Web Content Mining

Web Content Mining

- Pre-processing data before web content mining: *feature selection*
- Post-processing data can reduce ambiguous searching results
- Web Page Content Mining
  - Mines the contents of documents directly
- Search Engine Mining
  - Improves on the content search of other tools like search engines.

Web Content Data Structure

- Unstructured – free text
- **Semi-structured – HTML, XML**
  - Content is, in general, semi-structured, ex.,
    - Title
    - Author
    - Publication_Date
    - Length
    - Category
    - Abstract
    - Content
- More structured – Table or Database generated HTML pages
- Multimedia data – receive less attention than text or hypertext

Web Content Mining: IR View

- Unstructured Documents
  - ✓Bag of words, or phrase-based feature representation
  - ✓Features can be boolean or frequency based
  - ✓Features can be reduced using different feature selection techniques
  - ✓Word stemming, combining morphological variations into one feature

Table 3: An IR view on Web content mining for unstructured documents

Author	Document Representation	Process	Method	Application
Ahonen, et al. [3]	Bag of words and word positions	1 - 2 - 3 - 4	Episode rules	- Finding keywords and keyphrases - Discovering grammatical rules and collocations
Billsus and Pazzani [14]	Bag of words	1 - 2 - 3 - 4	- TFIDF - Naïve Bayes	Text classification
Cohen [27]	Relational	1 - 2 - 3 - 4	- Propositional rule based system Inductive Logic Programming	Text classification
Dumais, et al. [39]	- Bag of words - Phrases	1 - 2 - 3 - 4	- TFIDF - Decision trees - Naive Bayes - Bayes nets - Support Vector Machines	Text categorization
Feldman and Dagan [45]	Concept categories	1 - 2 - 3 - 4	Relative entropy	Finding patterns between concept distributions in textual data
Feldman, et al. [46]	Terms	1 - 2 - 3 - 4	Association rules	Finding patterns across terms in textual data
Frank, et al. [51]	Phrases and their positions	1 - 2 - 3 - 4	Naive Bayes	Extracting keyphrases from text documents
Freitag and McCallum [53]	Bag of words	1 - 3 - 4	Hidden Markov Models	Learning extraction models
Hofmann [63]	Bag of words	1 - 2 - 3 - 4	Unsupervised statistical method	Hierarchical clustering
Honkela, et al. [65]	Bag of words with n-grams	1 - 2 - 3 - 4	Self-Organizing Maps	Text and document clustering
Junker, et al. [69]	Relational	1 - 2 - 3 - 4	Inductive Logic Programming	- Text categorization - Learning extraction rules
Kargupta, et al. [71]	Bag of words with n-grams	1 - 2 - 3 - 4	- Unsupervised hierarchical clustering - Decision trees - Statistical analysis	Text classification and hierarchical clustering
Nahm and Mooney [95]	Bag of words	1 - 2 - 3 - 4	Decision trees	Predicting (words) relationship
Nigam, et al. [98]	Bag of words	1 - 3 - 4	Maximum entropy	Text classification
Scott and Matwin [108]	- Bag of words - Phrases - Hypernyms and synonyms	1 - 2 - 3 - 4	Rule based system	Text classification
Soderland [111]	Sentences, and clauses	1 - 2 - 3 - 4	Rule learning	Learning extraction rules
Weiss, et al. [121]	Bag of words	1 - 2 - 3 - 4	Boosted decision trees	Text categorization
Wiener, et al. [122]	Bag of words	1 - 2 - 3 - 4	- Neural Networks - Logistic Regression	Text categorization
Witten, et al. [124]	Named entity	1 - 2 - 3 - 4	Text compression	Named entity classifier
Yang, et al. [125]	Bag of words and phrases	1 - 2 - 3 - 4	- Clustering algorithms - k-Nearest Neighbor - Decision tree	Event detection and tracking

Web Content Mining: IR View

- Semi-Structured Documents
  - ✓Uses richer representations for features, based on information from the document structure (typically HTML and hyperlinks)
  - ✓Uses common data mining methods (whereas unstructured might use more text mining methods)

Table 4: An IR view on Web content mining for semi-structured documents

Author	Document Representation	Process	Method	Application
Craven, et al. [34]	Relational and ontology	1 - 2 - 3 - 4	- Modified Naive Bayes - Inductive Logic Programming	- Hypertext classification - Learning Web page relation - Learning extraction rules
Crimmins, et al. [35]	Phrase, URLs, and meta information	1 - 2 - 3 - 4	Unsupervised and supervised classification algorithms	- Hierarchical and graphical classification - Clustering
Fürnkranz [54]	Bag of words and hyperlinks information	1 - 2 - 3 - 4	Rule learning	Hypertext classification
Joachims, et al. [68]	Bag of words and hyperlinks information	1 - 2 - 3 - 4	- TFIDF - Reinforcement learning	Hypertext prediction
Muslea, et al. [94]	Bag of words, tags, and word positions	1 - 2 - 3 - 4	Rule learning	Learning extraction rules
Shavlik and Eliassi-Rad [40]	Localized bag of words, and relational.	1 - 2 - 3 - 4	Neural networks with reinforcement learning	Hypertext (homepage) classification
Singh, et al. [109]	Concepts and Named entity	1 - 2 - 3 - 4	- Modified association rule - Classification algorithm	Finding patterns in semi-structured texts
Soderland [111]	Sentences, phrases, and named entity	1 - 2 - 3 - 4	Rule learning	Learning extraction rules

Web Content Mining: DB View

- Tries to infer the structure of a Web site or transform a Web site to become a database
  - ✓ Better information management
  - ✓ Better querying on the Web
- Can be achieved by:
  - ✓ Finding the schema of Web documents
  - ✓ Building a Web warehouse
  - ✓ Building a Web knowledge base
  - ✓ Building a virtual database

Web Content Mining: DB View

- Mainly uses the Object Exchange Model (OEM)
  - ✓ Represents semi-structured data (some structure, no rigid schema) by a labeled graph
- Process typically starts with manual selection of Web sites for content mining
- Main application: building a structural summary of semi-structured data (schema extraction or discovery)

Table 5: Web content mining from a database view

Author	Document Representation	Process	Method	Application
Goldman and Widom [57]	OEM	1 - 2 - 3 - 4	Proprietary algorithms	Finding DataGuide in semi-structured data
Grumbach and Mecca [59]	Strings and relational	1 - 2 - 3 - 4	Proprietary algorithms	Finding schema in semi-structured data
Nestorov, et al. [96]	OEM	1 - 2 - 3 - 4	Proprietary algorithms	Finding type hierarchy in semi-structured data
Toivonen [116]	OEM	1 - 2 - 3 - 4	Upgraded association rules	Finding useful sub-structure in semi-structured data
Wang and Liu [70]	OEM	1 - 2 - 3 - 4	Modified association rules	Finding frequent sub-structures in semi-structured data
Zaiane and Han [127]	Relational	1 - 2 - 3 - 4	Attribute-oriented induction	Multilevel databases

Web Content Mining

- Web content mining is related to data mining and text mining. *[Bing Liu. 2005]*
  - It is related to data mining because many data mining techniques can be applied in Web content mining.
  - It is related to text mining because much of the web contents are texts.
  - Web data are mainly semi-structured and/or unstructured, while data mining is structured and text is unstructured.

Text Mining

- **Text mining**

  Application of data mining to nonstructured or less structured text files. It entails the generation of meaningful numerical indices from the unstructured text and then processing these indices using various data mining algorithms

Text Mining

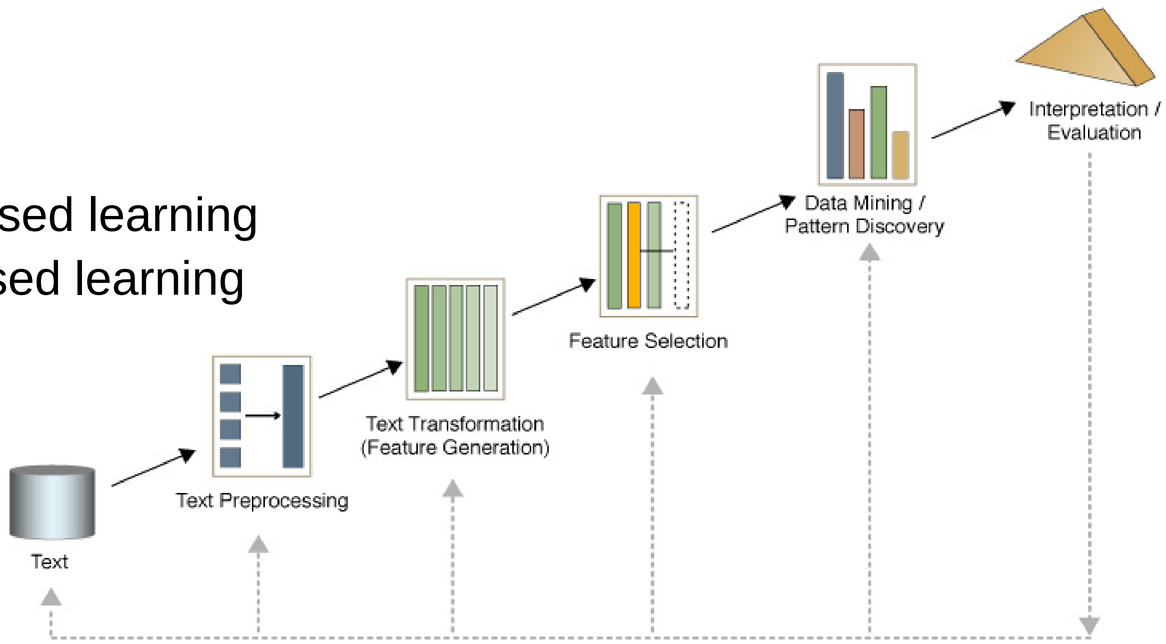
- **Applications of text mining**
  - Analysis of related scientific publications in journals to create an automated summary view of a particular discipline
  - Creation of a “relationship view” of a document collection
  - Qualitative analysis of documents to detect deception

Text Mining

- Text mining hierarchy
  - Natural language processing
  - Classification, clustering
  - Similarity search
  - Term association
  - keyword

Text mining process

- Text preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification- Supervised learning
  - Clustering- Unsupervised learning
- Analyzing results



Text Representation

- Basic idea:
  - Keywords are extracted from texts.
  - These keywords describe the (usually) topical content of Web pages and other text contributions.
- Based on the *vector space model* of document collections:
  - Each unique word in a corpus of Web pages = one dimension
  - Each page(view) is a vector with non-zero weight for each word in that page(view), zero weight for other words

➔ words become "features"

Data Preparation for Text Mining

- Eliminate commonly used words (stop-words)
- Replace words with their stems or roots (stemming algorithms)
- Consider synonyms and phrases
- Calculate the weights of the remaining terms

=>Select most representative terms as features

=>Represent texts (documents) by features

- Each text p is represented as a *k*-dimensional feature vector, where *k* is the total number of extracted features

Conceptually, the inverted file structure represents a document-feature matrix, where each row is the feature vector for a page and each column is a feature

Feature Extraction

Training documents

↓

Identification all unique words

↓

Removal stop words

- non-informative word
- ex.{the,and,when,more}

↓

Word Stemming

- Removal of suffix to generate word stem
- grouping words
- increasing the relevance
- ex.{*walker*,*walking*}→*walk*

↓

Term Weighting

- Naive terms
- Importance of term in Doc

Feature Extraction: Weighting Model(1)

- **tf weighting**

tf: term frequency

$$w_{ij} = Freq_{ij}$$

$Freq_{ij}$: := number of times j-th term occurs in document D_i.

Drawback: without reflection of importance factor for document discrimination.

- **Ex.**

D1 *ABRTSAQWAXAO*

D2 *RTABBAXAQSAK*

	A	B	K	O	Q	R	S	T	W	X
D1	4	1	0	1	1	1	1	1	1	1
D2	4	2	1	0	1	1	1	1	0	1

document-feature matrix

ex.,“This”

Ref:[11][22]

Feature Extraction: Weighting Model(2)

- **tf×idf weighting**

  Idf: Inverse Document Frequency

$$w_{ij} = Freq_{ij} * \log(N / DocFreq_j) .$$

N :: = number of documents in the training doc collection.

$DocFreq_j$ ::= number of documents in which the j-th term occurs.

Advantage: with reflection of importance for document discrimination.

Assumption: terms with low DocFreq are better discriminator than ones with high DocFreq in document collection

- Ex.

	A	B	K	O	Q	R	S	T	W	X
D1	0	0	0	0.3	0	0	0	0	0.3	0
D2	0	0	0.3	0	0	0	0	0	0	0

$$\log\left(\frac{10000}{10000}\right) = 0$$

$$\log\left(\frac{10000}{5000}\right) = 0.301$$

$$\log\left(\frac{10000}{20}\right) = 2.698$$

$$\log\left(\frac{10000}{1}\right) = 4$$

Ref:[13][22]

Example

N=10

D1

	A	B	K	O	Q	R	S	T	W	X
$DocFreq_j$	10	5	3	3	5	2	1	5	3	5
$N/DocFreq_j$	1.00	2.00	3.33	3.33	2.00	5.00	10.00	2.00	3.33	2.00
$\log_2(N/DocFreq_j)$	0.00	1.00	1.74	1.74	1.00	2.32	3.32	1.00	1.74	1.00
$Freq_{ij}$	4	1	0	1	1	1	1	1	1	1
tf×idf	0.00	1.00	0.00	1.74	1.00	2.32	3.32	1.00	1.74	1.00

N=10

D2

	A	B	K	O	Q	R	S	T	W	X
$DocFreq_j$	10	5	3	3	5	2	1	5	3	5
$N/DocFreq_j$	1.00	2.00	3.33	3.33	2.00	5.00	10.00	2.00	3.33	2.00
$\log_2(N/DocFreq_j)$	0.00	1.00	1.74	1.74	1.00	2.32	3.32	1.00	1.74	1.00
$Freq_{ij}$	4	2	1	0	1	1	1	1	0	1
tf×idf	0.00	2.00	1.74	0.00	1.00	2.32	3.32	1.00	0.00	1.00

Feature Extraction: Weighting Model(3)

- **Entropy weighting**

$$w_{ij} = \log(Freq_{ij} + 1) \times (1 - entropy(w_j))$$

where

$$entropy(w_j) = \frac{1}{\log(N)} \sum_{k=1}^{N} \left[-\frac{Freq_{kj}}{gf_j} \log\left(\frac{Freq_{kj}}{gf_j} \right) \right]$$

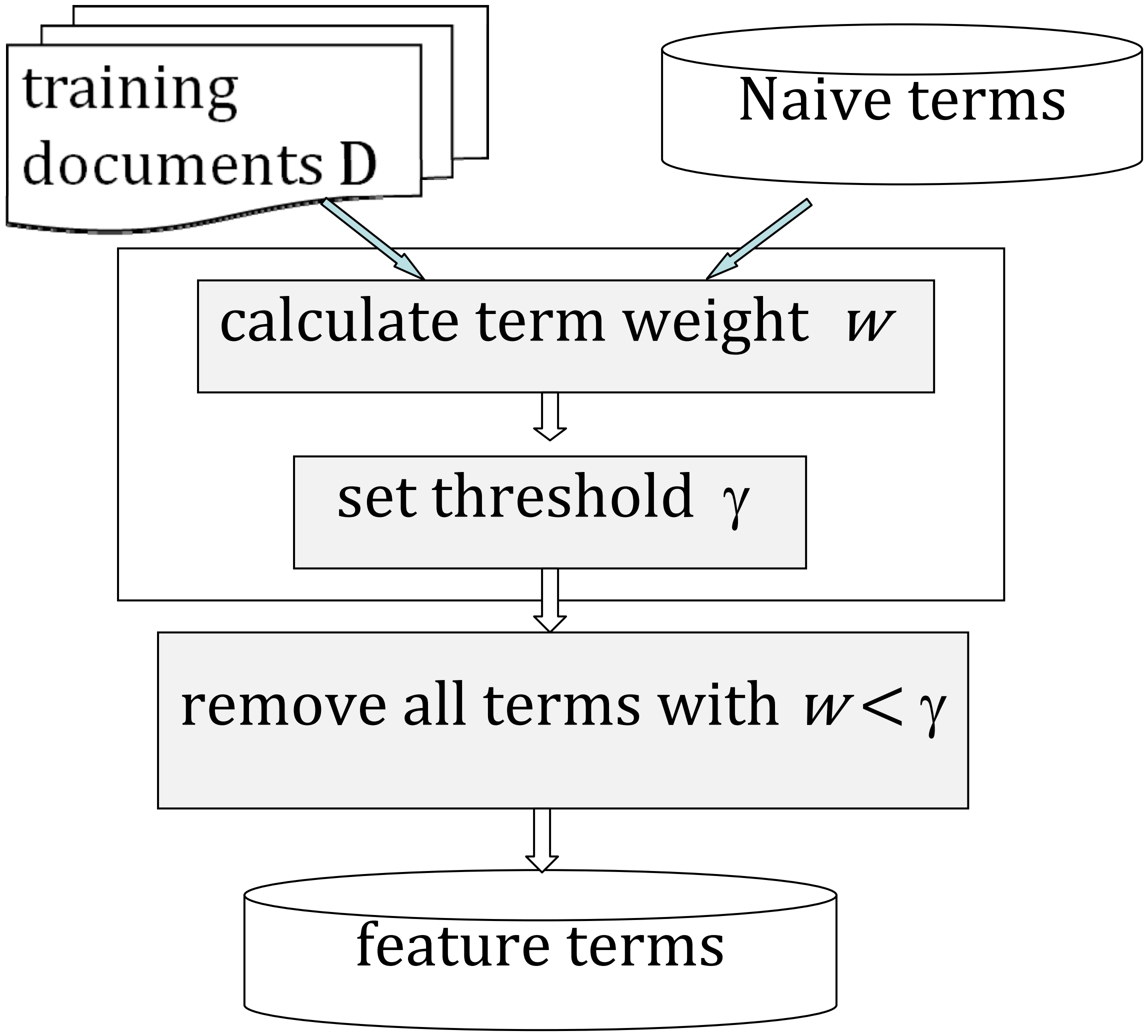
average entropy of j-th term

gf_j ::= number of times j-th term occurs in the whole training document collection

-1: if word occurs once time in every document

0: if word occurs in only one document

Feature Extraction and Dimension Reduction



Document Representation as Vectors

- Starting point is the raw term frequency as term weights
- Other weighting schemes can generally be obtained by applying various transformations to the document vectors

Document Ids **Features**

	nova	galaxy	heat	actor	film	role	diet
A	1.0	0.5	0.3				
B	0.5	1.0					
C	0.4	1.0	0.8		0.7		
D				0.9	1.0	0.5	
E	0.5	0.7			0.9		
F			0.6	1.0	0.3	0.2	0.8

a document vector

Term weights can be:
- Binary
- Raw Frequency in document (Text Freqency)
- Normalized Frequency
-TF x IDF
-Entropy weight

Computing Document Similarity

- Advantage of representing documents as vectors is that it facilitates computation of document similarities
- Example (Vector Space Model)
  - the dot product of two vectors measures their similarity
  - the normalization can be achieved by dividing the dot product by the product of the norms of the two vectors
  - given vectors $X = <x_1, x_2, \ldots, x_n>$ and $Y = <y_1, y_2, \ldots, y_n>$
  - the similarity of vectors X and Y is:

$$\boxed{sim(X,Y) = \frac{\sum_i (x_i \times y_i)}{\sqrt{\sum_i x_i^2 \times \sum_i y_i^2}}}$$

cosine

Learning Methods

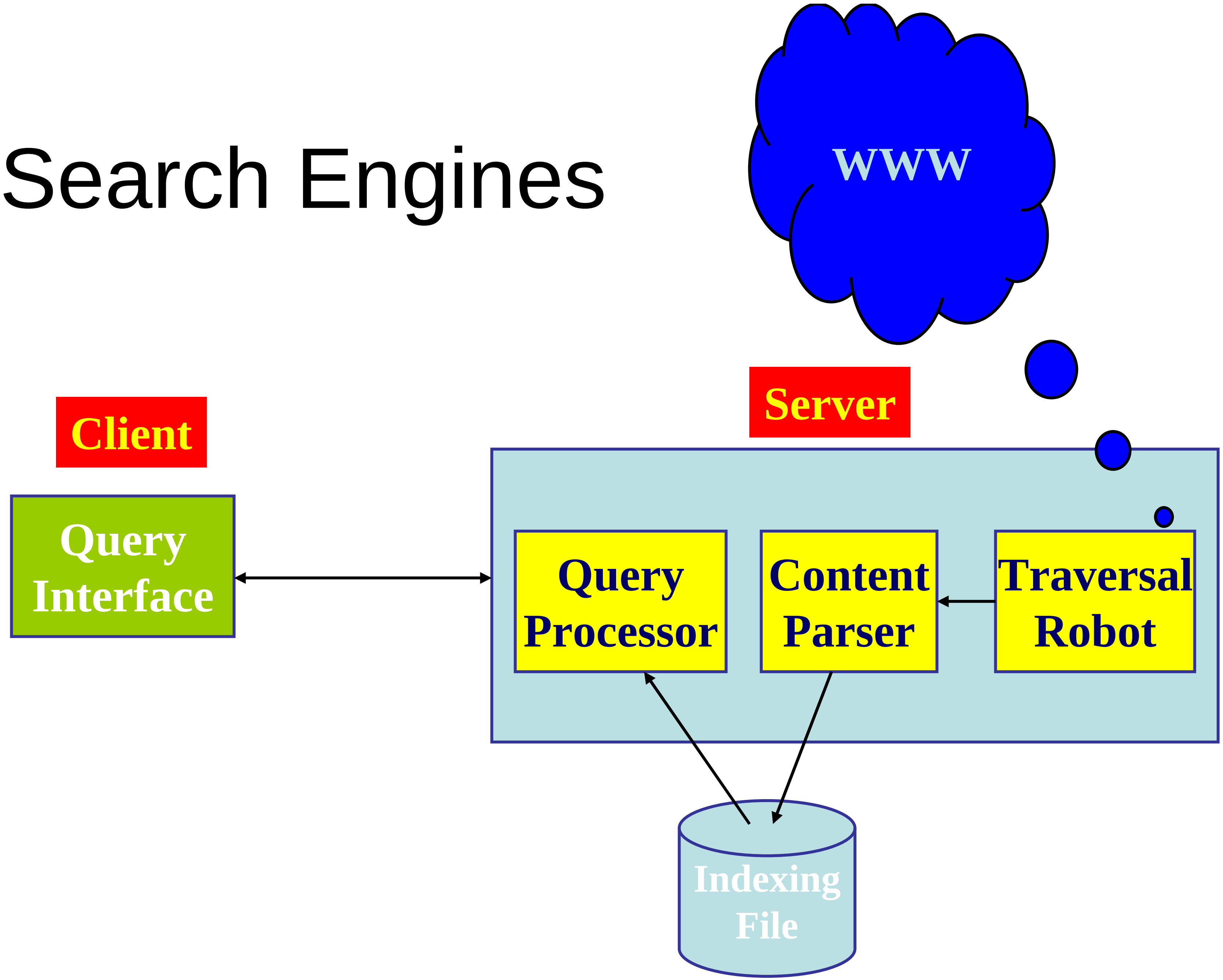
- Classification
  - Instance-Based Methods
  - Decision trees
  - Neural networks
  - Bayesian classification

- Clustering
  - Partitioning Methods
  - Hierarchical Methods

Search Engines

- Search tools
  - Keyword search
  - Hierarchically organized subject catalog or directory

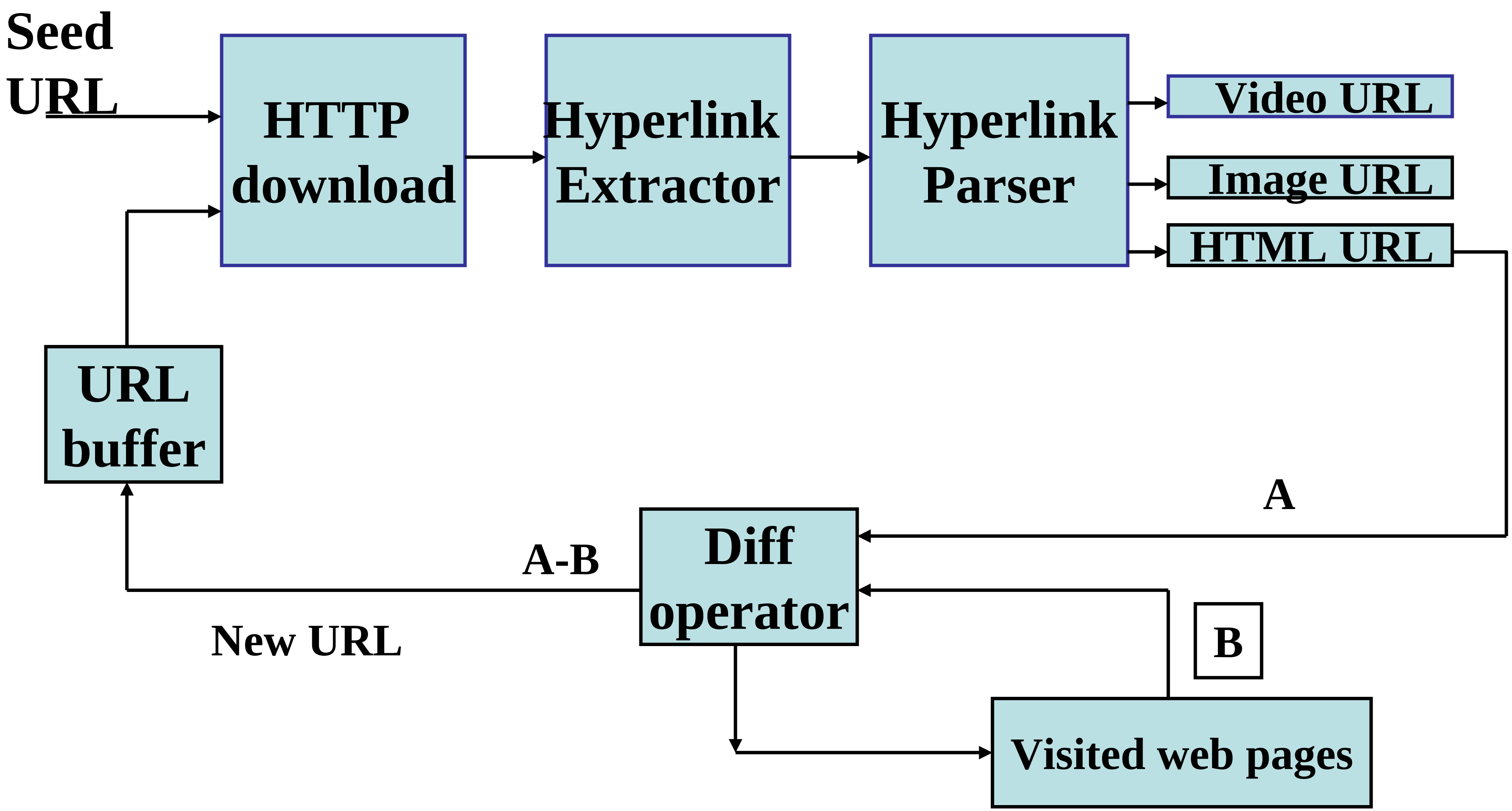
Search Engines



Crawl "all" Web pages?

- Problem: no catalog of all accessible URLs on the Web.
- Solution:
  - start from a given set of URLs
  - Progressively fetch and scan them for new outlinking URLs
  - fetch these pages in turn.....
  - Submit the text in page to a text indexing system

Web Traversal Robots (Crawlers)



Web Robots

- WWW robots (ex. spiders, wanders, crawlers, walkers, ants)
  - programs that traverse WWW
  - recursively retrieving pages hyperlinks by URL
  - goals: automate specific Web-related tasks, e.g.
    - retrieving Web pages for keyword indexing
    - maintaining Web information space at local site
    - Web mirroring
  - actually, robots never *moves*