Data Mining and Its Applications

Jiye Li
School of Computer Science
University of Waterloo
CS785 Talk, October 9, 2009

Agenda

- What is Data Mining
- What are the main approaches in Data Mining
  - Association rules
  - Classifications
- Applications of Data Mining

What is Data Mining

- Real world large data set
- Not enough knowledge
- Data mining (knowledge discovery in databases - KDD)
  - "The process of analyzing data from different perspectives and summarizing it into interesting (non-trivial, implicit, previously unknown and potentially useful) information." (Data Mining: Concepts and Techniques, 2nd ed., March 2006. ISBN 1-55860-901-6.)

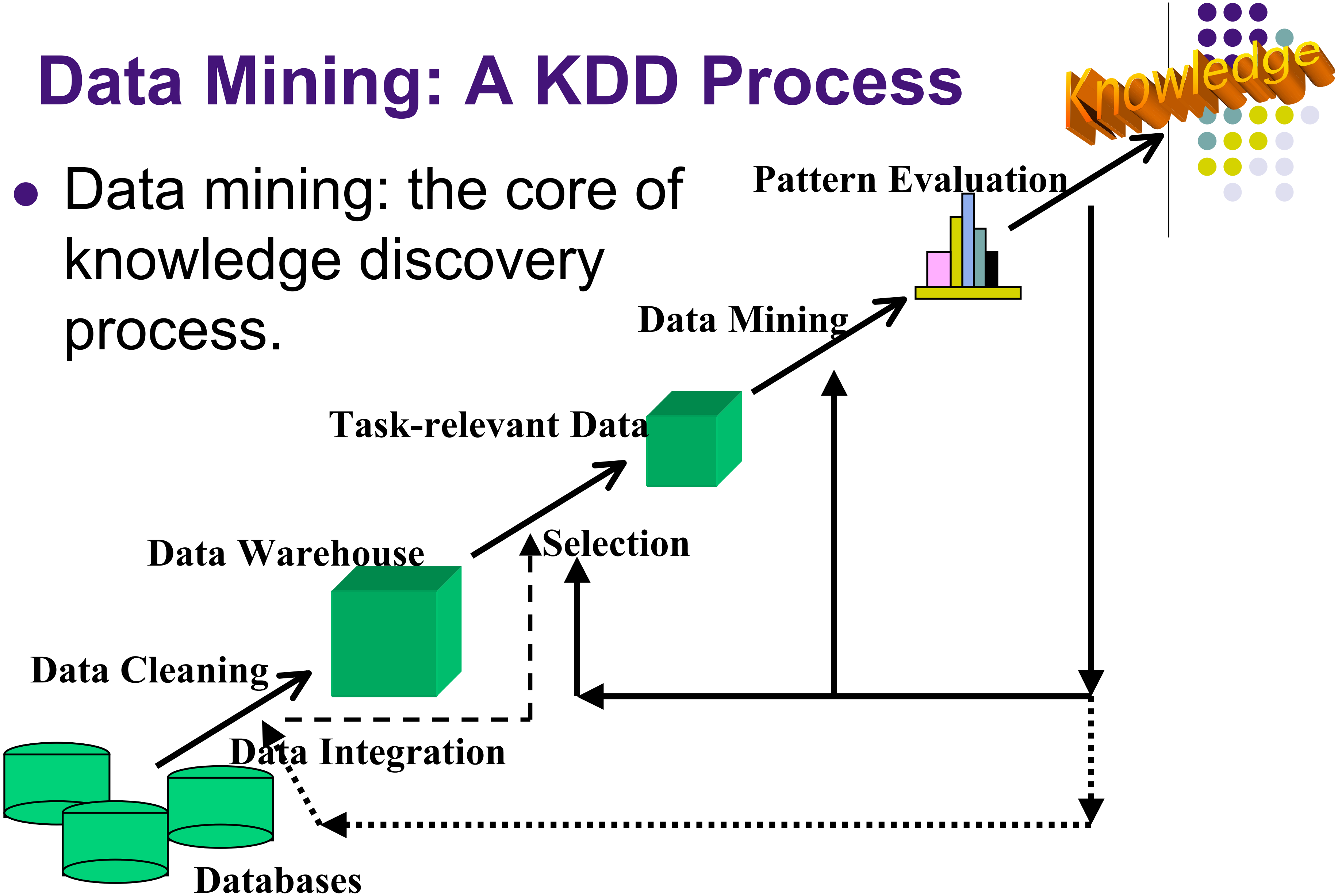
A brief history of data mining

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 1. Introduction. (http://www.cs.sfu.ca/~han/dmbook)

Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 1. Introduction. (http://www.cs.sfu.ca/~han/dmbook)

What can Data Mining do

- Decision Support
  - Market analysis
  - Revenue forecast, risk management
  - Fraud detection
- Medical Diagnosis
- Recommender System (books, movies, …)
- Web applications
- Blog mining …

Data Mining Algorithms

- Search for interesting patterns
- Well-known algorithms
  - Association rules
  - Classification rules
    - Decision Tree
    - Naïve Bayes
  - Other (from rough sets theory, text mining, …)

Association Rules

- Association Rules
  - Find frequent patterns and associations in transaction data
  - Help stores displaying items that are likely to be sold together
  - Business support, recommender system, medical diagnosis, etc.

Association Rules

- Introducted by Agrawal in 1994

For an association rule $\alpha \Rightarrow \beta$

$$\text{Support} = \frac{|\alpha \cup \beta|}{|D|}$$

$$\text{Confidence} = \frac{|\alpha \cup \beta|}{|\alpha|}$$

Association Rules

- To find shopping behaviors of customers
- Sample transaction list

Customer A:	bread, cheese, ketchup, mustard
Customer B:	juice, bread, cheese
Customer C:	cheese, orange, banana

$$bread \rightarrow cheese$$

(Support = 66%, Confidence =100%)

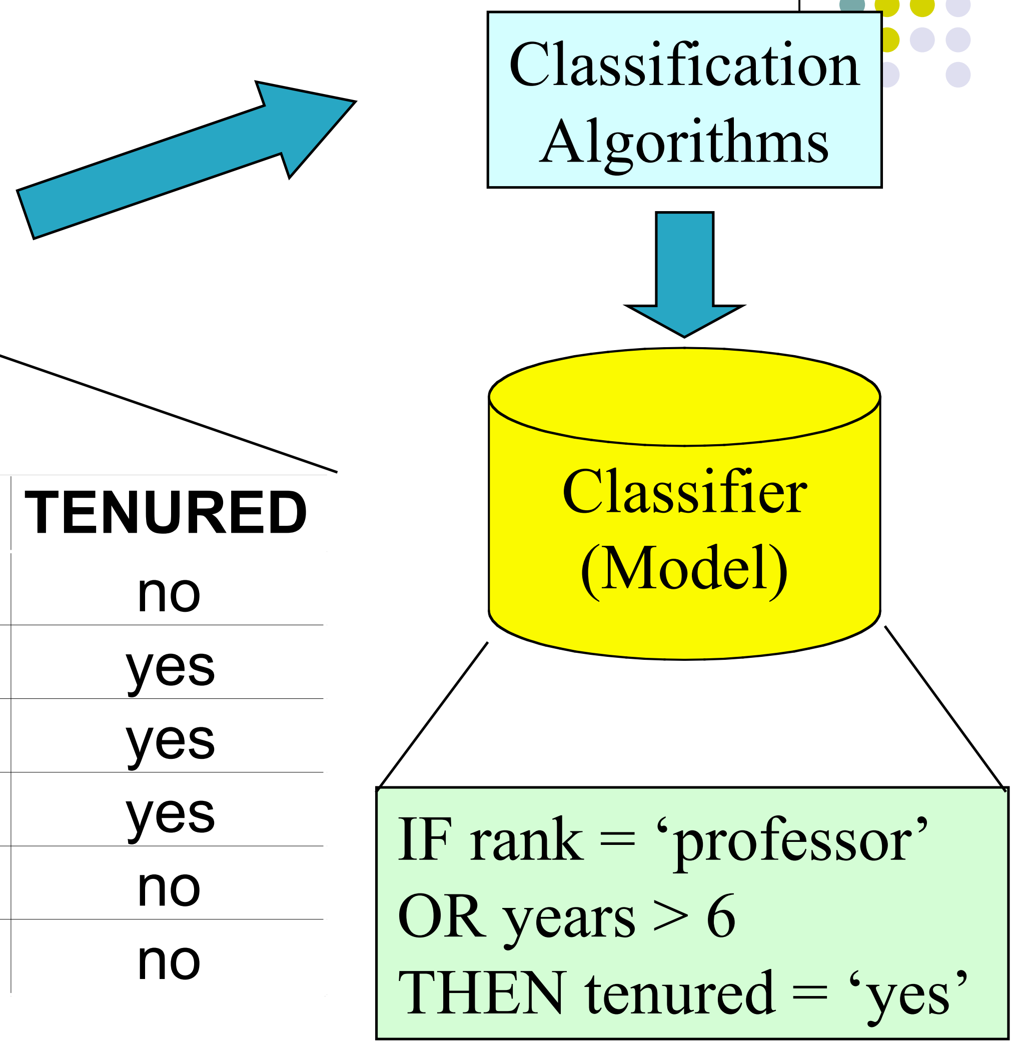
Association rules

- Problems
  - Huge amount of rules are generated
  - Difficult to extract useful rules
- Solutions
  - Post-processing
    - interestingness measures

Classification

- Predict categorical class labels
- Classifies data
  - Construct a model based on a training set
  - Classify new data based on the model

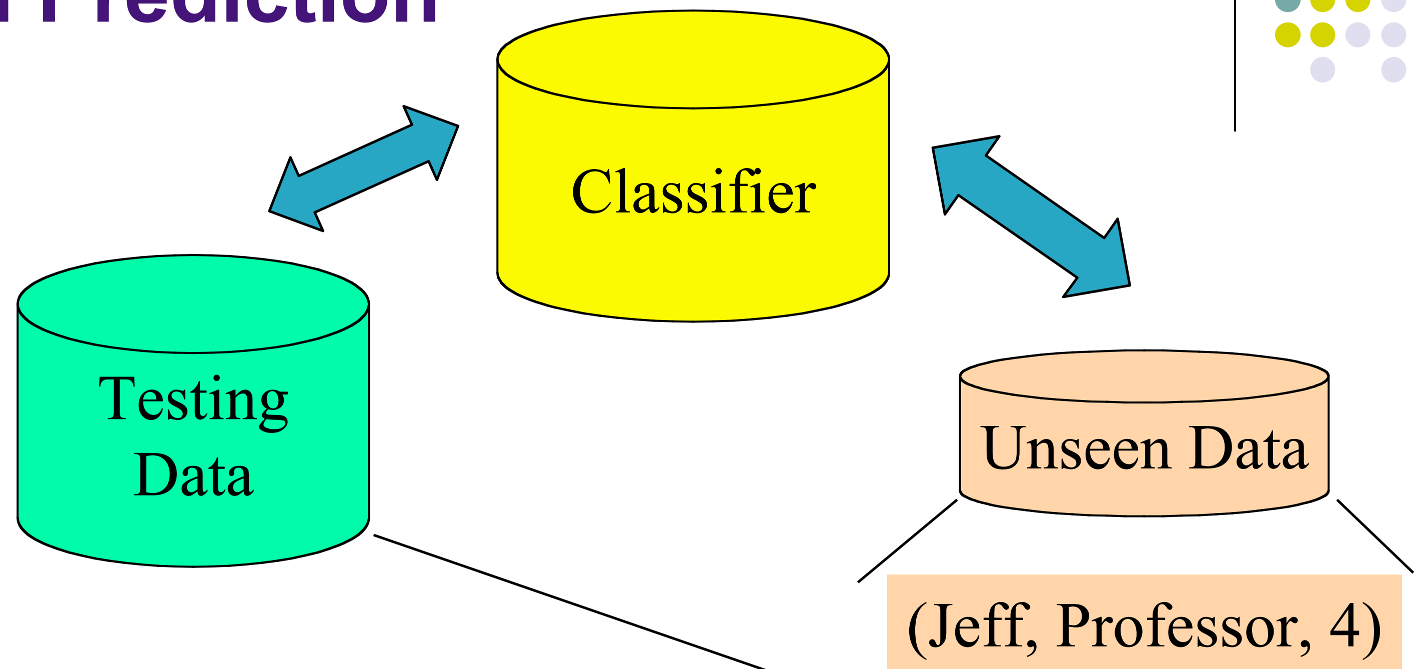
Classification Process: Model Construction



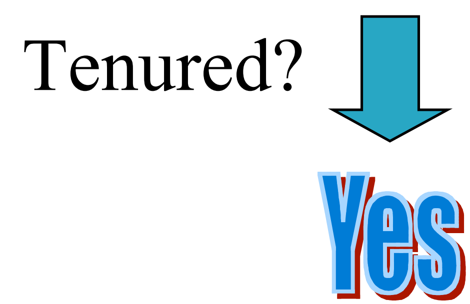
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 7. Classification and Prediction. (http://www.cs.sfu.ca/~han/dmbook)

Classification Process (2): Use the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes



This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 7. Classification and Prediction (http://www.cs.sfu.ca/~han/dmbook)

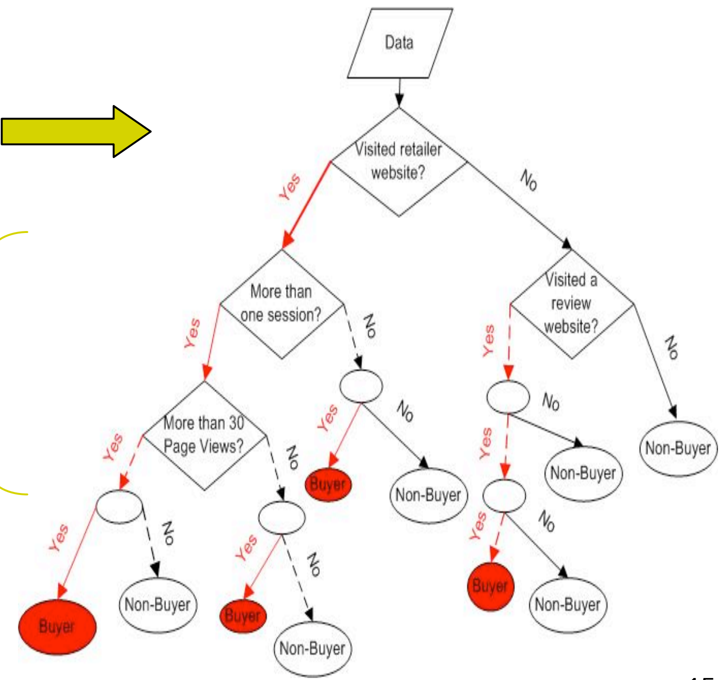
Decision Tree

- Classification
  - Simple to use
  - Easy to understand

Nodes represent features

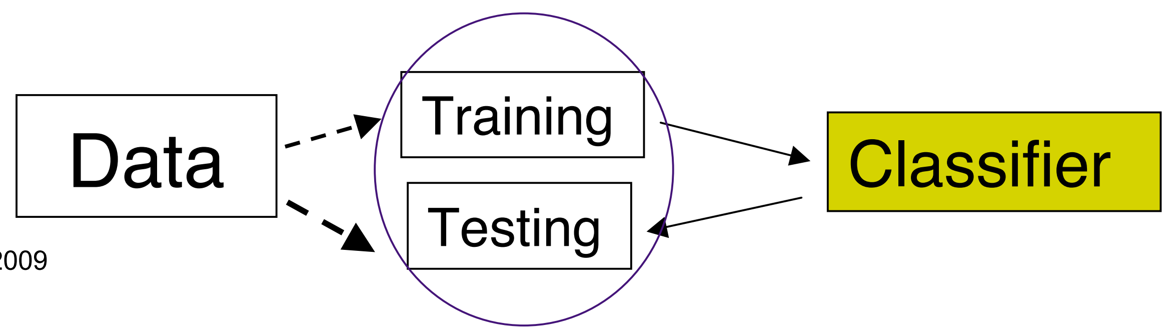
Branches represent conjunctions of features that lead to the predicted decisions

Leaves represent predicted decisions



Decision Tree

- Problems
  - Overfitting
    - Training and testing on the same data, achieve very high prediction precision
  - Branches are too deep
    - Branches containing 4 or 5 levels of leaves are too deep
- Solutions
  - Pruning and Post-processing
  - Cross-validations (training on unseen cases)



Decision Tree

- Sample Decision Table
  - T = (C, D)
  - C is condition attribute sets (feature sets)
  - D is decision attribute sets (buyer, non-buyer)

Panel ID	**Feature 1 (Whether searched “laptop” on google before purchase)**	**Feature 2 (Whether visited online manufacturer store before purchase)**	**Feature 3 (Whether made a purchase last month)**	**… …**	**Feature n (whether visited a review website before purchase)**	**Decision Attribute (Whether is a buyer)**
1	Yes	No	Yes	…	Yes	Buyer
2	No	No	No	…	No	Non-buyer
3	Yes	No	No	..	Yes	Non-buyer
…	…	…	…	…	…	…
83,635	No	No	Yes	…	No	Non-buyer

Naïve Bayes Classifier

- Given training data *D, posteriori probability of a hypothesis h, P(h|D)* follows the Bayes theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- MAP (maximum posteriori) hypothesis

$$h_{MAP} \equiv \underset{h \in H}{\arg\max}\, P(h|D) = \underset{h \in H}{\arg\max}\, P(D|h)P(h).$$

- Assume attributes are conditionally independent
  - $P(x_1,\ldots,x_k|C) = P(x_1|C)\cdot\ldots\cdot P(x_k|C)$

This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 7. Classification and Prediction (http://www.cs.sfu.ca/~han/dmbook)

Play-tennis example: estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14	
P(n) = 5/14	

outlook	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 2/5**
windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 7. Classification and Prediction (http://www.cs.sfu.ca/~han/dmbook)

Play-tennis example: classifying X

- An unseen sample
  - X = <rain, hot, high, false>
- $P(X|p) \cdot P(p) = P(rain|p) \cdot P(hot|p) \cdot P(high|p) \cdot P(false|p) \cdot P(p) = 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$
- $P(X|n) \cdot P(n) = P(rain|n) \cdot P(hot|n) \cdot P(high|n) \cdot P(false|n) \cdot P(n) = 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$
- Sample X is classified in class n (don't play)

This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 7. Classification and Prediction (http://www.cs.sfu.ca/~han/dmbook)

Logistic Regression

- Statistical regression model for binary dependent variables (e.g., buyer or non-buyer)
- Estimate the probability of a certain event occurring by measuring the predictive capabilities of the independent variables (features)

$$\ln\left(\frac{P}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

$$P = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n}}$$

- Produce the probability of product purchase
- Use as cutoff to classify buyer vs. non-buyer

Rough Sets Approach

- Proposed by Pawlak in 1980's
- Knowledge Discovery, Data Analysis, Medical Diagnoses, …
- Reduction of Knowledge
  - "A reduct of knowledge is its essential part, which suffices to define all basic concepts occurring in the considered knowledge, whereas the core is in a certain sense its most important part." – in *Pawlak, "Rough Sets", 1991.*

Rough Sets Theory

Decision Table T= (U, C, D)

Attribute 1	Attribute 2	…	Attribute 5	…	Attribute n	Decision Attributes

Given a reduct = {Attribute 1, Attribute 2, Attribute 5}

How to generate rules from reducts?

Sample Rule: $\boxed{\text{Attribute 1, Attribute 2, Attribute 5}}$ => Decision Attribute

TFIDF(Term frequency-inverse document frequency)

- Sample Ranked Terms

Apparel	Automotives	Computer Hardware	Watch and Jewelry
Granby	rotonda	Dell	Seiko
Coupon	Civic	Laptop	watches
Centreville	Eps	Pc	ebay
Coupons	Ifinder	Memory	movado
Shirts	Altima	hp	overstock.com
Wrightsville	Motorcycle	Computer	watche
Clothing	Airbag	Compaq	Xbox
Pajamas	turbonator	Notebook	Timex
Transat	Vlovo	Pentium	Watchband
shirt	Nissan	Acer	Necklaces

Term Significance captures more behavior terms instead of syntactically related terms.

Challenges in Data Mining

- Feature selection and User modeling
  - extract prominent features from data
- High performance
  - rapid response to assist users
- Understandability of patterns
- Interpretation
- Accommodate data types beyond numeric: e.g., multimedia

Case Study -
web personalization

Larger Motivation: automatic personalization

I wish to see more deals on digital camera …

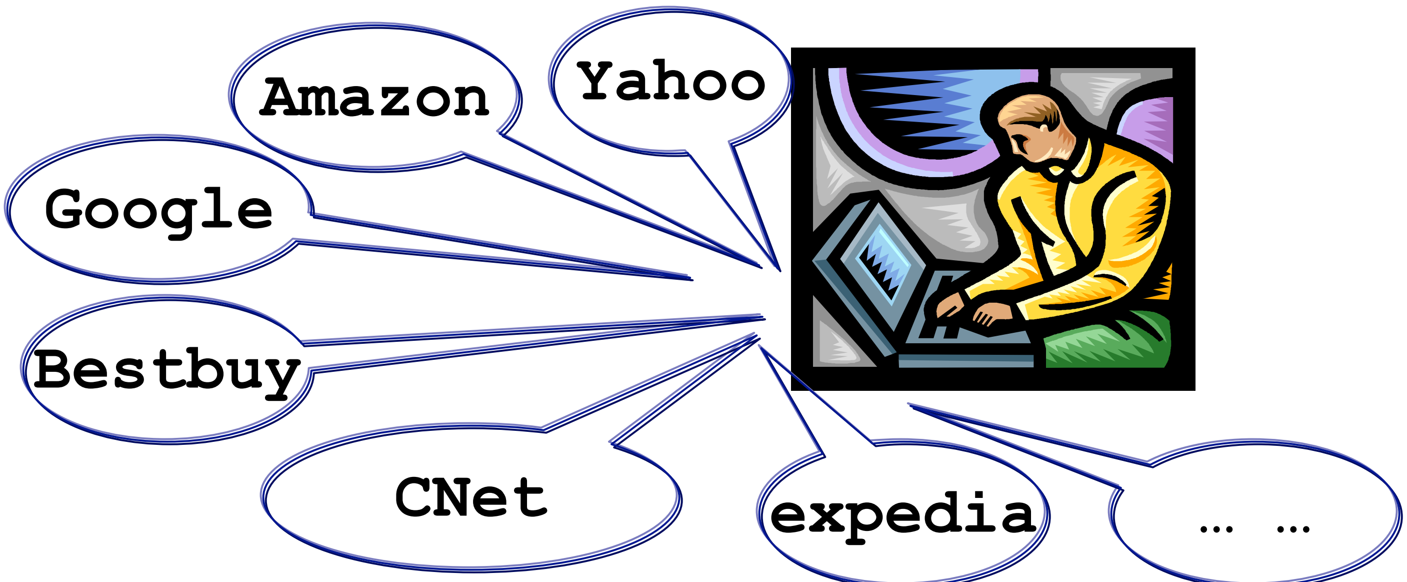
- This is my pc/laptop.
- **But**, am I using someone else's homepage??

http://hp.aol.com

Motivation

- Most existing web personalization systems rely on site-centric user data (user's behaviors on a specific site).
- We use a dataset supplied by a major audience measurement company that represents a complete **user-centric view** of clickstream behavior.

User-centric clickstream data



User-Centric vs. Site-Centric

- How is the data collected
- What is considered in a session
  - For Example

Google → hp → Google → Google → hp

↑ Purchase

  - Site-Centric data

$$(google_{time1}, google_{time2}, google_{time3})$$

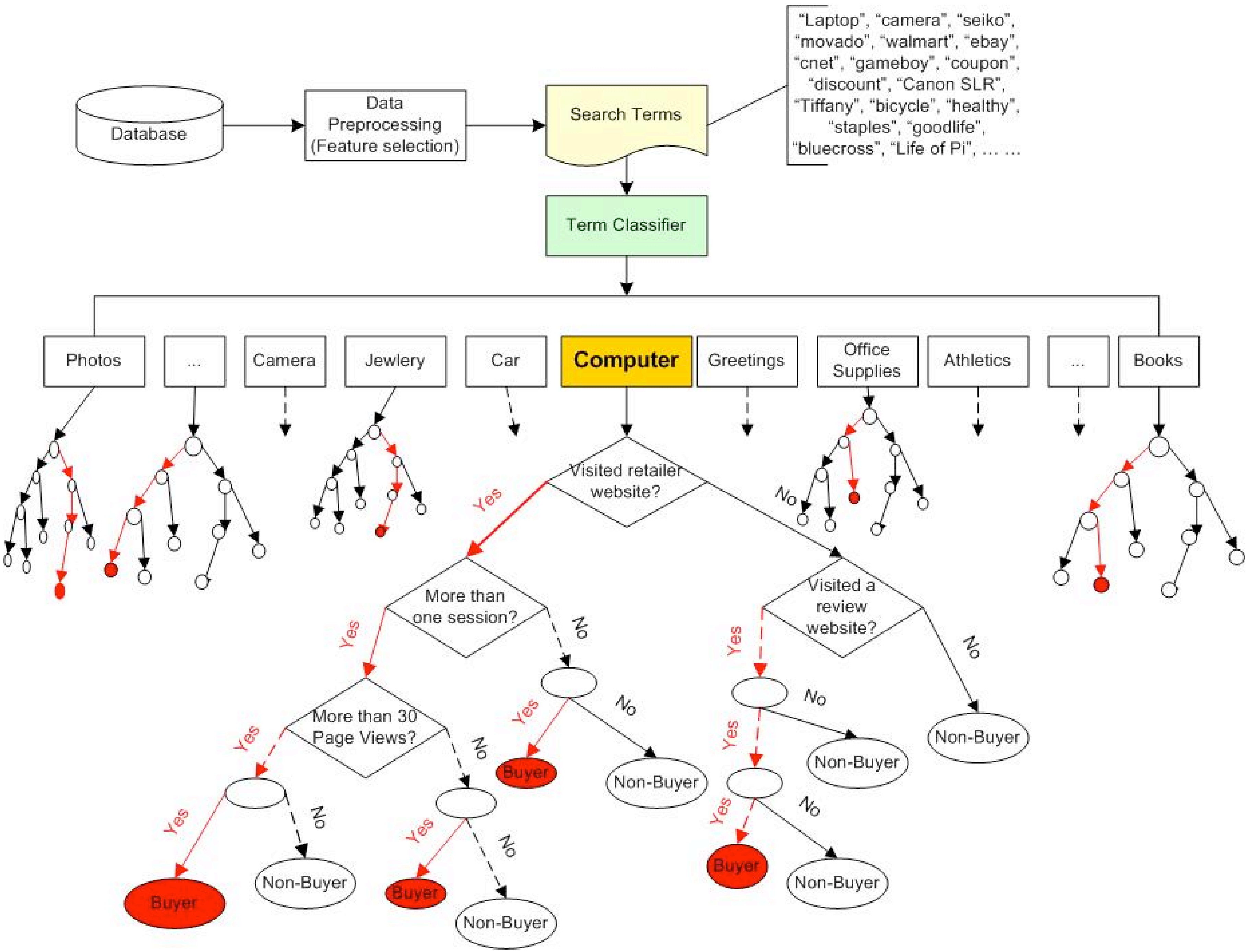
  - User-Centric data

$$(google_{time1}, HP_{time1}, google_{time2}, google_{time3}, HP_{time2})$$

Purchase

Motivation

- Personalize Users' Online Experiences
  - Predicting specific product category level purchases at any website
  - Developing algorithms for personalization based on user-centric behavioral data (Web browsing, Search, Purchase patterns, etc.)
  - Learning models of the user's probability of purchase within a time window
  - Quantifying the advantages of user-centric approaches to site-centric approaches



Database
Data Preprocessing (Feature selection)
Search Terms
"Laptop", "camera", "seiko", "movado", "walmart", "ebay", "cnet", "gameboy", "coupon", "discount", "Canon SLR", "Tiffany", "bicycle", "healthy", "staples", "goodlife", "bluecross", "Life of Pi", … …
Term Classifier
Photos
...
Camera
Jewlery
Car
Computer
Greetings
Office Supplies
Athletics
...
Books
Visited retailer website?
Yes
No
More than one session?
Visited a review website?
More than 30 Page Views?
Buyer
Non-Buyer

Experimental Data

- Experimental Data
  - Collected over 8 months amount to approximately 1 terabyte from more than 100,000 households (November 2005 ~ June 2006)
  - Clickstream data (URLs with timestamps for each panelist)
  - Retail transactional data (100+ leading online shopping destinations and retailer sites)
  - Travel transactional data (air, hotel, car and package)
  - Search terms (top search engines such as Google and Yahoo and comparison shopping sites such as Amazon and Bestbuy)

Feature Construction

Sample Decision Table with one feature

Panel ID	Condition Attribute (Feature)	Decision Attribute
ID	Whether searched “laptop” on google before purchase	buyer/ non-buyer

Panel ID	Condition Attribute (Feature)	Decision Attribute
1	Yes	Buyer
2	Yes	Non-buyer
…	…	…
83,635	Yes	Buyer

- Studying the predicting abilities for each feature
- Experiment
  - Data: December 2005 data (83,635 users, 10 features)
  - Preprocessing
    - Decision Table T = (C, D)
    - C is condition attribute set (features)
    - D is decision attribute set (buyer, non-buyer)
  - Evaluation Metrics
    - Precision
    - Reach

$$= \frac{\text{\# of people who searched “laptop” before purchasing and bought computers}}{\text{\# of people who searched “laptop”}}$$

$$= \text{\# of people who searched “laptop”}$$

Experimental Data

- Input Data (December 2005 with 83,635 users)

  T = (C, D), where C is the condition attribute sets (feature sets), D is the decision attribute sets (buyer, non-buyer)

Panel ID	**Feature 1 (Whether searched "laptop" related keywords on Google before purchase)**	**Feature 2 (Whether visited online manufacturer store before purchase)**	**Feature 3 (Whether made a purchase last month)**	**… …**	**Feature n (whether visited a review website before purchase)**	**Decision Attribute (Whether is a buyer)**
1	Yes	No	Yes	…	Yes	Buyer
2	No	No	No	…	No	Non-buyer
3	Yes	No	No	..	Yes	Non-buyer
…	…	…	…	…	…	…
83,635	No	No	Yes	…	No	Non-buyer

Experimental Design

- The goal is to predict whether a user is a potential online buyer or non-buyer for a given product category (computer)
- Experiments
  - Classification algorithms (C4.5, Naïve Bayes, Logistic Regression)
  - 10-fold cross-validation
  - Evaluation Metrics

	actual buyer	actual non-buyer
Predicted buyer	TP	FP
Predicted non-buyer	FN	TN

Precision = TP/(TP+FP)
(Conversion Rate)

Recall = TP/(TP+FN)

Classification results comparisons

User-Centric Classifier	Precision	Recall
C4.5 decision tree	29.47%	8.37%
Naïve Bayes	3.52%	23.2%
Logistic Regression (cutoff rate is 0.5)	18.52%	2.23%

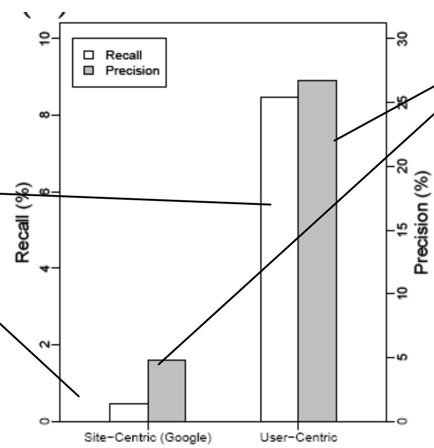
- Classifiers can be created based on user-centric features to predict potential buyers.
- C4.5 obtains the highest prediction precision.
- The branching nodes in the tree splitting a potential buyer and non-buyer can be detected and used for suggesting personalized product content.

User-Centric vs. Site-Centric Classifier

- We compare decision tree classifier against the best site-centric feature as a single classifier from a major search engine.
- "users who searched for laptop keywords on Google before purchasing and searched more than one session"

Precision 29.47% vs. 4.76%

Recall 8.37% vs. 0.45%



Classifier Comparisons

This demonstrates the rich value contained in user-centric data for widely applicable prediction problems.

Refer to the following paper for more details. "Learning User Purchase Intent From User-Centric Data". The 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Osaka, Japan. May 20-23, 2008.

Challenges We Experienced

- Terabytes of Data
- Feature Construction
  - Rich data allows infinite number of features
  - Requires a mix of domain knowledge and data miner's expertise
  - Features specific to the site design, product, …
- Search Term extraction
  - Different website uses different indications
    - Bestbuy.com "query=", Buy.com "qu=", staples.com "keyword=", …

References

- **Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. August 2000. 550 pages. ISBN 1-55860-489-8**
- Data mining and KDD (SIGKDD member CDROM):
  - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery
- Database field (SIGMOD member CD ROM):
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.

This slide is taken from **Data Mining: Concepts and Techniques** course slides for Chapter 1. Introduction (http://www.cs.sfu.ca/~han/dmbook)