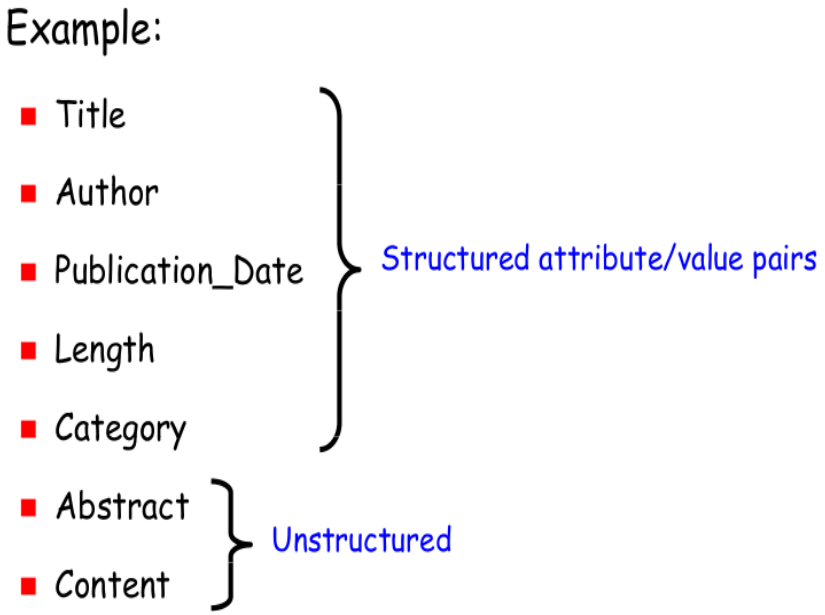
Web content mining:

vector space model, web search, latent semantic indexing (LSI), automatic topic extraction.

Introduction

Web Content Mining applies the principles of data mining to extract meaningful knowledge from the vast and often unstructured content found on web pages and web search results. By representing documents as sets of attributes or features, we can apply traditional data mining techniques to uncover patterns and insights from web data.

Unlike conventional data mining, which typically focuses on structured data (e.g., databases and spreadsheets), web content is predominantly semi-structured or unstructured. This introduces unique challenges in processing and analyzing web content effectively. As a result, Web Content Mining requires specialized methods to manage the inherent complexity and diversity of web-based information.



Preprocessing of text and web pages

The goal of preprocessing is to reduce the complexity of the data, minimizing the search space and improving the efficiency of mining and analysis processes. Effective preprocessing enhances the quality of input data, which in turn yields better insights during the content mining phase.

1. **Tokenization**: This step involves breaking the text into smaller units, called tokens, which are typically individual words or terms. It also includes: Handling digits, hyphens, and punctuations by either retaining, removing, or normalizing them depending on the analysis goal.
2. **Standardization**: Converts all characters to a consistent format, such as: Lowercasing all letters to ensure uniformity, so that "Apple" and "apple" are treated as the same word.
3. **Stopword Removal**: Eliminates common, frequently occurring words that typically add little value to the analysis, such as articles, prepositions, and conjunctions (e.g., "the", "and", "a", "is", "of", "that"). These words are filtered out to focus on more meaningful terms.
4. **Stemming and Lemmatization**: This process reduces words to their root forms, allowing words with the same meaning but different variants (e.g., "running", "runner", "ran") to be treated as the same base term ("run").

For **Web Pages**:

1. HTML Tag Removal: Since web pages contain a mix of content and HTML markup, it is essential to strip away HTML tags, scripts, and other non-informative elements to isolate the actual content.
2. Identification of Main Content Blocks: This step involves detecting and extracting the primary information from the web page, while ignoring irrelevant parts like advertisements, navigation bars, or headers/footers, which do not contribute to the main content.

Vector space model 1/2

How to represent a document?

The representation of a set of documents as vectors in a common vector space is known as a **vector space model** or “ bag of words ”

It uses statistics to add numerical dimensions to unstructured text.

Like: word frequency, document frequency, document length…

It is fundamental to a large number of information retrieval operations, ranging from evaluating documents for a query, to document classification and document clustering.

Vector Space Model 2/2

- **Boolean** representation
  - Each entry describes a document
  - Attribute indicating whether or not a term appears in the document
- The **frequency of terms** noted $tf_{t,d}$: The attributes represent the frequency with which a term t appears in the document d
- **Relative Frequency** : Number of occurrences/Number of words in the document

	word1	word2	word3
doc 1	1	0	0
doc 2	1	1	1
doc 3			

	word1	word2	word3
doc 1	5	0	1
doc 2	7	4	12
doc 3			

Term frequency weighting scheme:TF-IDF

- TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a term in a document relative to a collection of documents (corpus).
- The weighting scheme gives more weight to rare terms, making them more discriminative.

1. N: Total number of documents in the collection.
2. TF (Term Frequency): Measures the frequency of a term t in a document d.

   Increases weight for terms that appear frequently within a single document.
3. DF (Document Frequency) : Counts the number of documents that contain term t.
4. IDF (Inverse Document Frequency): Gives higher weight to rare terms across documents:

$$idf_t = log \frac{N}{df_t}$$

If a term occurs frequently in a large number of documents, it is not **discriminatory** .

- ***tf -idf :***

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

Locate relevant documents

- We can treat a request as a very short document.

Use similarity/distance measure to find similar/relevant documents.

Rank documents based on distance/similarity.

- Euclidean distance:

■ is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Cosine of the angle between the vectors representing the document and the query

Documents "in the same direction" are closely related.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

web search 1/2

Web search originated from information retrieval (**IR**)

The **unit** of information is a **document**, and a large collection of documents is available to form the text database.

A web **crawler** (web robot) is an automated program or script that crawls the web in a methodical and automated manner.

Applications such as business intelligence, website and interest page monitoring, and search engines.

Types of crawlers

- universal
- preferential (targeted) uses similarity and classification

web search 2/2

The operations of a search engine are

Parsing : Syntactic analysis

Indexing :

Searching and Ranking :

1. Preprocessing of query terms
2. Find pages that contain the query terms
3. Sort the pages and return them to the user.

Classify pages: in-links, Occurrence Type, Count, Position:,,

Matrix Decomposition

- Matrix decomposition transforms a large, complex matrix into simpler, smaller matrices.
- Singular Value Decomposition (SVD) $A = U\Sigma V^T$

$$UU^T = I$$
$$VV^T = I$$
$$\sum = diag(\sigma_1, \sigma_2, \cdots, \sigma_p)$$
$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$$
$$p = \min(m, n)$$

- **Benefits**:
  - Reduces Dimensionality: Keeps only the largest singular values to approximate A
  - Data Compression
  - Noise Reduction

Latent Semantic Indexing (LSI) 1/2

- The vector space representation has limitations in its ability to deal with two major **problems** : **synonymy** and **polysemy** .
- Latent semantic analysis (**LSA**) is a natural language processing technique that aims to extract and represent the **contextual meaning** of words through statistical computations applied to a large corpus of text. The fundamental idea is that the set of information about all the contexts in which a given word occurs and does not occur provides a set of mutual constraints that largely determine the semantic similarity between words and sets of words.
- LSA is based on the theory that words with similar meanings often appear in similar contexts.
- LSA uses a **term-document matrix** that describes the occurrences of terms in documents, typically based on the tf-idf method . LSA transforms this occurrence matrix into a relationship between **terms** and some **concepts** , as well as a relationship between these **concepts** and **documents** .

Latent Semantic Indexing (LSI) 2/3

Rank reduction

- LSA facilitates the discovery of a lower-rank matrix that approximates the occurrence matrix, allowing the merging of terms with similar meanings.
- By performing a singular value decomposition on X, we obtain two orthonormal matrices, U and V, as well as a diagonal matrix Σ such that$X = U\Sigma V^T$.
- By selecting the k most important singular values and the corresponding singular vectors in U and V, we can obtain a rank k approximation of the occurrence matrix.
- To compare a query q in the concept space, it is necessary to translate it into this space using the formula

$$\hat{q} = \Sigma_k^{-1} U_k^T q$$

Latent Semantic Indexing (LSI) 3/3

LSA

- the comparison of documents in the concept space (classification, clustering, etc.)
- searching for similar documents between different languages, (using a dictionary)
- searching for relationships between terms (resolution of synonymy and polysemy)
- translate query terms into the concept space, to find semantically related documents (IR)
- find the best similarity between small groups of terms, semantically

automatic theme extraction

- Automatic topic extraction, a natural language processing (NLP) method, seeks to identify and extract the main themes, topics or concepts present in a text corpus, without requiring direct human intervention.
- It is widely used in various fields such as sentiment analysis, document categorization, content recommendation, and social media monitoring. This method helps to merge and reduce manual workload by identifying relevant information within large sets of texts, thus facilitating decision-making and data analysis.

approaches for theme extraction

Several approaches are used to achieve automatic theme extraction:

1. **Frequency-based methods** : They identify the most frequent terms or expressions in a corpus of text, considering that frequently used words are often associated with key topics.
2. **LSA and vector models** : These mathematical methods seek to understand the relationships between words and identify themes from the semantic structure of the text.
3. **Supervised approaches** : These use machine learning models to train a system to recognize and extract specific themes from annotated documents.
4. **Clustering techniques** : These automatically group documents containing similar terms to identify common themes among them.
5. **Neural networks and pre-trained language models** : Recent advances in artificial intelligence have enabled the development of more sophisticated natural language processing models capable of extracting themes with increased accuracy.