Search User Experience**:
   - Search engines focus on simplicity and speed, with a focus on delivering **relevant results** quickly.
   - Google's success was largely due to emphasizing **precision over recall**, delivering accurate results at the top of the page rather than overwhelming users with many less relevant options.
   - **Three Categories of Search Queries**:
     - **Informational Queries**: Users seek general information on a topic.
     - **Navigational Queries**: Users are looking for a specific website, e.g., entering "Lufthansa" to reach the airline's homepage.
     - **Transactional Queries**: Users want to complete a transaction, such as purchasing a product or downloading a file.
3. **Index Size and Estimation**:
   - Search engines' ability to index a significant portion of the web affects their comprehensiveness, but there are challenges with accurately measuring index size.
   - **Dynamic Pages**: Some web pages are generated on demand, meaning there is an infinite number of possible web pages. These cannot all be indexed by search engines.
   - **Capture-Recapture Method**: A technique to estimate the size of one search engine's index relative to another by sampling pages from both.
4. **Duplicate Content and Shingling**:
   - **Duplicate Pages**: As much as 40% of web pages may be duplicates, where the content is identical or nearly identical.

- **Near-Duplicate Detection**: Search engines use a technique called **shingling** to identify near-duplicate content. **Shingles** are sequences of words (often 4-word sequences), and documents are compared based on their shingles to detect similarities.

Detailed Explanations of Various Topics:

1. **Advertising and the Economic Model**:

   - **Banner Ads**: Early internet advertising primarily consisted of graphical banner ads on websites. These ads were usually priced based on the number of views they generated, measured in **CPM (Cost per Mil)**, where advertisers paid per 1,000 impressions.
   - **CPM vs. CPC**: As web advertising evolved, a more performance-based pricing model emerged: **CPC (Cost per Click)**. In this model, advertisers only paid when users clicked on their ads, reflecting user engagement and intent.
   - **Sponsored Search (Search Advertising)**: Pioneered by Goto (later Overture), this model allowed advertisers to bid on keywords. When users searched for these terms, the highest-bidding advertisers' pages appeared in search results, and they were charged per click.
   - **Google's Model**: Google combined traditional search results (known as **algorithmic search**) with sponsored results, placing paid advertisements alongside organic search results. This blend of sponsored and algorithmic search became the standard for modern search engines.

2. **Search User Experience**:

   - **Relevance and Speed**: Google differentiated itself from early search engines by focusing on returning highly relevant results at the top of the page and by having an uncluttered interface. This ensured that users found the information they needed quickly, which became key to its success.
   - **User Behavior in Web Search**: Users generally type simple queries consisting of two to three keywords. Rarely do they use complex operators like Boolean expressions or wildcards, emphasizing the need for search engines to interpret natural language queries effectively.
   - **Query Types**:
     - **Informational Queries**: These involve users seeking knowledge about a topic. For example, a user searching for "leukemia" is likely gathering information from multiple sources.
     - **Navigational Queries**: When users already know what website or entity they are looking for, they use the search engine to navigate directly to that site. For example, typing "Lufthansa" to find the airline's homepage.

    - **Transactional Queries**: These queries precede an action, such as buying a product, booking a flight, or downloading software.

3. **Index Size Estimation**:

    - **Comprehensiveness vs. Relevance**: While larger search engine indexes are assumed to provide more comprehensive search results, this isn't always the case. Quality and relevance of indexed pages are more important than sheer size.
    - **Dynamic Pages**: Many web pages are **dynamic**, meaning they are generated by databases in response to user queries. For example, entering a query on a flight status website generates a page with the latest information. These dynamic pages complicate the task of determining the size of a search engine's index.
    - **Capture-Recapture Method**: A statistical method used to estimate the relative size of search engine indexes. It involves checking what proportion of pages from one search engine's index appear in the other's index, and vice versa.

4. **Duplicate Content and Shingling**:

    - **Duplication on the Web**: A significant percentage of web pages are duplicates or near-duplicates, which can waste search engine resources. Duplicates occur for various reasons, such as websites mirroring content for redundancy or offering identical products across different domains.
    - **Shingling for Near-Duplicate Detection**: **Shingling** is a technique that breaks documents into sequences of words (usually 4-word sequences called **k-shingles**). The similarity between two documents is measured by comparing their shingles. If two documents have a high overlap in their shingles, they are considered near-duplicates, and one of them is removed from the search engine's index.
    - **Jaccard Coefficient**: This is used to measure the similarity between two sets of shingles. If the coefficient is above a certain threshold (e.g., 0.9), the documents are considered near duplicates.

This section highlights how web search engines have developed beyond just indexing and retrieving content to incorporate economic models like advertising and more sophisticated user interaction. Additionally, it addresses the technical challenges of estimating index size and detecting duplicate content using techniques like shingling. This combination of economic and technical solutions has helped search engines become more efficient and user-friendly.