

# Challenging the KDD Cup by DSS datamining technology

---

Masao Okada,  
Database Communications, Inc.



<mailto:okada@database.co.jp>

# Profile of DCI

- DCI was established in 1987.
- Distributor of CCA and Sirius Software products
- Developing DSS (Decision Support System) datamining application
- Okada got the US patent, PN-6907415, a method for finding rules and exceptions from database.





# What is the KDD Cup ?

- KDD Cup is the annual knowledge discovery and data mining (KDDM) tools competition organized by ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining)
- Primary focus of the SIGKDD is to provide the premier forum for advancement and adoption of the "science" of KDDM.
- When the data for the task is provided, contestants can use any method to solve it.
- The task is very tough to solve.

# What is the KDD Cup ? (continued)

---

- The tasks from various areas are given:
  - KDD-Cup 1997, Direct marketing for lift curve optimization
  - KDD-Cup 1998, Direct marketing for profit optimization
  - KDD-Cup 1999, Computer network intrusion detection
  - KDD-Cup 2000, Online retailer website clickstream analysis
  - KDD-Cup 2001, Molecular bioactivity; plus Protein locale prediction
  - KDD-Cup 2002, BioMed document; plus Gene role classification
  - KDD-Cup 2003, Network mining and usage log analysis
  - KDD-Cup 2004, Particle physics; plus Protein homology prediction
  - KDD-Cup 2005, Internet user search query categorization
  - KDD-Cup 2006, Pulmonary embolisms detection from image data
- For more information, please visit
  - <http://www.acm.org/sigs/sigkdd/kddcup/index.php>

# 1998 KDD Cup task

---

- Profit maximization of Donation:  
To whom the promotion should be done to maximize the profit ?
  - Population: 96,367 donors
  - Donation history data for recent 2 years is given (Includes about 400 fields).
  - Package and mail cost is \$0.68 for each.
  - If the promotion is done to all donors, the profit will be \$10,777. But, if unlikely donors can be avoided, the profit will be increased.

# Results

---

- DCI's solution got the profit of \$15,013, the number of the target donors was 51,666 => **This is the world record (unofficial), higher than the GOLD award by \$301.**
- The official results of the KDD Cup 1998
  - GOLD award: GainSmarts of Urban Science Applications, Inc., the profit was \$14,712.
  - SILVER award: Enterprise Miner of SAS Institute, Inc.
  - BRONZE award: Decisionhouse of Quadstone Limited, the profit was \$13,954.

# What is Datamining ?

- Discovering hidden useful patterns or relations from large-scale database automatically.
- We call such patterns or relations as Database Rules.
  - Example: Relation between diapers & beers  
On Thursday nights, males who buy diapers buy also beers. (Walmart)



# How to represent Database Rules

- We represent database rules in IF-THEN format.
  - Easy to understand for human, and  
Easy to be handled by computer
  - Example: **IF** Form = 'BOX' AND Color = 'RED'  
**THEN** Weight IS  $\geq 200$   
(Support = 4, Confidence =  $4/5 = 80\%$ )

Row	Form	Color	Weight
1	BOX	RED	250
2	BOX	RED	210
3	BOX	RED	205
4	BOX	RED	225
5	BOX	RED	190
6	BOX	BLUE	195
7	CONE	BLUE	300



# We have to limit the number of fields for datamining

---

- If the number of fields is large, it requires huge amount of computing for datamining, and we can't process all database rules in limited time.  
→ We have to limit the number of fields for datamining.
- This is because ...(see the next page)

# We have to limit the number of fields for datamining (continued)

- For the case of KDD Cup data,
  - The file has 400 fields.
  - Assume each field has 4 segments. (That is, each it has 4 unique values.)
  - The number of FINDs which include 5 fields is
$$400C_1 * 4 + 400C_1 * 399C_1 * 4^2 + 400C_2 * 398C_1 * 4^3 + 400C_3 * 397C_1 * 4^4 + 400C_4 * 396C_1 * 4^5$$
$$= \text{about } 5 * 10^{14}$$
- Even if we can process one database rule per millisecond, it will take about 12 years.

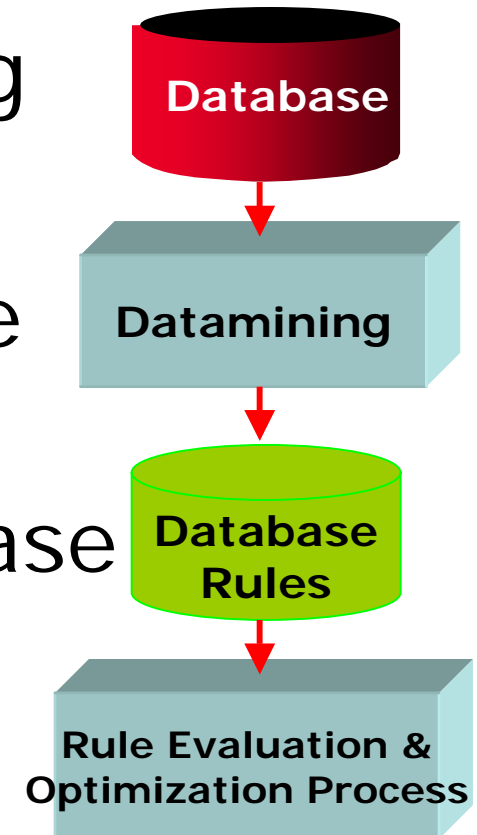
# We have to limit the number of fields for datamining (continued)

---

- From many trials and errors, we concluded the followings things :
  - Number of fields for datamining : 10-20
  - Maximum number of field combination: 5-10
  - Number of segments
    - Numeric field : 2
    - Character field: 3-5
  - At present, we can do the datamining, where the number of fields is 10 and the number of field combination is 5, in one minute.

# DCI's DSS datamining approach

1. Preparation for datamining
2. Execution of datamining
3. Evaluation of the database rules
4. Optimization of the database rules



# Step1: Preparation for datamining

---

- Selection of the key fields
  1. At first, viewed data, and cut off 200 fields which are not related to donation.
  2. Next, evaluated each field on certain criteria, selected the top 10 fields in evaluated value automatically.

# Step1:Preparation for datamining (continued)

## Top 10 fields with evaluated value

Evaluated value is maximum {average donation amount per donor} among segments

No.	FIELDNAME	Evaluated Value	Corresponding Num. Field Value
1	Donor rank	0.412	
2	Socio-Economic status of the neighborhood	0.276	
3	Maximum donation amount	0.263	17
4	Urbanicity level of the neighborhood	0.191	
5	Wealth level	0.181	1
6	Income level	0.180	3
7	Average donation amount	0.172	8
8	Total donation amount	0.160	40
9	Last donation amount	0.145	10
10	Minimum donation amount	0.129	3

# Step 1: Preparation for datamining (continued)

- Segmentation of the key fields
  - For numeric fields
    - Created the following 2 segments for each field:
      - field IS  $< x$
      - field IS  $\geq x$   
, where  $x$  is “Corresponding Field Value”  
described in just before page
  - For character fields
    - For each unique value ( $=x$ ) of each field,  
created one segment such as field =  $x$ .

# Step1: Preparation for datamining (continued)

- Segmentation for numeric fields

FIELDNAME	SEGMENT1	SEGMENT2
Average donation amount	IS < 8	IS >= 8
Income level	IS < 3	IS >= 3
Last donation amount	IS < 10	IS >= 10
Maximum donation amount	IS < 17	IS >= 17
Minimum donation amount	IS < 3	IS >= 3
Total donation amount	IS < 40	IS >= 40
Wealth level	IS < 1	IS >= 1

- Segmentation for character fields

FIELDNAME	SEG1	SEG2	SEG3	SEG4	SEG5
Donor rank	= 'D'	= 'E'	= 'F'	= 'G'	<del></del>
Socio-Economic status of the neighborhood	= '1'	= '2'	= '3'	= '4'	<del></del>
Urbanicity level of the neighborhood	= 'C'	= 'R'	= 'S'	= 'T'	= 'U' <sub>16</sub>

## Step2: Execution of datamining

---

- Generated all combinations of the key fields, where the number of the key fields is up to 5, and retrieved database rules automatically.
- In this execution, the number of the FINDs expected was about 200,000. But the number of the FINDs actually executed was about 9,000.
- Extracted 160 database rules in one minute.

# Step2: Execution of datamining (continued)

## ■ Extracted Rules Example

No.	Conditions	Result	Confidence	Support
1	DOMAIN_SES = '1' LASTGIFT IS >= 10 MAXRAMNT IS >= 17 RAMNTALL IS >= 40	DONATION AMOUNT IS >= 20	3%	431
2	AVGGIFT IS >= 8 DOMAIN_SES = '1' MAXRAMNT IS >= 17 RAMNTALL IS >= 40	DONATION AMOUNT IS >= 20	3%	430
3	DOMAIN_SES = '1' MAXRAMNT IS >= 17 RAMNTALL IS >= 40	DONATION AMOUNT IS >= 20	3%	437
4	DOMAIN_SES = '1' INCOME IS >= 3 LASTGIFT IS >= 10 MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	3%	363
5	DOMAIN_SES = '1' INCOME IS >= 3 MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	3%	367
6	AVGGIFT IS >= 8 DOMAIN_SES = '1' INCOME IS >= 3 MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	3%	361
7	DOMAIN_SES = '1' LASTGIFT IS >= 10 MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	3%	512

## Step3: Evaluation of the database rules

- When we calculated the profit for each database rule, we got the maximum profit of \$13,625.  
 → The profit increased from \$10,777 to \$13,625.

No.	Conditions	Result	Profit
1	AVGGIFT IS >= 8 RAMNTALL IS >= 40	DONATION AMOUNT IS >= 20	13,625
2	AVGGIFT IS >= 8 LASTGIFT IS >= 10 RAMNTALL IS >= 40	DONATION AMOUNT IS >= 20	13,514
3	MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	13,096
4	AVGGIFT IS >= 8 MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	13,078
5	AVGGIFT IS >= 8	DONATION AMOUNT IS >= 20	13,048
6	LASTGIFT IS >= 10 MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	13,002
7	AVGGIFT IS >= 8 LASTGIFT IS >= 10	DONATION AMOUNT IS >= 20	12,999
8	AVGGIFT IS >= 8 LASTGIFT IS >= 10 MAXRAMNT IS >= 17	DONATION AMOUNT IS >= 20	12,987
9	MAXRAMNT IS >= 17 RAMNTALL IS >= 40	DONATION AMOUNT IS >= 20	12,904
10	AVGGIFT IS >= 8 MAXRAMNT IS >= 17 RAMNTALL IS >= 40	DONATION AMOUNT IS >= 20	12,890

## Step4: Optimization of the database rules

- Combination of the database rules
  - Combined (the IF-parts of) the database rules up to 5 by logical OR, and found the target donor conditions that makes the profit maximal.
  - This requires huge amount of computing: The number of combinations is  
 ${}_{160}C_5 = \text{about } 10^9$
  - Using the backtracking and tree-pruning method of the AI technology, the number of combinations actually processed became about 200,000.

## Step4: Optimization of the database rules (continued)

---

- DCI got the profit of \$14,855, which is the (unofficial) world record. It took about 4 days.
- The target donor condition was complex, it consisted of union of (the IF-parts of) 5 database rules. Please see the next page.

## Step4: Optimization of the database rules (continued)

- (Average donation amount IS  $\geq$  \$8  
AND Income level IS above middle  
AND Total donation amount IS  $\geq$  \$40 )
- OR**  
(Average donation amount IS  $\geq$  \$8  
AND Maximum donation amount IS  $\geq$  \$17  
AND Total donation amount IS  $\geq$  \$40  
AND Wealth level IS above middle)
- OR**  
(Average donation amount IS  $\geq$  \$8  
AND Socio-Economic status of the neighborhood IS high  
AND Total donation amount IS  $\geq$  \$40)
- OR**  
(Socio-Economic status of the neighborhood IS high  
AND Maximum donation amount IS  $\geq$  \$17)
- OR**  
(Income level IS above middle  
AND Maximum donation amount IS  $\geq$  \$17  
AND Donor Rank IS 'F')

## Step4: Optimization of the database rules (continued)

- Usage of human experience
  - We found the following human rule from past experiences:

```
IF (Residence state IS Florida  
    AND Average donation amount IS >= $20  
    AND Donor Rank IS 'G')  
THEN Donation amount IS >= $20
```

- Combining this rule to the previous condition, we got the profit of \$15,013. The profit was more increased by \$158. **This is the NEW world record.**

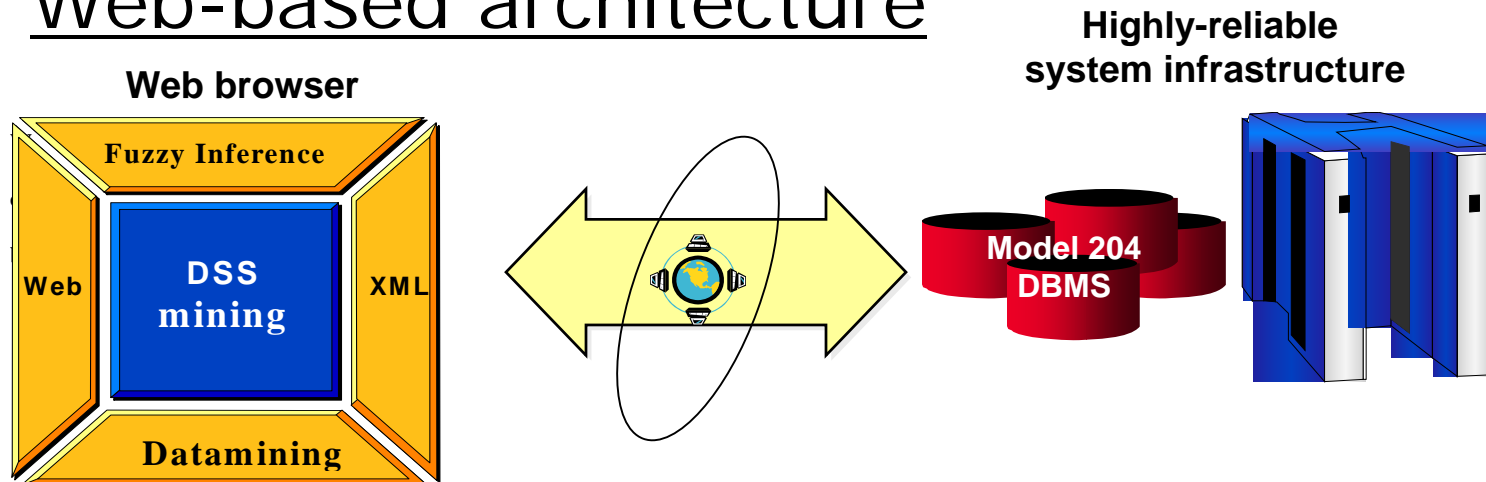
## Step4: Optimization of the database rules (continued)

---

- Please note the followings:
  - In this case, human rule improved the profit. But some human rules may decrease the profit. This is because human rule is not absolute but relative.
  - To handle human rules systematically, we use “fuzzy theory” to incorporate them into our application.

# System Architecture of DCI's DSS datamining

## ■ Web-based architecture



## ■ DSS system configuration

- Datamining Subsystem
- Rule evaluation/Fuzzy inference Subsystem
- UL-based
- XML-capable

# Challenge to KDD Cup 2007

---

- Prediction of customer's preference:  
There given 100 million records of Netflix customers's preference for rental DVDs during last 8 years.  
Predict their preference of this year.
- You have to register by June 1.
- You have to submit results by July 2.
- <http://www.cs.uic.edu/~liub/Netflix-KDD-Cup-2007.html>

# Challenge to KDD Cup 2007 (continued)

---

- DCI will challenge the KDD Cup with using the DSS datamining technology.
- We are discussing with Dr. Takahara for the details and start solving soon.
- We will do the best to get the award!
- Related to this year's KDD Cup, Netflix is holding another competition, whose grand prize is \$1,000,000.

# At the Last

---

- DCI's DSS datamining is ready for use.
- DCI's DSS datamining can be applied to your Business Intelligence projects.
- Let's try to use DCI's DSS datamining, and improve your business activity.
- Please contact me, [okada@database.co.jp](mailto:okada@database.co.jp)

# Question ?

---