Web Usage Mining

- Recommender Systems
  - Introduction
  - Memory-Based Recommender Systems
  - Model-Based Recommender Systems
- Web Log Mining

Recommender Systems

- Scenario:
  - Users have a potential interest in certain items
- Goal:
  - Provide recommendations for individual users
- Examples:
  - recommendations to customers in an on-line store
  - movie recommendations

	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Customer A	X			X		
Customer B		X	X		X	
Customer C		X	X			
Customer D		X				X
Customer E	X				X	

Recommender Systems

- User provide recommendations
  - implicit
    (buying decisions, click streams, reading time of articles,...)
  - explicit
    (feedback forms, texts, mining public sources, ...)
- The recommender system
  - computes recommendations
  - can direct them to the right users
    - filter out items with negative recommendations
    - sort items
    - present evaluations
    - place ads tailored to the user's interests

Example: amazon.com

- "If I have 2 million customers on the Web, I should have 2 million stores on the Web" (Jeff Bezos, CEO)

- Types of recommendations:
  - display of customer comments
  - personalized recommendations based on buying decisions
  - customers who bought also bought.... (books/authors/artists)
  - email notifications of new items matching pre-specified criteria
  - explicit feedback (rate this item) to get recommendations
  - customers provide recommendation lists for topics

amazon.de

WUNSCHZETTEL | MEIN KONTO | HILFE | IMPRESSUM

NEU! DVD-VERLEIH Mehr dazu

HOME | JOHANNES' SHOP | BÜCHER | ENGLISH BOOKS | ELEKTRONIK & FOTO | MUSIK | LEIHEN DVD | VHS | SOFTWARE | PC- & VIDEO-SPIELE | KÜCHE, HAUS & GARTEN | SPIELWAREN & KINDERWELT

INTERNATIONAL | TOPSELLER | PREIS-HITS | GUTSCHEINE | JETZT VERKAUFEN | FILM-HIGHLIGHTS

SCHNELLSUCHE Alle Produkte LOS

Kostenlose Lieferung ab 20 EUR. Bücher versandkostenfrei! Mehr dazu.

ANGEBOT DER WOCHE

Exklusive CDs

Diese CDs gibt's nur bei Amazon.de!

UNSERE SHOPS

Buch, Musik & DVD
- Bücher
- English Books
- Zeitschriften
- Musik
- DVD
- Video

Elektronik & PC
- Elektronik
- Kamera & Foto
- Computer & Zubehör
- PC- & Videospiele
- Software

Haus & Garten
- Küche & Haushalt
- Garten & Freizeit
- Heimwerken
- Körperpflege & Bad

Spielwaren & Kinderwelt
- Spielwaren
- Kinderbücher
- Kinder-DVDs

Geschenke & E-Cards
- Geschenke
- E-Cards

Günstige Angebote
- Gebraucht-Shop
- Auctions
- zShops

Hallo, Johannes Fürnkranz! Hier sind Ihre persönlichen Empfehlungen.
(Wenn Sie nicht Johannes Fürnkranz sind, klicken Sie bitte hier.)

Neue und künftige Veröffentlichungen

Python von Marc Balmer
(Warum wurde mir das empfohlen?)

City of God (2 DVDs) ~Alexandre Rodrigues (Darsteller), u. a.
(Warum wurde mir das empfohlen?)

Harry Potter and the Half-Blood Prince (Harry Potter 6) von J.K. Rowling
(Warum wurde mir das empfohlen?)

▶ **Mehr Empfehlungen**

Empfehlungen für Sie

Um die Ecke gedacht von Eckstein
(Warum wurde mir das empfohlen?)

AI Game Engine Programming with CDROM (Charles River Media Game Development (Paperback)) von Brian Schwab
(Warum wurde mir das empfohlen?)

Game Programming Gems 5 (Charles River Media Game Development (Hardcover)) von Kim Pallister (Herausgeber)
(Warum wurde mir das empfohlen?)

▶ **Mehr Empfehlungen**

AI for Game Developers

Intended for C/C++ programmers new to artificial intelligence, this book shows how to give game characters believable intelligence by employing a mix of deterministic and newer AI techniques. Bourg (New Orleans School of Marine Engineering) and Seemann (Crescent Vision Interactive) explain the... Mehr dazu | (Warum wurde mir das empfohlen?)

Top Produkte bis zu **40%** reduziert
Hier klicken!

DVDs reduziert
Entdecken Sie die Vielfalt deutscher Filme: 400 deutsche Filme bis zu 40% reduziert!

NEU FÜR SIE

Johannes, das gibt es heute **Neu für Sie:**

(Sind Sie nicht Johannes Fürnkranz, klicken Sie bitte hier)

Ihr Einkaufswagen

Sie haben 0 Artikel in Ihrem Einkaufswagen.

Ihre Neuerscheinungen

Robin Gibb with the Neue Philharmonie Frankfurt Orchestra - Live

Weitere Kategorien
- Originalfassungen
- Kochen & Lifestyle
- Originalfassungen
- Kinder & Familie

amazon.de

WUNSCHZETTEL | MEIN KONTO | HILFE | IMPRESSUM

NEU! DVD-VERLEIH Mehr dazu

HOME | JOHANNES' SHOP | BÜCHER | ENGLISH BOOKS | ELEKTRONIK & FOTO | MUSIK | LEIHEN DVD | VHS | SOFTWARE | PC- & VIDEO-SPIELE | KÜCHE, HAUS & GARTEN | SPIELWAREN & KINDERWELT

IHRE PERSÖNLICHE SEITE | IHRE PERSÖNLICHEN EMPFEHLUNGEN | IHRE LIEBLINGSSHOPS | NEU FÜR SIE

Schnellsuche: Alle Produkte LOS | Stöbern: Bücher LOS

IHRE EMPFEHLUNGEN

Alle Produkte
Alles gebraucht

Ihre Favoriten Ändern

English Books
Software
DVD
Bücher

Mehr Shops

Zeitschriften
Musik
Klassik
VHS
PC- & Videospiele
Spielwaren & Kinderwelt
Elektronik
Computer & Zubehör
Kamera & Foto
Küche & Haushalt
Heimwerken
Garten
Körperpflege & Bad

Persönliche Empfehlungen

Hallo, Johannes Fürnkranz. Entdecken Sie die heute vorgestellten Empfehlungen. (Wenn Sie nicht Johannes Fürnkranz sind, klicken Sie hier.)

Software Empfehlungen

Lernspaß - 1. Klasse

Aus der Amazon.de-Redaktion
Verheißungsvoll klingt der Titel, bei dem sich wohl alle Eltern erträumen, es möge den eigenen Kindern zeitlebens so ergehen: *Lernen macht Spaß*. Diese Software unterstützt Erstklässler in den Fächern Mathematik und Deutsch, steigert ihr Konzentrationsvermögen... Mehr dazu

Mehr gibt es in Kinder & Familie, Schule & Studium, und anderen Software Empfehlungen

DVD-Empfehlungen

The King And I [UK IMPORT]

Aus der Amazon.de-Redaktion
Der König und ich ist der dritte Broadway-Hit des berühmten Komponistenduos Rogers & Hammerstein. Der Film zeigt eine schauspielerische Leistung Yul Brynners, die seiner Karriere einen Schwung nach oben verlieh. Brynner wiederholte seinen Bühnenerfolg in der Hauptrolle und bewies den... Mehr dazu

Mehr gibt es in Originalfassungen, und anderen DVD-Empfehlungen

Buch-Empfehlungen

Guck mal, was hier passiert!

Kurzbeschreibung
Ein Wimmelbilderbuch zum Schauen, Entdecken, Wiedererkennen und natürlich zum Geschichtenerfinden und -erzählen. (Ab 2 Jahren.)

Mehr gibt es in Kochen & Lifestyle, und anderen Buch-Empfehlungen

Verbessern Sie Ihre Empfehlungen

Haben wir mit den empfohlenen Artikeln Ihren Geschmack noch nicht ganz getroffen? Lassen Sie uns genauer wissen, was Sie interessiert:

Ändern Sie Ihre bisherigen Angaben

Wählen Sie Ihre bevorzugten Interessensgebiete

Bewerten Sie Artikel, die Sie schon haben

Empfohlene Autoren, Künstler & Regisseure

Eckstein

- Samba kurz & gut
- Um die Ecke gedacht
- Agile Softwareentwicklung im Großen

Mehr Autoren

Uriah Heep

- Anthology [DOPPEL-CD]

PRODUKTINFO

Mehr zu diesem Buch

- Überblick
- Inhaltsverzeichnis
- Amazon.de-Redaktion

Mehr von ...

- Soumen Chakrabarti

Kunden kauften auch

- diese Produkte

Was meinen Sie?

- Ihre Meinung zu diesem Buch
- Empfehlen Sie das Buch per E-Mail weiter

mining the web

von Soumen Chakrabarti

US-Preisempfehlung*: $57.95
Amazon-Preis: EUR 53,90 **Kostenlose Lieferung.** Siehe Details.

Versandfertig bei Amazon in 10 bis 12 Tagen.

Noch schneller geht's mit DHL Expressversand.

Alle Angebote ab EUR 40,70

Größeres Bild

*suggested retail price

Kategorie(n): Computers & Internet

BEI AMAZON.DE KAUFEN

In den Einkaufswagen

ODER

Hier anmelden, um 1-Click® zu aktivieren.

ALLE ANGEBOTE

26 Neu ab EUR 40,70
3 gebraucht ab EUR 40,72

Möchten Sie verkaufen?
Diesen Artikel verkaufen

Auf meinen Wunschzettel

Meinen Wunschzettel ansehen

Sprache: Englisch
Gebundene Ausgabe - 344 Seiten - Morgan Kaufmann Publishers
Erscheinungsdatum: 1. Oktober 2002
ISBN: 1558607544
Amazon.de-Verkaufsrang 226.869
Erhöhen Sie Ihre Verkäufe!

Schreiben Sie die erste Online-Rezension zu diesem Produkt, und gewinnen Sie mit etwas Glück einen Amazon.de Einkaufsgutschein über 50 EUR.

Kunden, die dieses Buch gekauft haben, haben auch diese Bücher gekauft:

- *Modern Information Retrieval* von Ricardo Baeza-Yates, Berthier Ribeiro-Neto
- *Einführung in die Kryptographie* von Johannes Buchmann

▸ **Entdecken Sie verwandte Produkte**

Unser Vorschlag

Kaufen Sie jetzt diesen Artikel zusammen mit Wenn der Partner geht. Wege zur Bewältigung vo

+

Amazon-Preis: EUR 411,80

Beide jetzt kaufen

Kunden, die diesen Artikel gekauft haben, kauften auch:

Wenn der Partner geht. Wege zur Bewältigung von Trennung und Scheidung von Doris Wolf

★★★★★ (48) EUR 12,80

Quelle: http://fun.sdinet.de/pics/german/waschmaschine.jpg, gefunden von Erik Tews

Recommendation Techniques

- non-personalized recommendations
  - most frequently bought items (Harry Potter)
- attribute-based recommendations
  - books of the same authors
  - books with similar titles
  - books in same category

- item-to-item correlations
  - users who bought this book, also bought...
  - items are similar if they are bought by the same users
- user-to-user correlations
  - people like you also bought...
  - users are similar if they buy the same items

Attribute-Based Recommendations

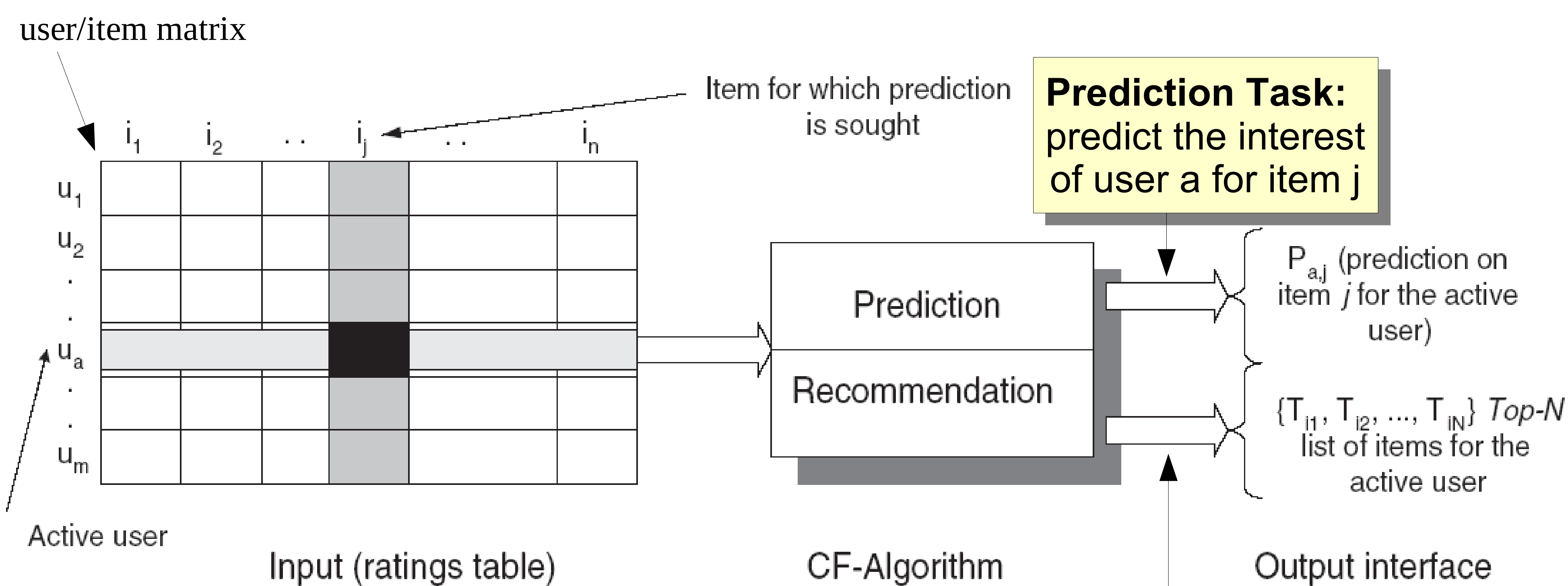
- Recommendations depend on properties of the items
- Each item is described by a set of attributes
  - Movies: e.g director, genre, year, actors
  - Documents: bag-of-word
- Similarity metric defines relationship between items
  - e.g. cosine similarity

Collaborative Filtering

- Recommends products to a target customer based on opinions of other customers

- Representation:
  - user/item matrix (customer/product matrix)
  - similar to document/term matrix
- Neighborhood formation:
  - identify similar customers based on similar buying decisions / recommendations (e.g., cosine similarity), may be optional (i.e., all users are neighborhood)
- Recommendation System:
  - derive a recommendation based on the information obtained from similar customers (e.g., most frequent items in neighborhood, weighted sum,...)

Collaborative Filtering (CF)



Prediction Task:
predict the interest of user a for item j

Recommendation Task:
predict a list of items that are most interesting for user a

the matrix contains recommendations of each user for each product, e.g.
1 if the user bought the item
0 if the user did not buy the item

Source: Sarwar, Karypis, Konstan, Riedl, WWW-10, 2001

Memory-Based Collaborative Filtering

- Simple approach:
  - The weight that user u_a attributes to an item i is the sum of the votes that the item receives from other users
  - weighted by the similarity of the user to the other users

$$v_p(u_a, i) = \kappa \sum_{u \in U} w(u_a, u) \cdot v(u, i)$$

$v(u,i)$ ...... vote of user u for item i

v_p .... predicted vote

$w(u_1,u_2)$ ... weight between user u_1 and user u_2

u_a .... active user

κ............. normalization factor for weights in the sum $\kappa = \frac{1}{\sum_{u \in U} w(u_a, u)}$

Memory-Based Collaborative Filtering

- Problem with the simple approach:
  - different users may have different scales
  - a recommendation of 6 out of 10 may be pretty good for critical users, or quite bad for others
- Solution:
  - Only consider deviations from the mean
    - normalize each vote with the average vote $m(u)$ of that user so that a vote of 0 is an average vote
    - add the predicted average deviation to the average vote of the active user

$$v_p(u_a, i) = m(u_a) + \kappa \sum_{u \in U} w(u_a, u)(v(u, i) - m(u))$$

$m(u)$ ... expected value (mean) over all votes of user

$$\kappa = \frac{1}{\sum_{u \in U} w(u_a, u)}$$

Memory-Based Collaborative Filtering

- The weight matrix $w(u_1,u_2)$ user-to-user correlations
- can be measured in different ways, e.g.:
  - cosine similarity:

$$w(u_1,u_2)=\frac{\sum_{i\in I} v(u_1,i)\cdot v(u_2,i)}{\sqrt{\sum_{i\in I_{u_1}} v(u_1,i)^2\cdot \sum_{i\in I_{u_2}} v(u_2,i)^2}}$$

  - correlation:

$$w(u_1,u_2)=\frac{\sum_{i\in I_{u_1}\cap I_{u_2}} v_m(u_1,i)\cdot v_m(u_2,i)}{\sqrt{\sum_{i\in I_{u_1}\cap I_{u_2}} v_m(u_1,i)^2\cdot \sum_{i\in I_{u_1}\cap I_{u_2}} v_m(u_2,i)^2}}$$

= cosine similarity of adjusted votes $v_m(u,i)=v(u,i)-m(u)$
restricted to all items where both users vote

Extensions

- Default Voting
  - default votes for items without explicit votes
  - allows to compute correlation from union instead of intersection (more items → more reliable)
- Inverse user frequency
  - reduce weights for objects popular with many users
  - *assumption:* universally liked items are less useful
    - cf. IDF
- Combine collaborative filtering with content-based similarities
  - user similarities:
    based on user profiles
  - item similarities:
    e.g., product categories, textual similarities, etc.

Extensions (Ctd.)

- Addition of pseudo users
  - use background knowledge (e.g., musical genres)
  - generate pseudo users that comment positively on all items of the genre
  - might be extracted automatically by wrappers (Cohen & Fan 2000)

Item Correlations

- Past purchases are transformed into relationships of common purchases

	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Customer A	X			X		
Customer B		X	X		X	
Customer C		X	X			
Customer D		X				X
Customer E	X				X	
Customer F			X		X	

Customers who bought…	Also bought…					
	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Book 1				1	1	
Book 2			2		1	1
Book 3		2			2	
Book 4	1					
Book 5	1	1	2			
Book 6		1				

Item Correlations

- Such correlation tables can then used to made recommendations
- If a visitor has some interest in Book 5, he will be recommended to buy Book 3 as well

		Also bought…					
		Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Customers who bought…	Book 1				1	1	
	Book 2			2		1	1
	Book 3		2			2	
	Book 4	1					
	Book 5	1	1	2			
	Book 6		1				

Original Slide by Arnaud De Bruyn, Penn State University

Problems with Memory-Based Collaborative Filtering

- Cold Start:
  - There needs to be enough other users already in the system to find a match.
- Sparsity:
  - If the user/ratings matrix is sparse, it is hard to find users that have rated the same items (likely to happen with many items)
- First Rater:
  - Cannot recommend an item that has not been previously rated (e.g., New items, Esoteric items, ...)
- Popularity Bias:
  - Cannot recommend items to someone with unique tastes.
  - Tends to recommend popular items.

Model-Based Collaborative Filtering

- learn an explicit model that predicts ratings and/or items
- examples
  - clustering of users
    - each user is characterized by her recommendations
    - apply any clustering algorithm that works for clustering documents
  - clustering of items
    - each item is characterized by the users that recommend it
    - apply any clustering algorithm that works for clustering documents
  - clustering of both users and items (*co-clustering*)
    - advantage: items and users are mutually dependent, a good clustering needs to consider both dimensions.
  - association rules
    - model associations between items
    - advantage: explicit, understandable representation

Clustering

	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Customer A	X			X		
Customer B		X	X		X	
Customer C		X	X			
Customer D		X				X
Customer E	X				X	

- Two Clusters based on similarity on bought items
  - Customers B, C and D are clustered together
  - Customers A and E are clustered into another group
- « Typical » preferences for **CLUSTER BCD** are:
  - Book 2, very high
  - Book 3, high
  - Books 5 and 6, may be recommended
  - Books 1 and 4, not recommended at all

Clustering

	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Customer A	X			X		
Customer B		X	X		X	
Customer C		X	X			
Customer D		X				X
Customer E	X				X	
Customer F			X		X	

- How do we recommend within a cluster?
- Any customer that will be classified as a member of **CLUSTER BCD** will receive recommendations based on preferences of the group:
  - Book 2 will be highly recommended to Customer F
  - Book 6 will also be recommended to some extent

Problems

- Customers may belong to more than one cluster
  - in our example: Customer F could fit to both clusters
- there may be overlap in items between clusters
  - clusters may be overlapping (one example may belong to different clusters)
- Possible solution:
  - average predictions of all fitting clusters
  - weighted by their importance

Co-Clustering

- Cluster users and items simultaneously
  - Mutual reinforcement of similarity
  - separate clusterings might be suboptimal
- Need advanced clustering techniques
  - e.g., (Ungar & Foster, 1998)

	Batman	Rambo	Andre	Hiver	Whispers	StarWars
Lyle			1			1
Ellen			1	1		1
Jason				1	1	
Fred	1					1
Dean	1	1				1
Karen	?	?	1	?	?	?

From *Clustering methods in collaborative filtering,* by Ungar and Foster

Association Rule Discovery

- Association Rules describe frequent co-occurences in sets
  - generalize correlation tables to correlations between more than two values
- Example Problems:
  - Which products are frequently bought together by customers? (*Basket Analysis*)
    - DataTable = Receipts x Products
    - Results could be used to change the placements of products in the market
  - Which courses tend to be attended together?
    - DataTable = Students x Courses
    - Results could be used to avoid scheduling conflicts....
  - Which words co-occur in a text?
    - cf. efficient generation of n-grams

Association Rules

- General Form:

  $$A_1, A_2, ..., A_n \Rightarrow B_1, B_2, ..., B_m$$

- Interpretation:
  - When items A_i appear, items B_i also appear with a certain probability
- Examples:
  - `Bread, Cheese => RedWine.`
    Customers that buy bread and cheese, also tend to buy red wine.
  - `MachineLearning => WebMining, MLPraktikum.`
    Students that take 'Machine Learning' also take 'Web Mining' and the 'Machine Learning Praktikum'

Basic Quality Measures

- **Support** $s(A \rightarrow B) = \frac{n(A \cup B)}{n}$
  - relative frequency of examples for which both the head and the body of the rule are true
- **Confidence** $c(A \rightarrow B) = \frac{n(A \cup B)}{n(A)}$
  - relative frequency of examples for which the head is true among those for which the body is true
- Example:
  - `Bread, Cheese => RedWine` (S = 0.01, C = 0.8)

    80% of all customers that bought bread and cheese also bought red wine.
    1% of all customers bought all three items.

Using Association Rules for Recommendations

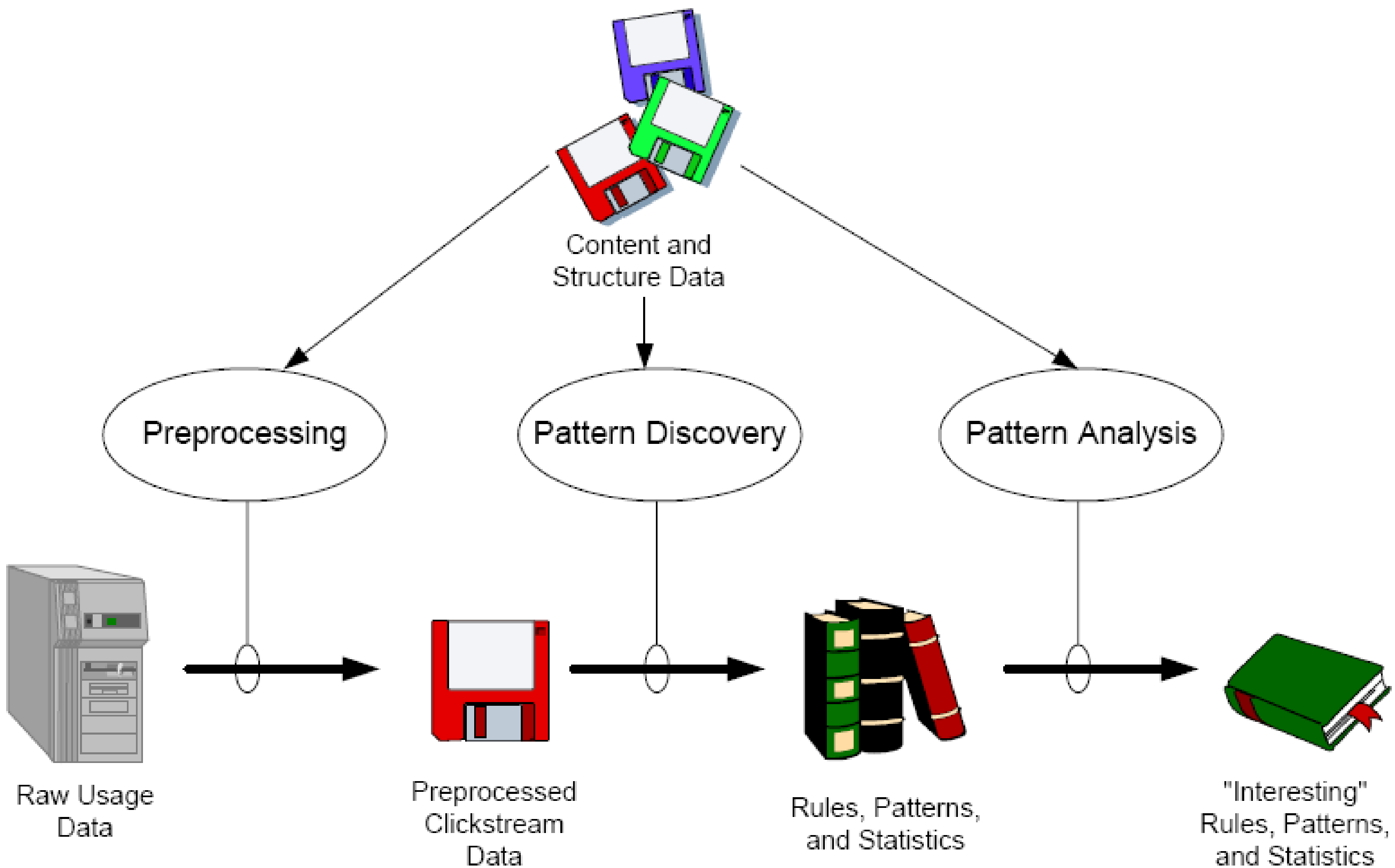
- APRIORI:
  - efficient algorithm for finding all rules that have a given *mimimum support* and a given *minimum confidence*
  - phase 1: find frequent item sets (→ n-grams)
  - phase 2: construct all rules with min confidence from item set

- Simple Use of APRIORI for recommendations:
  1. Input: database of all customers x all items they have bought
  2. Find association rules
  3. Find all rules whose conditions match the items previously bought by the active user
  4. Sort these rules by their confidence
  5. Predict the first N items on the top of the list

Web Log Mining

- Applying Data Mining techniques to the discovery of usage patterns in Web sites
  - e.g.: Find association rules that capture which pages are frequently visited in succession to each other
- Goals
  - improvement of site design and site structure
  - generation of dynamic recommendations
  - improving marketing
- Phases
  - data collection
  - pre-processing
  - pattern discovery
  - pattern analysis

Web Log Mining Process



Raw Data: Web Logs

#	IP	Id	Acces	Time	Method/URL/Protocol	Status	Bytes	Referer	Agent
1	165.182.168.101	-	-	16/06/2002:16:24:06	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
2	165.182.168.101	-	-	16/06/2002:16:24:10	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
3	165.182.168.101	-	-	16/06/2002:16:24:57	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
4	204.231.180.195	-	-	16/06/2002:16:32:06	GET p3.htm HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
5	204.231.180.195	-	-	16/06/2002:16:32:20	GET C.gif HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
6	204.231.180.195	-	-	16/06/2002:16:34:10	GET p1.htm HTTP/1.1	200	3821	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
7	204.231.180.195	-	-	16/06/2002:16:34:31	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
8	204.231.180.195	-	-	16/06/2002:16:34:53	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
9	204.231.180.195	-	-	16/06/2002:16:38:40	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
10	165.182.168.101	-	-	16/06/2002:16:39:02	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
11	165.182.168.101	-	-	16/06/2002:16:39:15	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
12	165.182.168.101	-	-	16/06/2002:16:39:45	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
13	165.182.168.101	-	-	16/06/2002:16:39:58	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
14	165.182.168.101	-	-	16/06/2002:16:42:03	GET p3.htm HTTP/1.1	200	4036	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
15	165.182.168.101	-	-	16/06/2002:16:42:07	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
16	165.182.168.101	-	-	16/06/2002:16:42:08	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
17	204.231.180.195	-	-	16/06/2002:17:34:20	GET p3.htm HTTP/1.1	200	2342	out.htm	Mozilla/4.0 (MSIE 6.0; Win98)
18	204.231.180.195	-	-	16/06/2002:17:34:48	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 6.0; Win98)
19	204.231.180.195	-	-	16/06/2002:17:35:45	GET p4.htm HTTP/1.1	200	3523	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
20	204.231.180.195	-	-	16/06/2002:17:35:56	GET D.gif HTTP/1.1	200	3231	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)
21	204.231.180.195	-	-	16/06/2002:17:36:06	GET E.gif HTTP/1.1	404	0	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)

Slide by Juan D. Velasquez

Preprocessing

- Identify user sessions in the log
  - so that we can see what individual users are doing
- Problems:
  - User Identification
    - Same IP does not need to be the same user
  - Session Time
    - Does a long break mean the user's session has ended?
  - Missing pages
    - not all retrieved pages appear in user log (e.g., might have been retrieved from user cache)

Some Heuristics for Session Identification

- Timeout:
  - if the time between pages requests exceeds a certain limit, it is assumed that the user is starting a new session
- IP/Agent
  - Different agent types for an IP address represent different sessions
- Referring page:
  - If the referring page for a request is not part of an open session, it is assumed that the request is coming from a different session.
- Same IP-Agent/different sessions (Closest):
  - Assigns the request to the session that is closest to the referring page at the time of the request.
- Same IP-Agent/different sessions (Recent):
  - In case of a tie, assign the request to the session with the most recent referrer access in terms of time

Data Analysis

Session traces can be mined for various useful patterns

- Basic statistics
  - Which pages are most frequently accessed?
  - Feedback about interestingness of content/products on these pages
- Association Rules
  - Which pages are accessed together?
    - products/contents of related interest
  - Which paths are frequently taken?
    - maybe provide a shortcut link to improve user satisfaction
- Clustering
  - find clusters of similar pages or clusters of similar users