Data Warehouse and Data Mining

Dhruv Gupta

dhruv.gupta@ntnu.no

14-March-2023

NTNU | Norwegian University of Science and Technology

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Classification
   - Evaluation in Classification
   - Nearest Neighbor Classifiers
   - Other Classification Methods

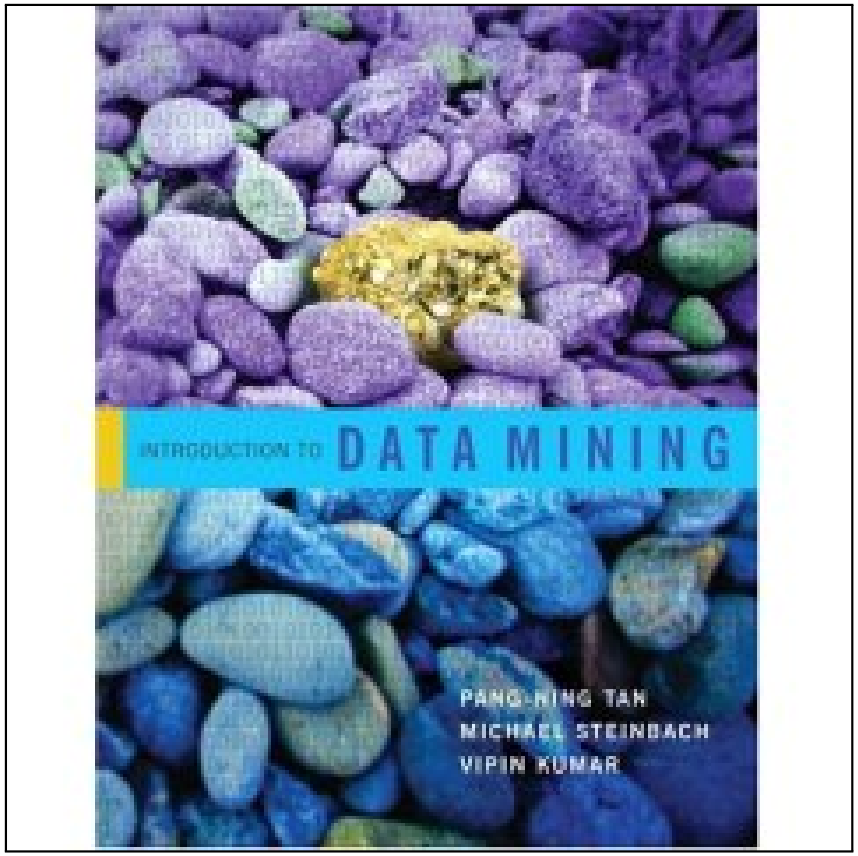
3. Web Usage Mining
   - Web Usage Data
   - Data Modeling
   - Data Mining

Administrative

1. Fourth Assignment
   - Available and due by 23.March.2023.

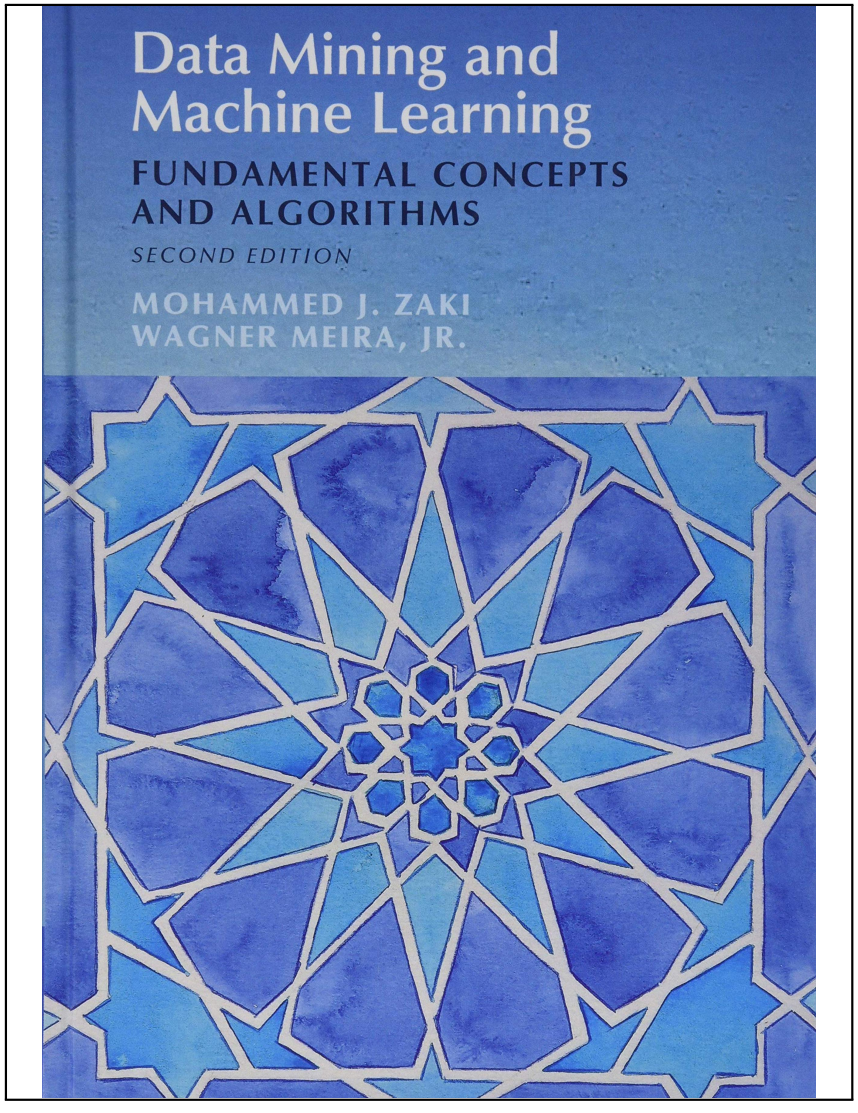
References for "Classification"

1. Book: Tan et al. *"Introduction to Data Mining"*, 1st Edition, 2006, Pearson Education Inc.
2. Text and images for majority of slides in "Classification" are based on the book by Tan et al.



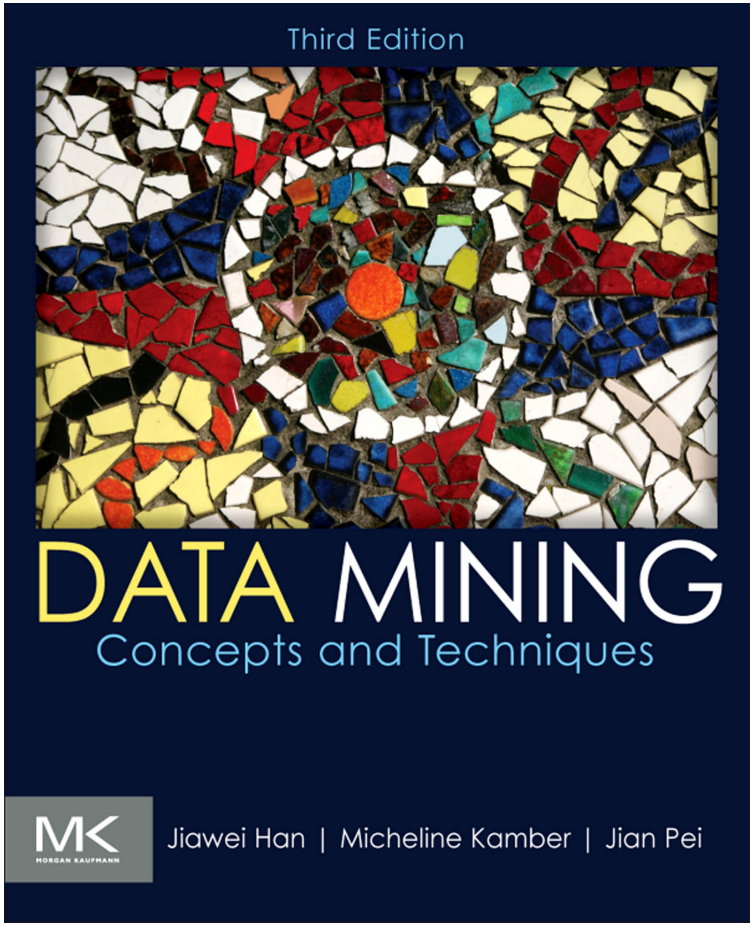
References for "Classification"

1. Book: Zaki and Meira. *"Data Mining and Machine Learning: Fundamental Concepts and Algorithms"*, 2nd Edition, 2020, Cambridge University Press.
2. All text and images for some slides in "Classification" are based on the book by Zaki and Meira et al.



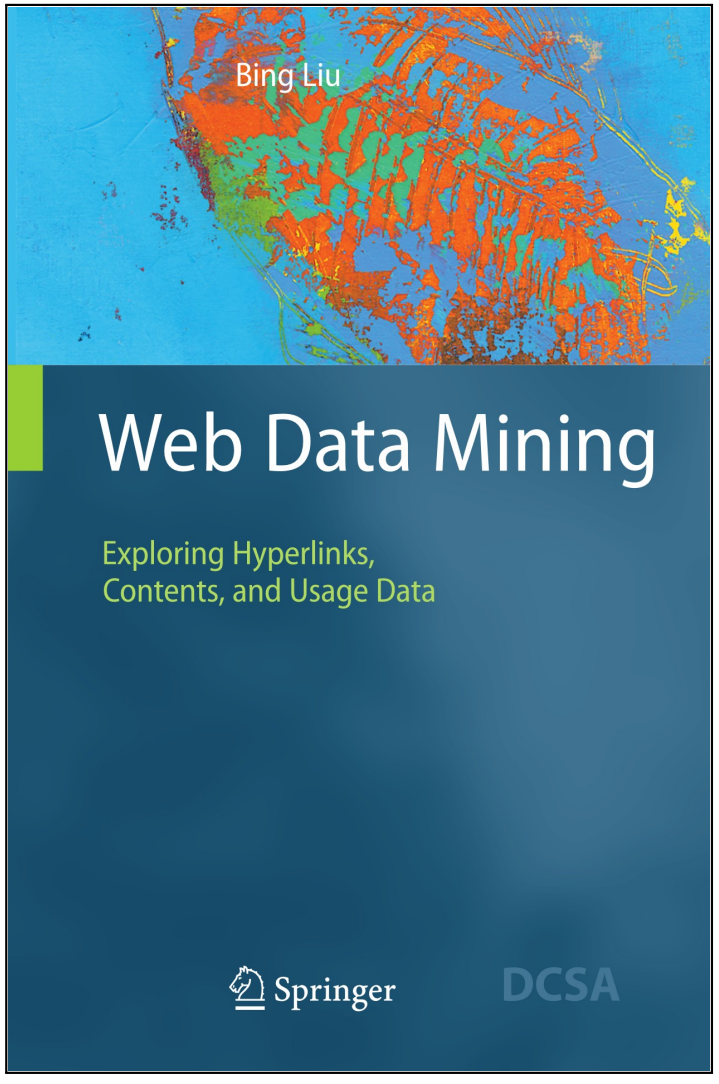
References for "Classification"

1. Book: Han et al. *"Data Mining Concepts and Techniques"*, 3rd Edition, 2012, Morgan Kaufmann Publishers.
2. All text and images for some slides in "Classification" are based on the book by Han et al.



References for "Web Usage Mining"

1. Book: Bing Liu *"Web Data Mining"*, 2nd Edition, 2011, Springer.
2. All text and images for slides in "Web Usage Mining" are based on the book by Bing Liu.



1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Classification
   - Evaluation in Classification
   - Nearest Neighbor Classifiers
   - Other Classification Methods

3. Web Usage Mining
   - Web Usage Data
   - Data Modeling
   - Data Mining

Evaluation Metrics

$$\begin{aligned}\text{Accuracy} &= \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= \frac{a + d}{a + b + c + d}.\end{aligned}$$

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a (True Positive)	b (False Negative)
	Class = No	c (False Positive)	d (True Negative)

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Evaluation Metrics

$$\text{Error Rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}}$$
$$= \frac{FN + FP}{TP + TN + FP + FN}$$
$$= \frac{b + c}{a + b + c + d}.$$

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a (True Positive)	b (False Negative)
	Class = No	c (False Positive)	d (True Negative)

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Evaluation Metrics

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{a}{a + c}.$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{a}{a + b}.$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot a}{2 \cdot a + b + c}.$$

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a (True Positive)	b (False Negative)
	Class = No	c (False Positive)	d (True Negative)

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Methods of Estimation

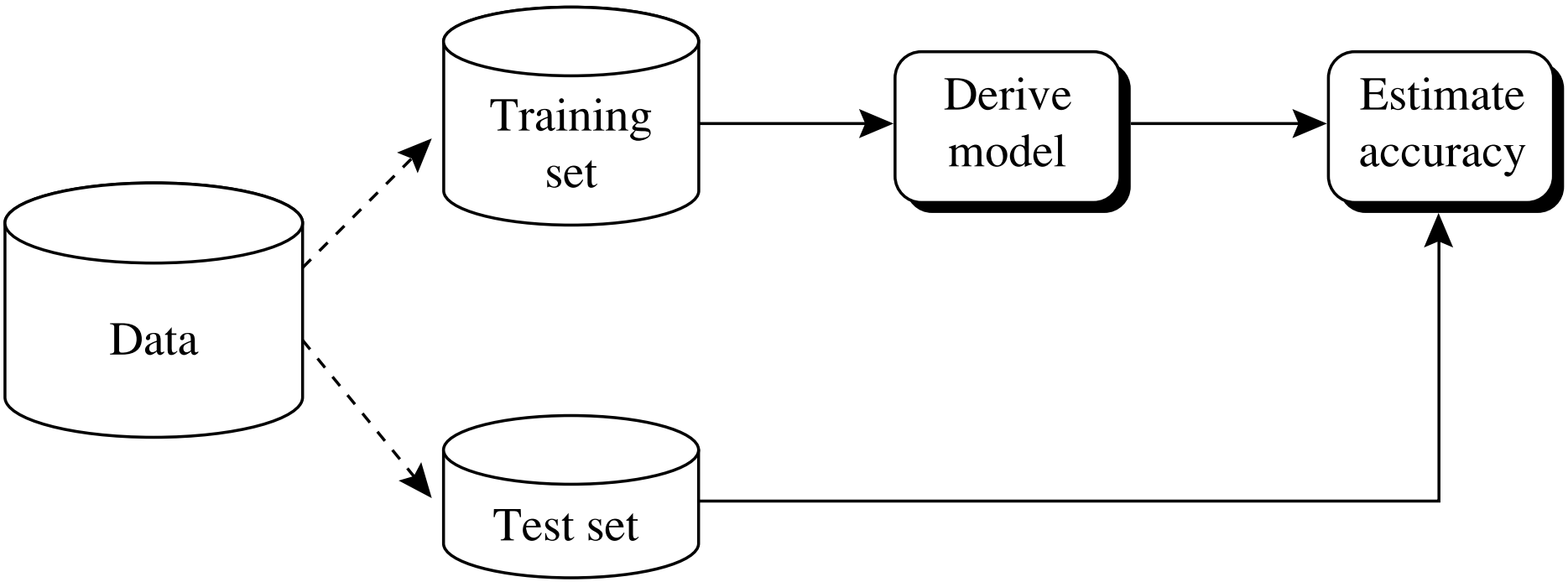


Figure 8.17 Estimating accuracy with the holdout method.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Methods of Estimation

- Holdout:
  - Reserve 2/3 for training and 1/3 for testing.
- Random subsampling:
  - Repeated holdout.
- Cross validation:
  1. Partition data into k disjoint subsets.
  2. k-fold: train on $k-1$ partitions, test on the remaining one.
  - Leave-one-out: $k = n$.
- Bootstrap
  - Sampling with replacement.

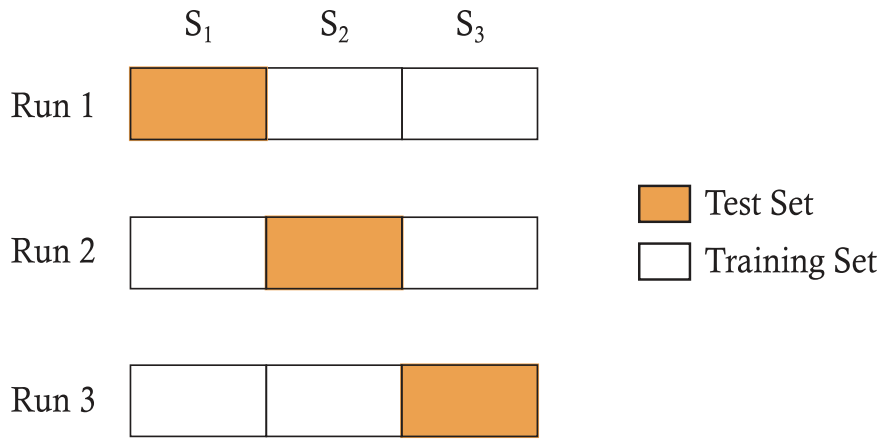


Figure 3.33. Example demonstrating the technique of 3-fold cross-validation.

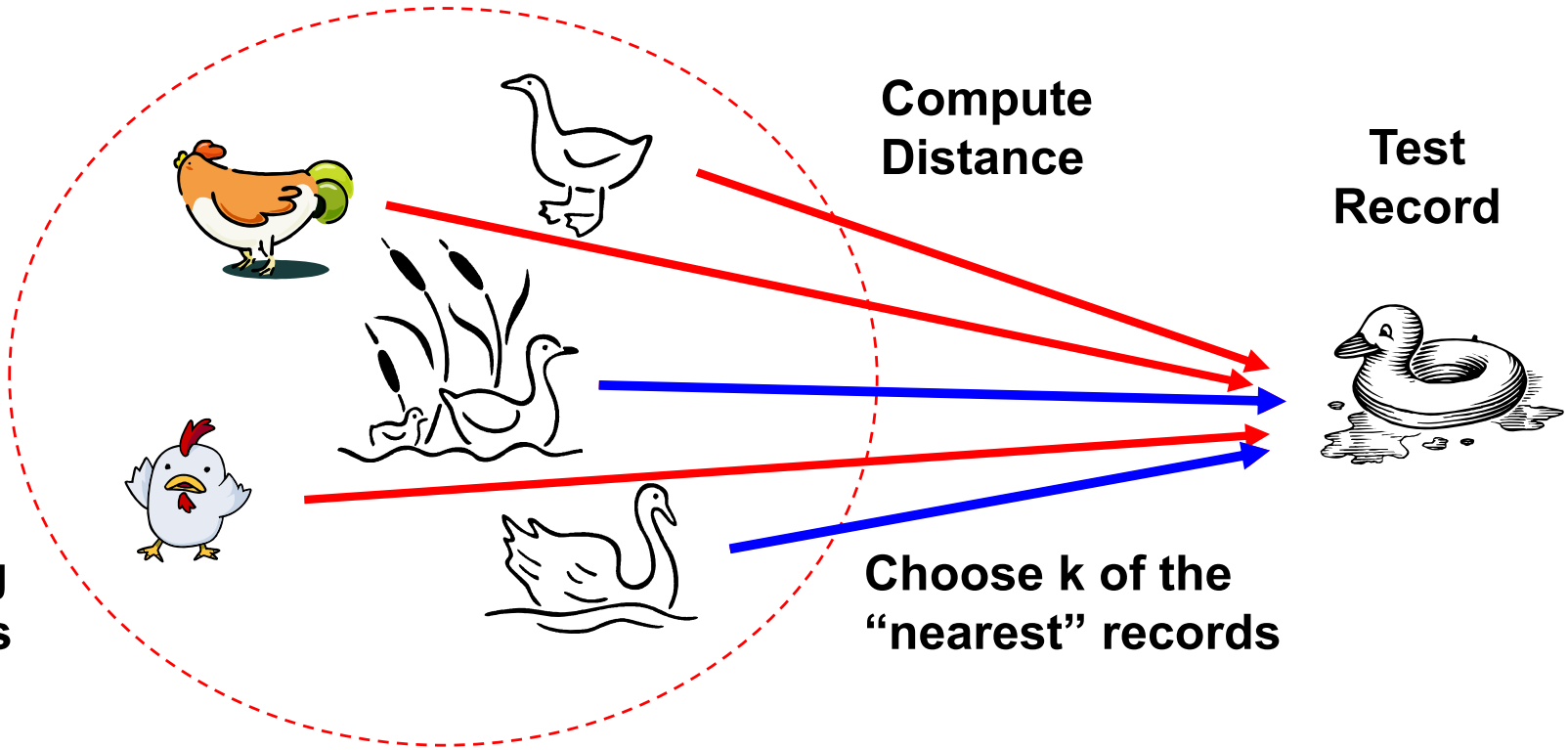
Nearest Neighbor Classifiers

- Eager Learners: Learn a model that maps the input attributes to the class label from the training data.
  - Decision Trees.
- Lazy Learners: Delay modeling the training data until classification of test examples is needed.
  - Rote Classifier: Memorize entire training data and perform classification only if the attributes of a test instance match one of the training examples exactly.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classifiers

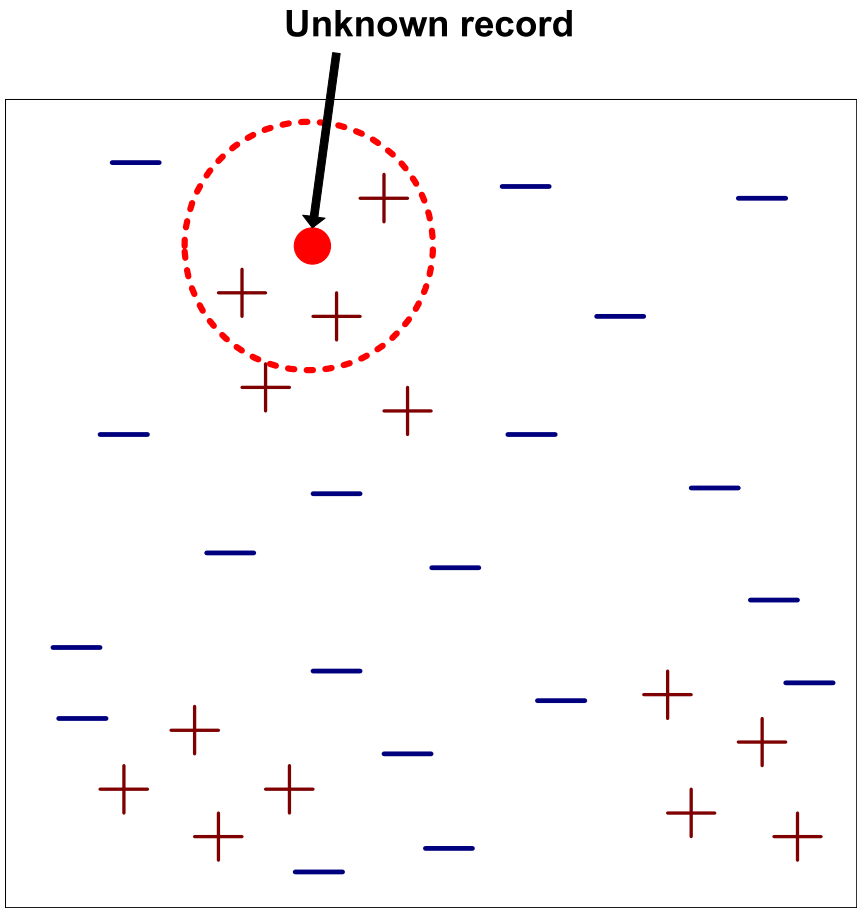
- Basic idea: "If it walks like a duck, quacks like a duck, then it's probably a duck."



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classifiers

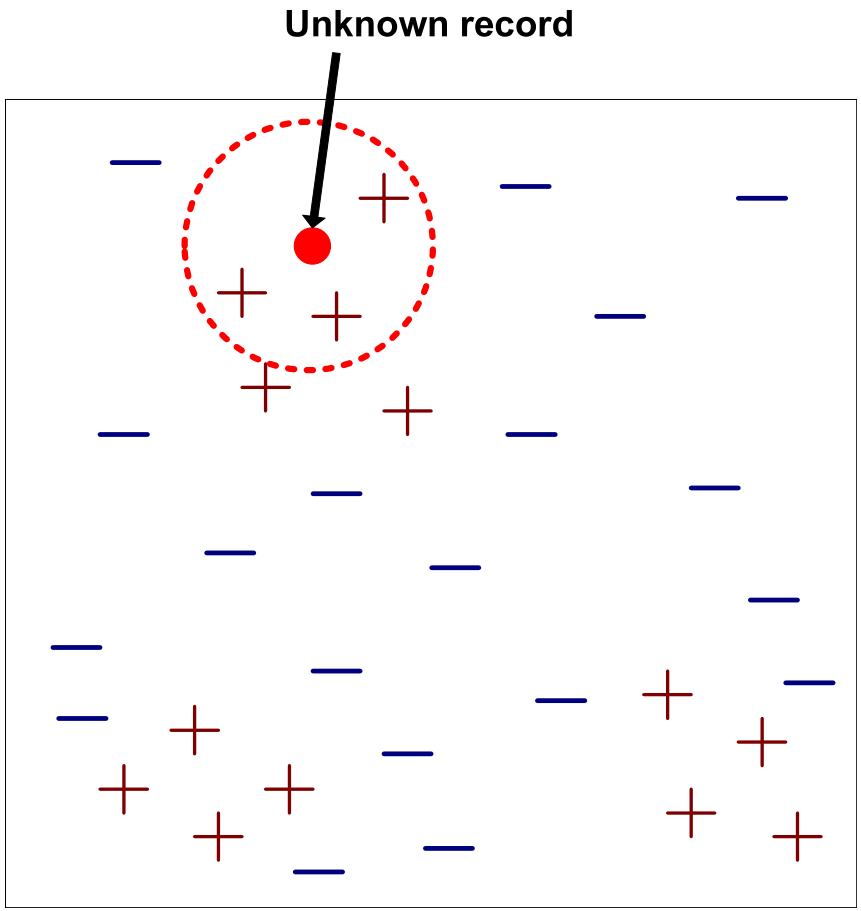
- Requires the following:
  - A set of labeled records.
  - Proximity metric to compute distance/similarity between a pair of records (e.g., Euclidean distance).
  - The value of *k*, the number of nearest neighbors to retrieve.
  - A method for using class labels of *k* nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote).



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Determine the Class Label of a Test Sample?

- Take the majority vote of class labels among the k-nearest neighbors.
- Weight the vote according to distance (e.g., with a weight factor, $w = \frac{1}{d^2}$).



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Choice of Proximity Measure Matters

- For documents, cosine is better than correlation or Euclidean.

1 1 1 1 1 1 1 1 1 1 1 0		0 0 0 0 0 0 0 0 0 0 0 1
	vs	
0 1 1 1 1 1 1 1 1 1 1 1		1 0 0 0 0 0 0 0 0 0 0 0

Euclidean distance = 1.4142 for both pairs, but the cosine similarity measure has different values for these pairs.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

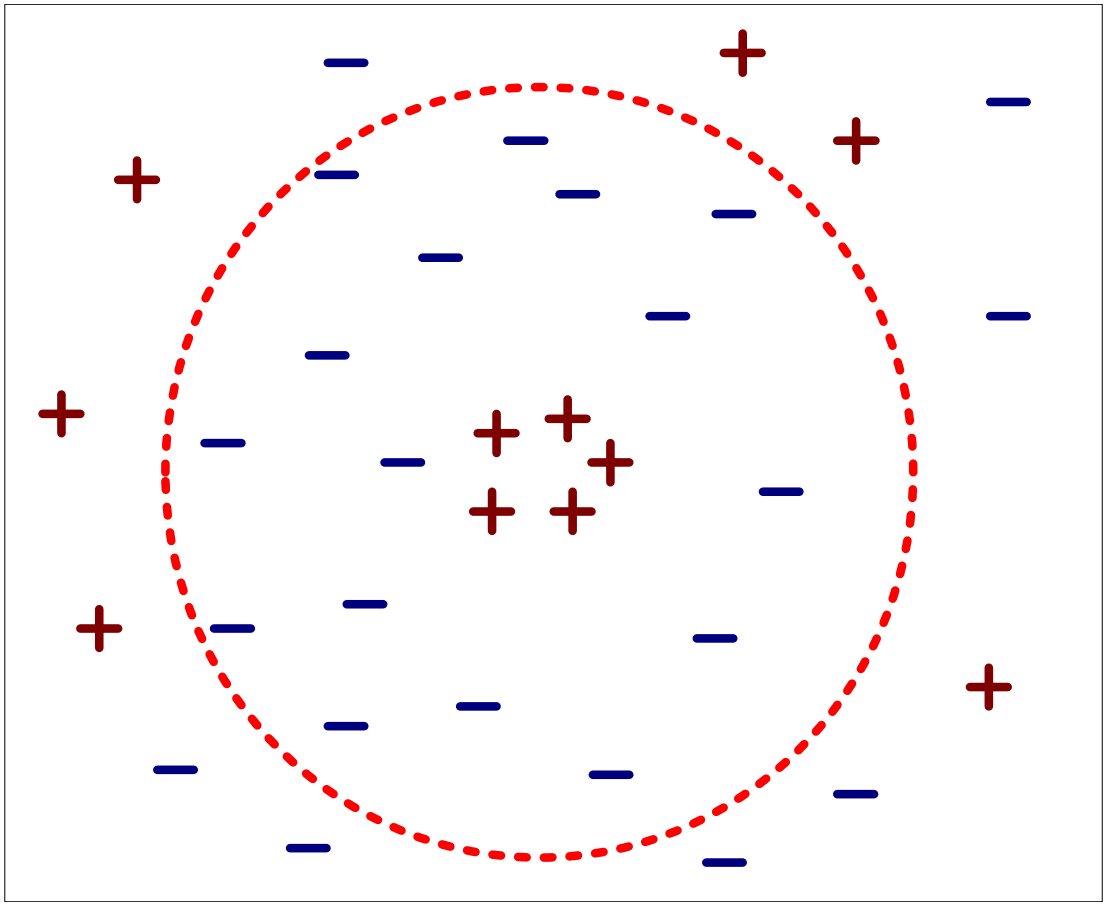
Nearest Neighbor Classification

- Data preprocessing is often required:
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes.
  - Example:
    - Height of a person may vary from 1.5 m to 1.8 m.
    - Weight of a person may vary from 90 lb to 300 lb.
    - Income of a person may vary from 10 K to 1 M.
  - Time series are often standardized to have
    0 means a standard deviation of 1.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classification

- Choosing the value of k:
  - If k is too small, sensitive to noise points.
  - If k is too large, neighborhood may include points from other classes.

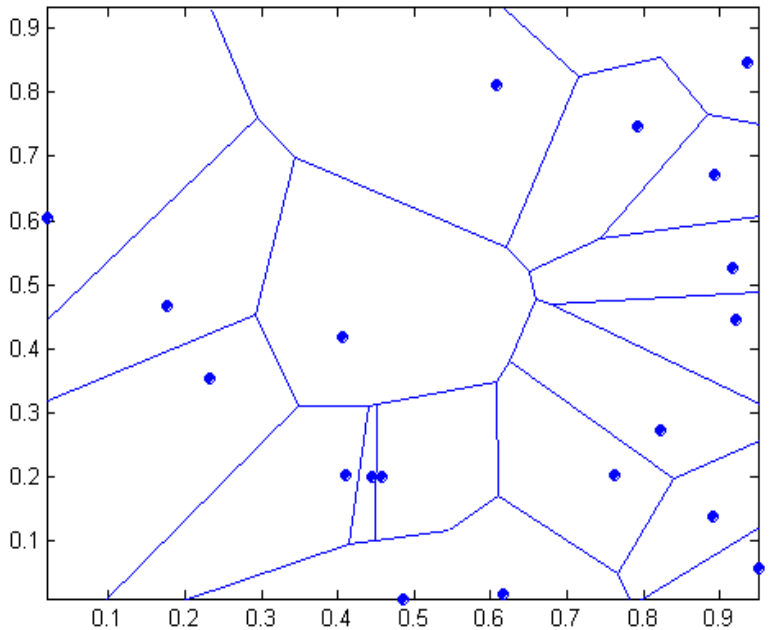


All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classification

- Nearest neighbor classifiers are local classifiers.
- They can produce decision boundaries of arbitrary shapes.

1-NN decision boundary is a Voronoi Diagram.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Classification Methods — Rule-Based Classifier

- Classify records by using a collection of "if...then..." rules.
- Rule: $(\text{Condition}) \rightarrow y$
  - Condition is a conjunction of tests on attributes.
  - y is the class label.
- Examples of classification rules:
  - $(\text{BloodType} = \text{Warm}) \wedge (\text{LayEggs} = \text{Yes}) \rightarrow \text{Birds}$.
  - $(\text{TaxableIncome} < 50\text{K}) \wedge (\text{Refund} = \text{Yes}) \rightarrow (\text{Evade} = \text{No})$.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Classification Methods — Naïve Bayes

- In many applications, the class label of a test record cannot be predicted with certainty even though its attribute set is identical to some of the training examples.
- This situation may arise because of noisy data or the presence of certain confounding factors that affect classification but are not included in the analysis.
- Naïve Bayes is based on the Bayes Theorem (with additional assumption on independence):

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}. \tag{1}$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Classification Methods — Support Vector Machines

- Support vector machines (SVMs), is a method for the classification of both linear and nonlinear data.
- It uses a nonlinear mapping to transform the original training data into a higher dimension.
- Within this new dimension, it searches for the linear optimal separating hyperplane (i.e., a "decision boundary" separating the tuples of one class from another).
- With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.
- The SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors).

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Other Classification Methods — Support Vector Machines

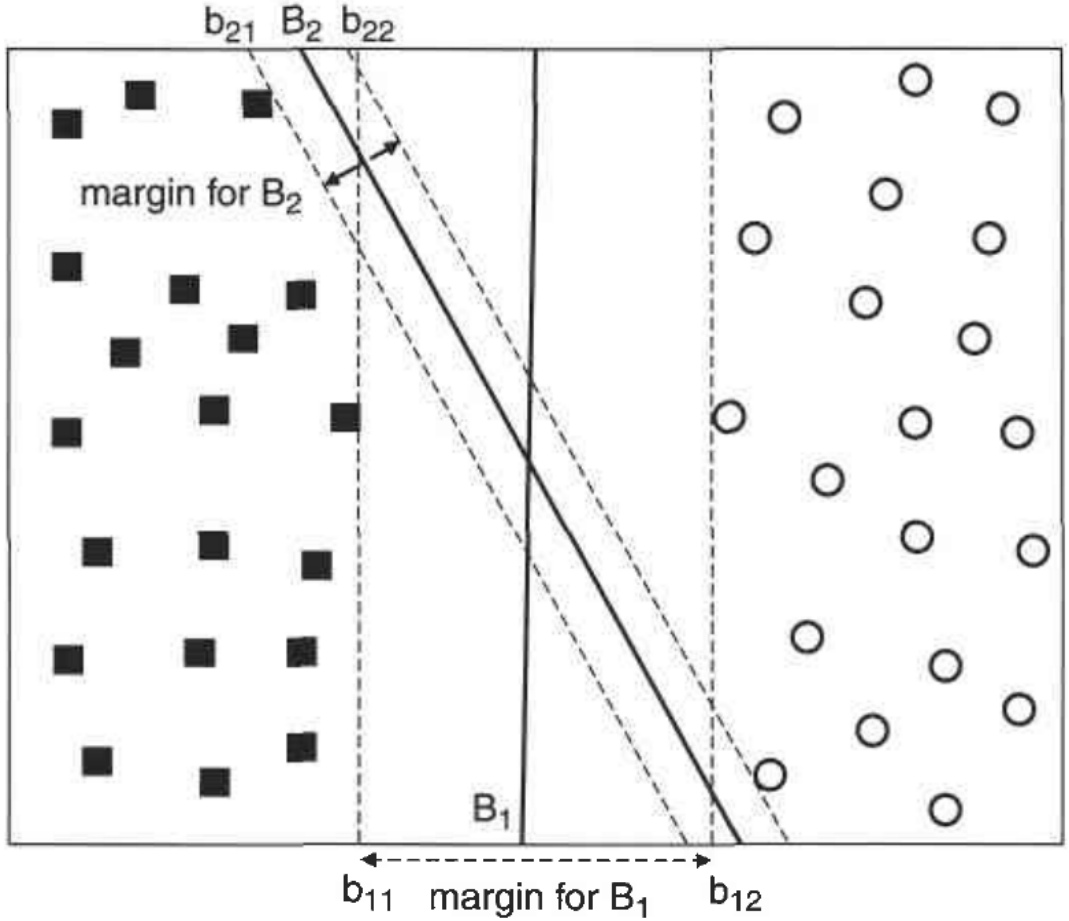


Figure 5.22. Margin of a decision boundary.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Classification Methods — Artificial Neural Networks

- Artificial neural networks (ANNs) or simply neural networks are inspired by biological neuronal networks.
- ANNs are comprised of abstract neurons that try to mimic real neurons at a very high level.
- ANNs can be described via a weighted directed graph $G = (V, E)$, with each node representing a neuron, and each directed edge representing a synaptic to dendritic connection.
- The weight of the edge denotes the synaptic strength.

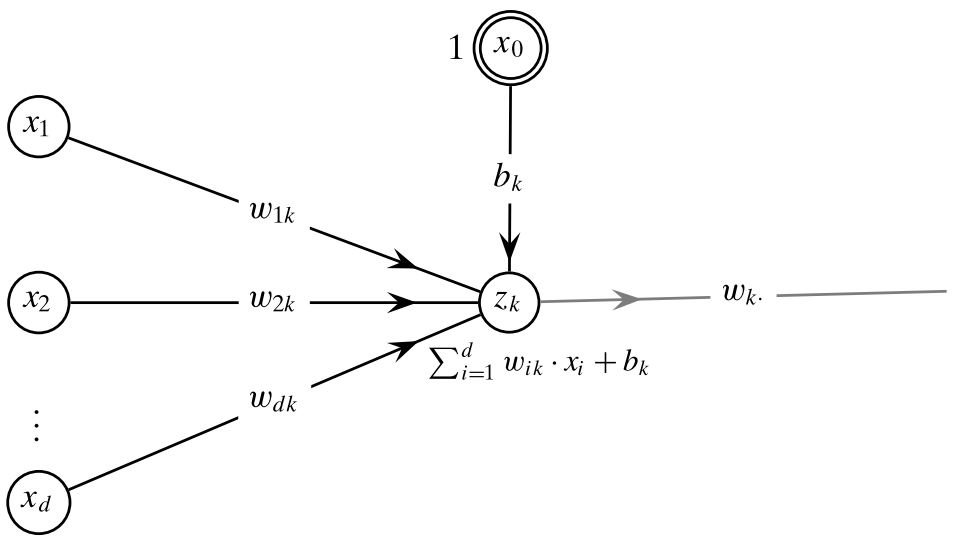


Figure 25.1. Artificial neuron: aggregation and activation.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

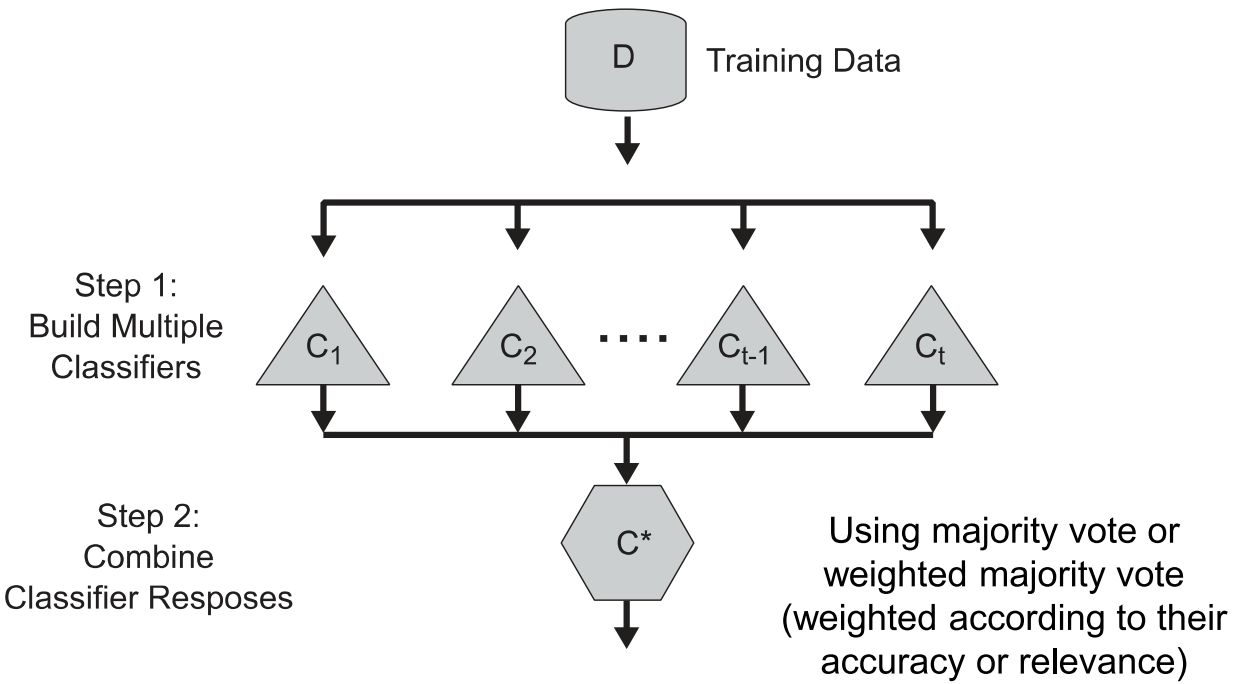
Other Classification Methods — Artificial Neural Networks

- ANNs are characterized by the type of activation function used to generate an output, and the architecture of the network in terms of how the nodes are interconnected.
- For example,
  - Is the graph is a directed acyclic graph or has cycles?
  - Is the graph is layered or not, and so on.
- It is important to note that a neural network is designed to represent and learn information by adjusting the synaptic weights.
- ANNs given enough hidden units and enough training samples, can closely approximate any function.

All from: 1. Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press. 2. Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Other Classification Methods — Ensemble Methods

- Ensemble methods improve classification accuracy by aggregating the predictions of multiple classifiers.
- An ensemble method constructs a set of base classifiers from training data and performs classification by taking a vote on the predictions made by each base classifier.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Classification
   - Evaluation in Classification
   - Nearest Neighbor Classifiers
   - Other Classification Methods

3. Web Usage Mining
   - Web Usage Data
   - Data Modeling
   - Data Mining

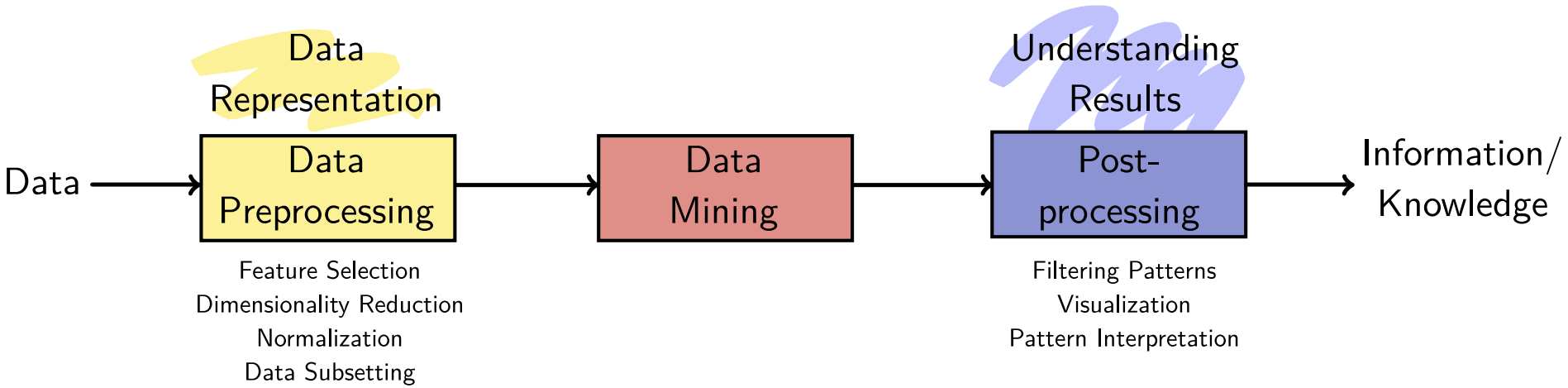
Web Usage Mining — Data

- Data from e-commerce, Web services, and Web-based information systems:
  - Volumes of Clickstream
  - Transaction Data
  - User Profile Data
- Analyzing such data can help organizations determine:
  - Value of clients.
  - Marketing strategies across products.
  - Effectiveness of promotional campaigns.
  - Optimize Web-based applications.
  - Personalized content.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Web Usage Mining — Data Mining Process

- Web usage mining process:
  1. Data Collection and Pre-Processing
  2. Pattern Discovery
  3. Pattern Analysis



All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Web Usage Mining — Data Mining Process

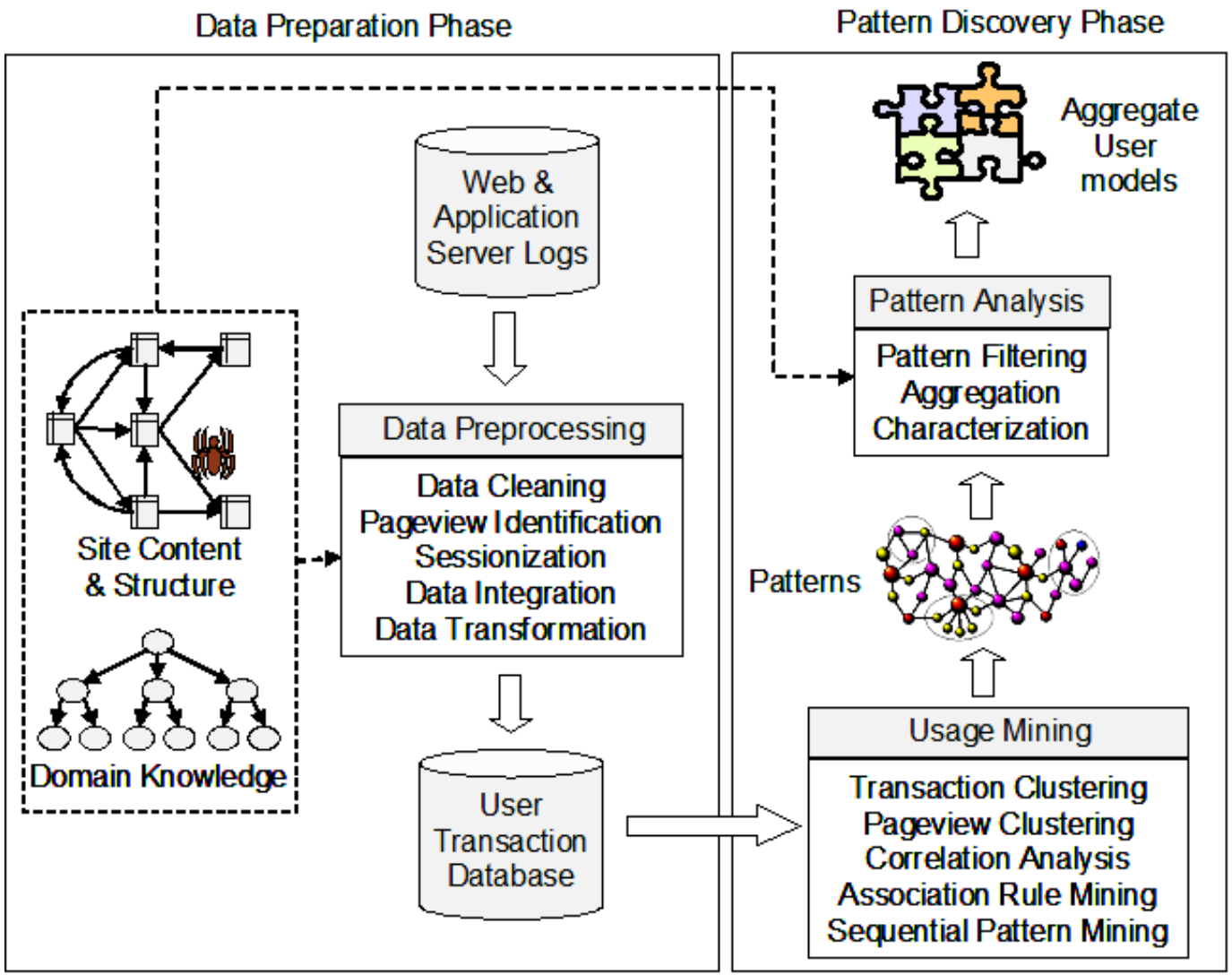


Fig. 12.1. The Web usage mining process

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Collection

- Web Usage Data:
  - Server Log Files (e.g., Web Server Access Logs and Application Server Logs)
  - Site Files
  - Metadata
  - Operational Databases
  - Application Templates
  - Domain Knowledge
  - Demographics Data
- Four primary groups of data sources:
  1. Usage Data (Server Access Logs)
  2. Content Data (HTML/XML pages)
  3. Structure Data (Hyperlink Structure as Sitemaps)
  4. User Data (User Profiles)

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Collection — Usage Data

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu **RESOURCE** Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu **IP ADDRESS** Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) **REFERRER FIELD** http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) **AGENT FIELD** http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Fig. 12.3. Portion of a typical server log

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Collection — Usage Data

- Usage data needs to be transformed and aggregated at different levels of abstraction for different analysis:
  1. Pageview — aggregation of resources representing user events (e.g., viewing a page).
  2. Session — sequence of pageviews by a single user during a single visit.
- Key steps:
  - Data Fusion and Cleaning
  - Pageview Identification
  - User Identification
  - Sessionization

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — Data Fusion and Cleaning

- Data cleaning is site-specific, involving tasks such as:
  - Removing extraneous references (e.g., style, graphics, and sound files).
  - Removing useless data fields (e.g., HTTP protocol).
  - Removing references due to crawler navigations.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — Pageview Identification

- Each pageview is a collection of resources representing a specific "user event," e.g.,
  - Clicking on a link.
  - Viewing a product page.
  - Adding a product to the shopping cart.
- Static single frame site: each HTML file has a one-to-one correspondence with a pageview.
- Multi-framed sites: several files make up a given pageview.
- Dynamic sites: combination of static templates and content generated by application servers.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — User Identification

- Authentication mechanisms.
- Client-side cookies.
- Using a combination of IP addresses and other information such as user agents and referrers.

Method	Description	Privacy Concerns	Advantages	Disadvantages
IP Address + Agent	Assume each unique IP address / Agent pair is a unique user	Low	Always available. No additional technology required.	Not guaranteed to be unique. Defeated by rotating IPs.
Embedded Session Ids	Use dynamically generated pages to associate ID with every hyperlink	Low to medium	Always available. Independent of IP addresses.	Cannot capture repeat visitors. Additional overhead for dynamic pages.
Registration	User explicitly logs in to the site.	Medium	Can track individuals not just browsers	Many users won't register. Not available before registration.
Cookie	Save ID on the client machine.	Medium to high	Can track repeat visits from same browser.	Can be turned off by users.
Software Agents	Program loaded into browser and sends back usage data.	High	Accurate usage data for a single site.	LIkely to be rejected by users.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — User Identification

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Fig. 12.4. Example of user identification using IP + Agent

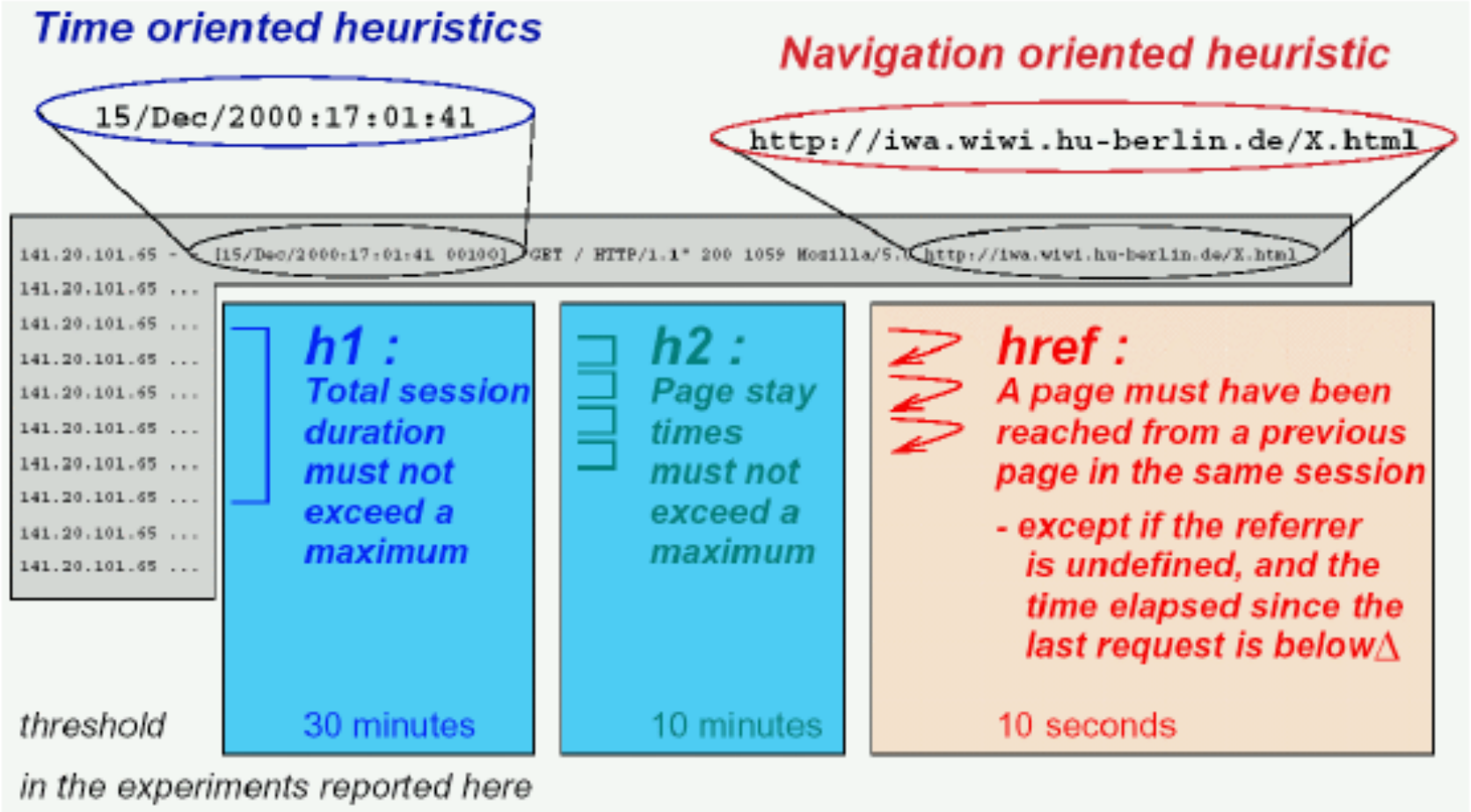
All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — Sessionization

- Sessionization: segmenting user activity record of each user into sessions, each representing a single visit to the site.
- Without mechanisms such as embedded session ids must rely on heuristics methods for sessionization.
- Heuristic categories:
  1. Time-Oriented: global or local time-out estimate to distinguish between consecutive sessions.
  2. Structure-Oriented: static site structure or implicit linkage structure captured in the referrer fields of the server logs.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — Sessionization



All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — Sessionization

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Fig. 12.5. Example of sessionization with a time-oriented heuristic

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — Sessionization

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:26	1.2.3.4	F	C

Session 2

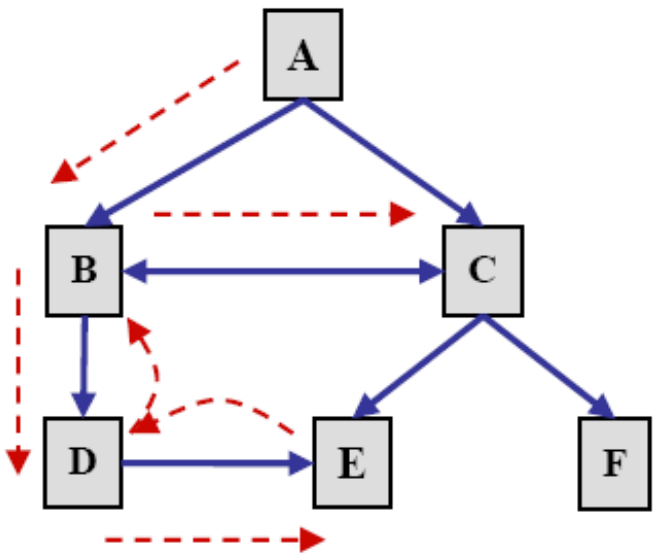
1:15	1.2.3.4	A	-
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Fig. 12.6. Example of sessionization with the h-ref heuristic

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Pre-processing — Path Completion

- Client- or proxy-side caching can often result in missing access references to those pages or objects that have been cached.
- Effective path completion requires extensive knowledge of the link structure within the site.
- Referrer information in server logs can also be used in disambiguating the inferred paths.



User's actual navigation path:

A → B → D → E → D → B → C

What the server log shows:

URL	Referrer
A	--
B	A
D	B
E	D
C	B

Fig. 12.7. Missing references due to caching.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Modeling

- Data pre-processing results in a set of n pageviews: $P = \{p_1, p_2, ..., p_n\}$.
- Also, a set of m user transactions: $T = \{t_1, t_2, ..., t_m\}$, where $t_i \in T$ is a subset of P.
- Pageviews are semantically meaningful entities to which mining tasks are applied (such as pages or products).
- Set of all user transactions can be viewed as an $m \times n$ user-pageview matrix (also called the transaction matrix), denoted by UPM.

Pageviews

Sessions / users	A	B	C	D	E	F
user0	15	5	0	0	0	185
user1	0	0	32	4	0	0
user2	12	0	0	56	236	0
user3	9	47	0	0	0	134
user4	0	0	23	15	0	0
user5	17	0	0	157	69	0
user6	24	89	0	0	0	354
user7	0	0	78	27	0	0
user8	7	0	45	20	127	0
user9	0	38	57	0	0	15

Amount of time (e.g., in seconds) a user spents on a particular page.

Fig. 12.8. An example of a user-pageview matrix (or transaction matrix)

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Modeling

- We can also integrate other knowledge sources, such as semantic information from the Web page contents for mining.
- Generally, the textual features from the content of Web pages represent the underlying semantics of the site.
- Each pageview p can be represented as a r-dimensional feature vector, where r is the total number of extracted features (words or concepts) from the site in a global dictionary.
- This vector, denoted by p, can be given by:
  $p = (fw^p(f_1), fw^p(f_2), ..., fw^p(f_r))$, where, $fw^p(f_j)$ is the weight of the jth feature (i.e., f_j) in pageview p, for $1 \leq j \leq r$.
- For the whole collection of pageviews in the site, we then have an $n \times r$ pageview-feature matrix $PFM = \{p_1, p_2, ..., p_n\}$.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Modeling

- Goal of transformation: represent user sessions / user profiles as vectors of semantic / textual / concept features rather than as vectors of pageviews.
- Thus, user's session reflects the significance of various concepts / context features relevant to user's interaction.
- Formally, transformation is the multiplication of the user-pageview matrix UPM with the pageview-feature matrix PFM.
- The new matrix is: $TFM = \{t_1, t_2, ..., t_m\}$, where each t_i is a r-dimensional vector over the feature space.

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Modeling

	A.html	B.html	C.html	D.html	E.html
user1	1	0	1	0	1
user2	1	1	0	0	1
user3	0	1	1	1	0
user4	1	0	1	1	1
user5	1	1	0	0	1
user6	1	0	1	1	1

User-Pageview Matrix

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Modeling

	A.html	B.html	C.html	D.html	E.html
web	0	0	1	1	1
data	0	1	1	1	0
mining	0	1	1	1	0
business	1	1	0	0	0
intelligence	1	1	0	0	1
marketing	1	1	0	0	1
ecommerce	0	1	1	0	0
search	1	0	1	0	0
information	1	0	1	1	1
retrieval	1	0	1	1	1

Term-Pageview Matrix

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Modeling

	web	data	mining	business	intelligence	marketing	ecommerce	search	information	retrieval
user1	2	1	1	1	2	2	1	2	3	3
user2	1	1	1	2	3	3	1	1	2	2
user3	2	3	3	1	1	1	2	1	2	2
user4	3	2	2	1	2	2	1	2	4	4
user5	1	1	1	2	3	3	1	1	2	2
user6	3	2	2	1	2	2	1	2	4	4

Content Enchanced Transaction Matrix

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Classification
   - Evaluation in Classification
   - Nearest Neighbor Classifiers
   - Other Classification Methods

3. Web Usage Mining
   - Web Usage Data
   - Data Modeling
   - Data Mining

- Clustering using a standard clustering algorithm such as k-means, results in three clusters of user transactions.
- This example, indicates that the resulting user segment is clearly interested in items B and F and to a lesser degree in item A.

		A	B	C	D	E	F
Cluster 0	user 1	0	0	1	1	0	0
	user 4	0	0	1	1	0	0
	user 7	0	0	1	1	0	0
Cluster 1	user 0	1	1	0	0	0	1
	user 3	1	1	0	0	0	1
	user 6	1	1	0	0	0	1
	user 9	0	1	1	0	0	1
Cluster 2	user 2	1	0	0	1	1	0
	user 5	1	0	0	1	1	0
	user 8	1	0	1	1	1	0

Aggregated Profile for Cluster 1	
Weight	Pageview
1.00	B
1.00	F
0.75	A
0.25	C

Fig. 12.11. Derivation of aggregate profiles from Web transaction clusters

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

Data Mining

- Clustering enhanced transaction matrix may reveal segments of users that have common interests in different concepts as indicated from their navigational behaviors.

	web	data	mining	business	intelligence	marketing	ecommerce	search	information	retrieval
user1	2	1	1	1	2	2	1	2	3	3
user2	1	1	1	2	3	3	1	1	2	2
user3	2	3	3	1	1	1	2	1	2	2
user4	3	2	2	1	2	2	1	2	4	4
user5	1	1	1	2	3	3	1	1	2	2
user6	3	2	2	1	2	2	1	2	4	4

Content Enchanced Transaction Matrix

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.

- Preferences of user are matched on the left-hand side X of each rule (e.g., $X \rightarrow Y$), and the items on the right-hand side of the matching rules can be used as potential recommendations.
- This also enables Web sites to organize the site content more efficiently, or to provide effective cross-sale product recommendations.

Transactions
A, B, D, E
A, B, E, C, D
A, B, E, C
B, E, B, A, C
D, A, B, E, C

Size 1		Size 2		Size 3		Size 4	
Itemset	Supp.	Itemset	Supp.	Itemset	Supp.	Itemset	Supp.
A	5	A,B	5	A,B,C	4	A,B,C,E	4
B	5	A,C	4	A,B,E	5		
C	4	A,E	5	A,C,E	4		
E	5	B,C	4	B,C,E	4		
		B,E	5				
		C,E	4				

Fig. 12.14. Web transactions and resulting frequent itemsets (minsup = 4)

All from: Bing Liu. "Web Data Mining", 2nd Edition, 2011, Springer.