Module -3

Introduction to Web-Mining: Inverted indices and Compression, Latent Semantic Indexing, Web Search, Meta Search: Using Similarity Scores, Rank Positons
Web Spamming: Content Spamming, Link Spamming, hiding Techniques, and Combating Spam

Introduction to Web-Mining: Web mining involves extracting useful information from web data, which includes web documents, web structure data, and web usage data. It is a critical aspect of web intelligence, enabling organizations to understand and leverage the vast amount of information available on the internet.

- Inverted Index: Definition: An inverted index is a data structure used in information retrieval systems, like search engines, to map content to its location within a database or document collection.

Components:

- Dictionary: A list of all unique terms (words or tokens) found in the dataset.
- Postings Lists: For each term in the dictionary, a postings list contains the document IDs where the term appears, often accompanied by additional information such as the frequency of the term in each document or the positions within the document.

Example: If we have two documents:

Doc1: "the quick brown fox"

Doc2: "the lazy dog"

The inverted index might look like:

"the" -> {Doc1, Doc2}

"quick" -> {Doc1}

"brown" -> {Doc1}

"fox" -> {Doc1}

"lazy" -> {Doc2}

"dog" -> {Doc2}

- Compression: Purpose: To reduce the size of the inverted index, thereby improving storage efficiency and retrieval speed.

Techniques:

- Dictionary Compression:

1. Front Coding: Compresses terms with common prefixes by storing the prefix once and encoding the suffixes.
2. Block Compression: Groups terms into blocks and compresses each block individually.

- Postings List Compression:

1. Delta Encoding: Instead of storing document IDs directly, stores the difference (delta) between consecutive document IDs.
2. Golomb Coding: An efficient encoding scheme for compressing integers, particularly useful for sparse postings lists.
3. Variable Byte Encoding: Encodes integers using a variable number of bytes, which can save space for smaller numbers.

Latent Semantic Indexing (LSI): LSI is a technique in natural language processing and information retrieval that uncovers the underlying structure in the data by reducing the dimensionality of the term-document matrix.

Process:

- Term-Document Matrix Construction: Create a matrix where each row represents a term and each column represents a document. The entries in the matrix represent the frequency of the terms in the documents.
- Singular Value Decomposition (SVD): Decompose the term-document matrix into three matrices: U, Σ, and V. Σ contains singular values, U contains the term vectors, and V contains the document vectors.
- Dimensionality Reduction: Reduce the number of dimensions by keeping only the top k singular values and corresponding vectors, capturing the most significant patterns in the data.

Advantages:

- Handles Synonymy: Identifies that different terms can have similar meanings.
- Reduces Noise: By focusing on the most significant relationships, it reduces the impact of irrelevant details.

- Applications: Enhancing search results, document clustering, and information retrieval.

Web Search

Components:

- Crawling: Automated bots, or spiders, systematically browse the web to collect and index web pages.
- Indexing: Organizing the data collected by crawlers into an efficient structure (e.g., inverted index) for quick retrieval.
- Query Processing: Interpreting user queries to match them against the indexed data.
- Ranking: Ordering the search results by relevance, typically using algorithms like PageRank, which consider the number and quality of links to a page, or machine learning models that incorporate various relevance signals.

Challenges:

- Scalability: Managing and searching through billions of web pages requires highly scalable systems.
- Relevance: Ensuring that the search results are highly relevant to the user's query.
- Freshness: Keeping the index up-to-date with the rapidly changing web content.
- Spam Detection: Filtering out low-quality or spammy content that attempts to game the ranking algorithms.

Meta Search: Using Similarity Scores, Rank Positions

Meta search engines aggregate search results from multiple search engines to provide users with a comprehensive set of results. They combine the strengths of various search engines and provide an integrated list of results, often enhancing the search experience by delivering more diverse and relevant information.

Similarity Scores: Similarity scores are numerical values that represent how closely a document matches a user's query. Each search engine assigns these scores based on its own algorithms. To merge results from different search engines by evaluating how relevant each document is to the search query.

1. **Techniques**:

Vector Space Model: Represents documents and queries as vectors in a multi-dimensional space. The similarity between a document and a query is often measured using cosine similarity, which calculates the cosine of the angle between the document and query vectors.

1. Cosine Similarity

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

where A and B are vectors representing the document and the query, respectively.

2. TF-IDF (Term Frequency-Inverse Document Frequency): Weighs terms based on their frequency in the document and their rarity across all documents. It helps in identifying the importance of terms within a document relative to the entire document collection.

- **TF-IDF Formula**:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

where $\text{TF}(t, d)$ is the term frequency of term t in document d, and $\text{IDF}(t)$ is the inverse document frequency of term t across the document corpus.

3. **BM25:** An improvement over the TF-IDF model that takes into account term frequency saturation and document length normalization.

$$\text{BM25}(d, q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{\text{avgdl}})}$$

where $f(q_i, d)$ is the frequency of query term q_i in document d, k_1 and b are free parameters, $|d|$ is the length of document d, and avgdl is the average document length.

Rank Positions:

Rank positions refer to the placement of documents in the search results list based on their relevance to the query. To combine and re-rank results from multiple search engines, presenting the most relevant documents at the top.

Techniques:

- Borda Count: A rank aggregation method where each result is assigned points based on its position in the individual search engines' rankings. The points are summed across all search engines, and results are re-ranked based on the total score.
- Borda Count Example: If a result is ranked 1st in one search engine (receives the highest points) and 3rd in another, the total points will be the sum of points for 1st and 3rd positions.
- Condorcet Method: An election-based method where each result is compared pairwise with every other result. The result that wins the most pairwise comparisons is ranked highest.
- Weighted Scoring: Assigns different weights to the ranks from different search engines based on their perceived reliability or relevance. The weighted scores are then combined to produce the final ranking.
- Score Normalization: Adjusts the scores from different search engines to a common scale before combining them. This is important because different search engines may use different scoring ranges and methods.
- Rank Aggregation: Combines the ranked lists from different search engines into a single ranked list using algorithms such as CombSUM (sum of scores), CombMNZ (sum of scores multiplied by the number of non-zero scores), and others

Min-Max Normalization

- **Min-Max Normalization**:

$$\text{normalized\_score} = \frac{\text{score} - \text{min\_score}}{\text{max\_score} - \text{min\_score}}$$

- **Web Spamming:** Web spamming involves manipulating web content and links to deceive search engines and achieve higher rankings or visibility. This practice undermines the quality of search results and negatively impacts user experience. There are various techniques used in web spamming, and combating them requires a combination of algorithmic and manual interventions.
- **Content Spamming:** Content Spamming involves manipulating the content on a web page to mislead search engines about its relevance to certain queries. Techniques include:
- **Keyword Stuffing:** Overloading a webpage with keywords or phrases to manipulate its ranking in search results.

Example: Repeating the same keyword excessively in the content, meta tags, or alt text of images. Impact: Search engines may detect unnatural keyword density and penalize the website.

- **Cloaking:** Showing different content to search engines than what is displayed to users.

Example: Serving an optimized, keyword-rich page to search engines while showing a more user-friendly page to visitors. This deceptive practice is against search engine guidelines and can result in severe penalties or delisting.

- **Hidden Text and Links:** Placing text or links on a webpage in such a way that they are invisible to users but visible to search engines.

Example: Using white text on a white background or placing text behind images. Once detected, search engines will penalize the website for attempting to manipulate rankings.

- **Link Spamming** : Link Spamming involves creating or manipulating hyperlinks to affect a website's ranking. Techniques include:
- **Link Farms:** A group of websites that all link to each other, often using automated scripts to create a large number of backlinks.
- **Impact:** While link farms may temporarily boost a site's ranking, search engines have become adept at identifying and devaluing these unnatural links.
- **Paid Links:** Purchasing links on other websites to artificially inflate the site's ranking.
- **Impact:** Google and other search engines can penalize both the buyer and seller of such links, as this violates their guidelines.
- **Comment Spamming:** Posting links in the comment sections of blogs, forums, and other platforms to generate backlinks.
- **Impact:** Most modern platforms use nofollow attributes for user-generated links, rendering this technique ineffective for SEO. Excessive comment spamming can also lead to domain blacklisting.
- **Hiding Techniques:** Hiding Techniques are methods used to conceal spammy practices from search engines and users. These include:
- **Invisible IFrames**: Embedding iframes that load spammy content but are not visible to users. Search engines can detect and penalize websites using invisible iframes for deceptive purposes.
- **CSS Tricks**: Using CSS to hide text or links from users while still making them visible to search engines.

Example: Display: none; or visibility: hidden; properties in CSS. Search engines can identify hidden content through CSS analysis and may penalize sites for these tactics.

- **Doorway Pages:** Creating pages optimized for specific keywords that redirect users to a different page. Search engines penalize doorway pages because they provide a poor user experience and manipulate search rankings.

- **Combating Spam**: Combating web spam requires a multi-faceted approach involving both automated and manual processes:
- **Algorithm Updates**: Search Engine Algorithms: Regular updates (e.g., Google's Panda, Penguin) aim to detect and reduce the impact of spammy techniques. These updates improve the quality of search results by penalizing websites that use deceptive practices.

Manual Review:

Human Evaluators: Teams of human reviewers assess websites for compliance with search engine guidelines. Manual reviews help identify and penalize sophisticated spamming techniques that automated systems might miss.

User Reports:

Feedback Systems: Search engines often allow users to report spammy websites. User reports can lead to further investigation and penalization of offending sites.

Machine Learning and AI:

Automated Detection: Advanced machine learning models can identify patterns indicative of spam. AI enhances the ability of search engines to detect new and evolving spamming techniques.

Spam Databases:

- Blacklists: Maintaining databases of known spammy domains and IP addresses.
- Impact: Search engines can quickly block or penalize known offenders.