Statistics 215a D. R. Brillinger 11/29/04

Association rules / market basket analysis

"... among data mining's bissest successes."

Data base of individuals' purchases

"In 80% of the cases when people buy bread, they also buy milk."

"70% of the people who buy spaghetti, wine, and sauce also buy garlic bread."

Concerned with co-occurrence of items in large volumes of transactions.

Looking for interesting patterns and trends

Web access: 80% of times user visited pages A and B also visited C

Set up a dynamic link

Market-basket data

$n \times p$ matrix of 0's and 1's

rows: baskets

columns: items

$n \sim$ millions

$p \sim$ tens of thousands

generally sparse matrix

Hastie, Tibshirani, Friedman example

p. 444 $n = 6876$, $p = 50$

Mall customers in Bay Area

Table 14.1

Demographic	#values	type
Sex	2	categorical
marital status	5	categorical
age	7	ordinal
education	6	ordinal
⋮		

Categoricals. Created k dummy variables

$Z_{21} = I(\text{income} < \$40k)$, $Z_{22} = I(\text{income} > \$40k)$

Ordinals. Cut at median

Jargon

itemset. set of items

Eg. A: $z_1 = 1$ male

B: $z_3 = 1$ unmarried

C: $z_1 = 1$ and $z_3 = 1$

support $f_r(C)$ fraction

Association rule. $A \Rightarrow B$

$Z_1 = 1$ implies $Z_3 = 1$

Support of $A \Rightarrow B$ is $fr(A \cap B)$

Confidence of rule $A \Rightarrow B$ is $\frac{fr(A \cap B)}{fr(A)} = \frac{n}{n_2}$

Lift of $A \Rightarrow B$ is $\frac{fr(A \cap B)}{fr(A)\, fr(B)}$

basket	A	B
1	1	0
2	1	1
3	1	0
4	0	0
5	0	1
6	1	1
7	1	0
8	0	1
9	1	0
10	0	1

Support. $fr(A \cap B) = \frac{2}{10} \equiv 20\%$

Confidence. $\frac{fr(A \cap B)}{fr(A)} = \frac{2}{6} \equiv 33.3\%$

Lift. $\frac{fr(A \cap B)}{fr(A)\, fr(B)} = \frac{20}{30} = .67$

$r = -.41$

6+

Example 1

A: number in household = 1
and
number of children = 0

B: language in home ≠ English

support 25%

confidence 99.7%

lift 1.03

single + no children means anglophone 99.7%
of the time (high association!)

{solo household, no children, anglophone}

Rule: solo and no children ⇒ anglophone

Support. For 25% of the cases solo, no children, anglophone together

Confidence. When solo and no children, 99.7% anglophone

Lift. If anglophone 96.8% of time
lift = 1.03

Might ask for all cases where:

anglophone, confidence > 80%,
support > 25 %

(solo and no children would be discovered)

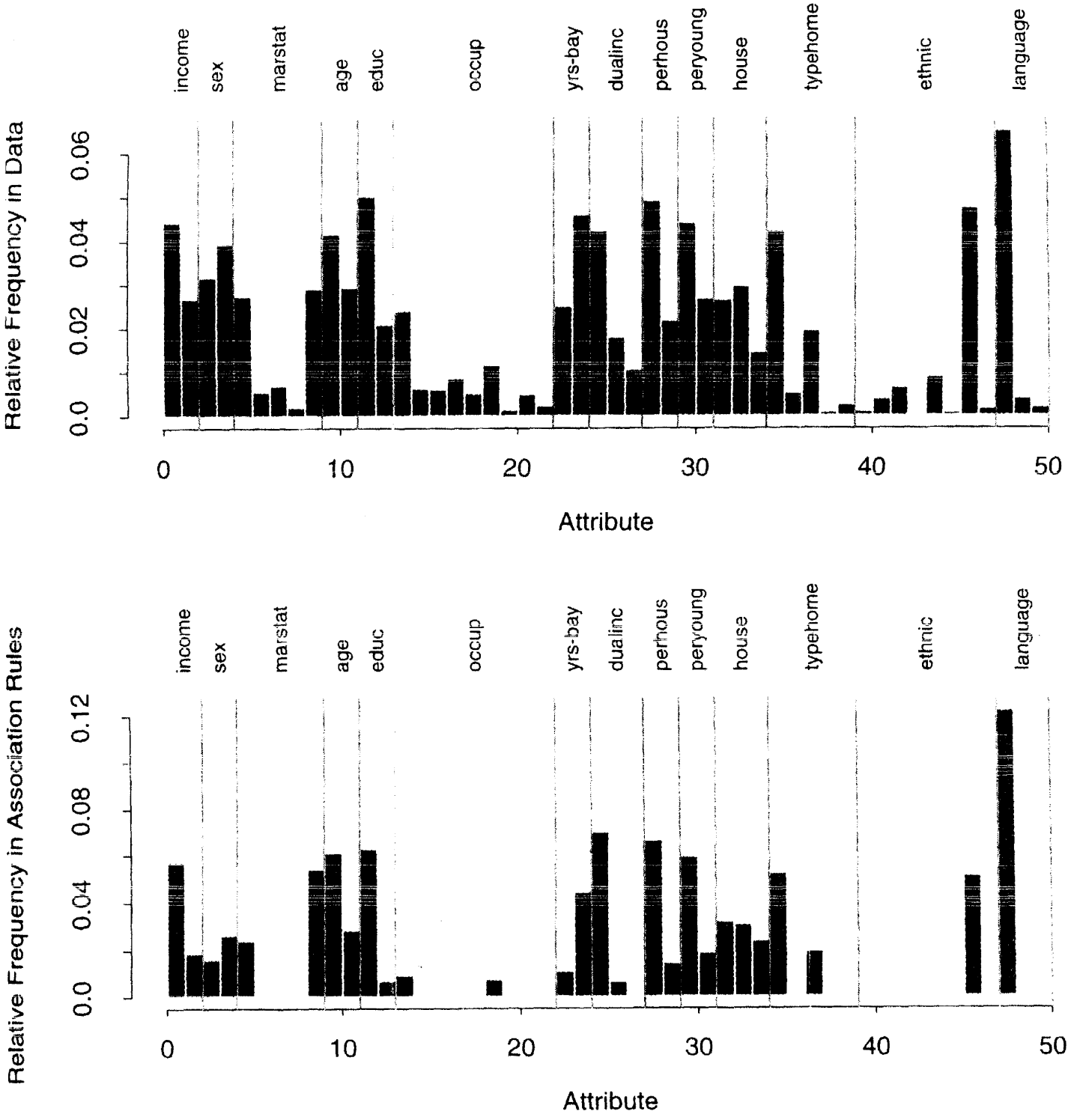
Not worth considering cases with small support.

Might ask for all cases where confidence $> p_c$ and support $> p_s$

How to find them?

Apriori algorithm

Agrawal & Srikani (1994)



GURE 14.2. *Market basket analysis: relative frequency of each dummy va e (coding an input category) in the data (top), and the association rules for the Apriori algorithm (bottom).*

"Personalization of supermarket product recommendations" Lawrence et al. (2001)

Data Mining and Knowledge Discovery 5, 11-32

IBM, Safeway (UK)

SmartPad remote shopping system

Personal digital assistant (PDA)

e.g. Palm Pilot

To build shopping list

Customers do not walk the aisles

Question.

Substitute for "spontaneous purchase"

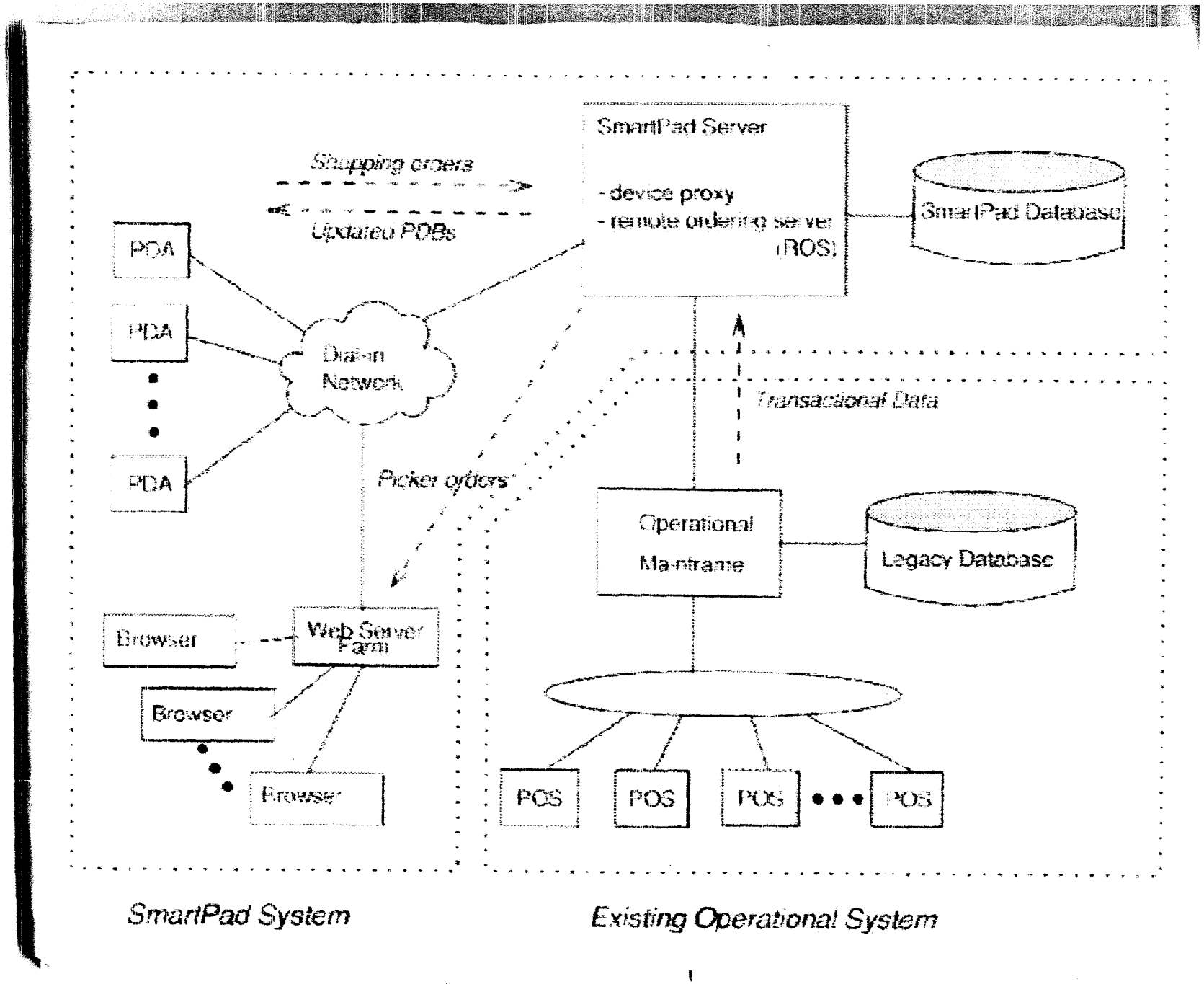
Solution

Compute recommendations

Deliver to individual's PDA

Associations mining + clustering

II

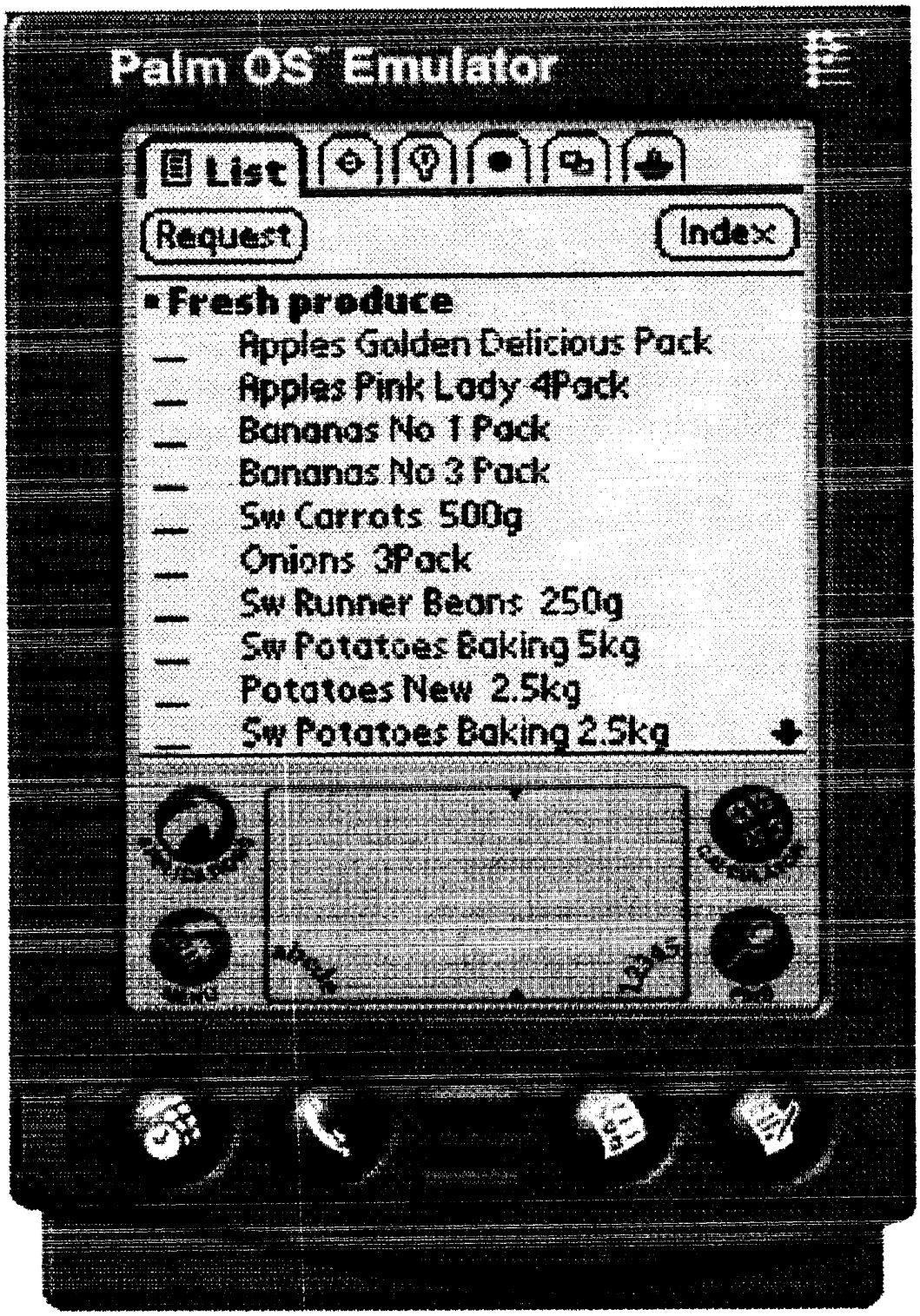

SmartPad Server
- device proxy
- remote ordering server
(ROS)
SmartPad Database
Shopping orders
Updated PDBs
PDA
PDA
PDA
Dial-in
Network
Picker orders
Transactional Data
Operational
Mainframe
Legacy Database
Browser
Web Server
Farm
Browser
Browser
POS
POS
POS
POS
SmartPad System
Existing Operational System

Personal databases for product choice

1. personal catalog
* 2. recommendations
3. promotions

Downloaded at customer call

III



nartPad main screen.

The recommender system

30,000 products

20,000 customers

8 weeks data for 8000 customers

10-20 products with highest scores

IV

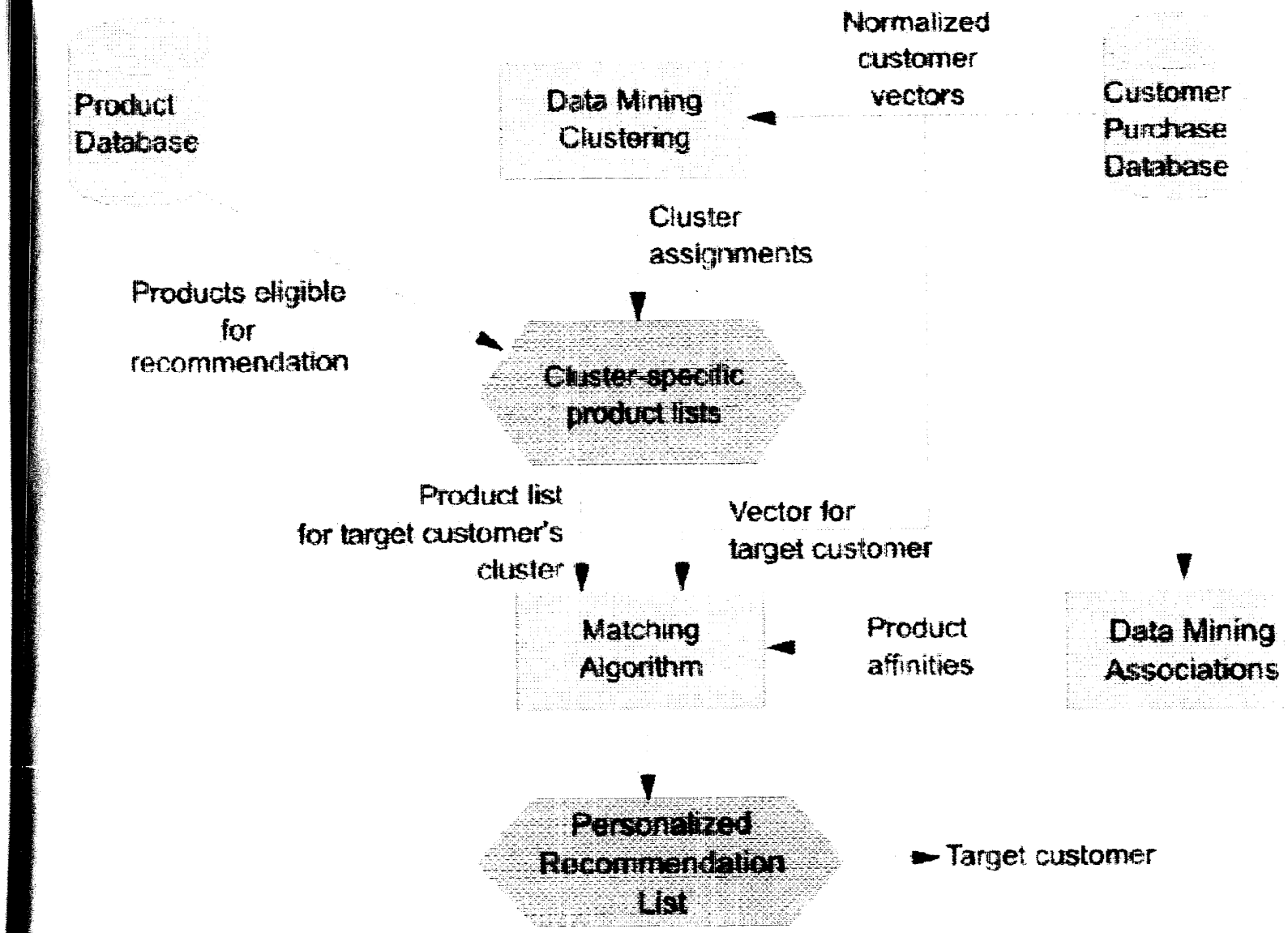


Figure 3. Overview of the personalized recommender system.

Product taxonomy

99 classes

2302 subclasses

Determine products "best" matched to customer's spending profile

Construct customer and product vectors

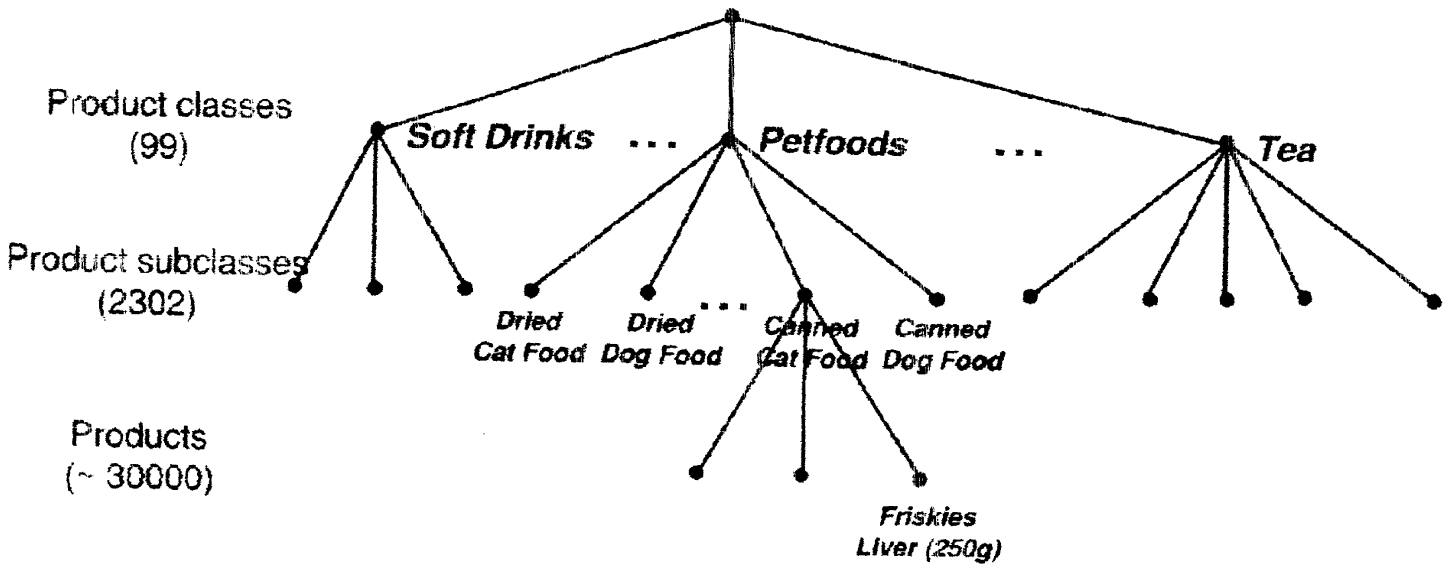


Figure 4. Safeway product taxonomy.

Product model

Idea: purchase of product in subclass implies interest in other products in same subclass

VI

Associations mining.

minimum support 1-4%

minimum confidence 30-40%

minimum lift 2-3

Table 1. Sample associations computed at the product-class and product-subclass levels.

Sup	Conf	Lift	Class or subclass	Relevant affinities
0.059	0.41	2.4	20(Baby products)	⇒ 41(Canned pasta)
0.082	0.47	2.2	66(Table wines)	⇒ 68(Beer/Lager/Spirits)
0.125	0.50	2.0	90(Fresh beef)	⇒ 91(Pork/Lamb)
0.025	0.38	9.0	2010(Baby:Disposable nappies)	⇒ 2007(Baby:Wipes)
0.016	0.33	4.9	2010(Baby:Disposable nappies)	⇒ 1012(Dairy:Childrens' yogurt)
0.01	0.33	4.9	2010(Baby:Disposable nappies)	⇒ 3115(Instore:Babysitting center)
0.012	0.37	3.4	1020(Dairy:Childrens' fromage)	⇒ 3115(Instore:Babysitting center)
0.016	0.52	5.2	2306(Biscuits:Kids biscuits)	⇒ 3115(Instore:Babysitting center)
0.022	0.30	4.9	9015(Fresh beef:Beef joints)	⇒ 9120(Pork/Lamb:Pork joints)

Summary.

1.8% boost in revenue

Substantial fraction accepted recommendations

"statisticians need to focus on interpretation of the rules that the algorithms produce, not the computational efficiency"

Hand, Mannila & Smyth (2001)