Using Data Mining Methods to Build Customer Profiles (2001)

김지연
데이터베이스 특강 (추천시스템)
Dept. of CSE
Seoul National University

Overview

- The paper has developed an approach that uses information learned from customers' **transactional histories** to construct accurate, comprehensive **individual profiles**.

- **Discovering rules and validating them** are two important key phases of the idea it proposes and the system it implemented.

Outline

- **Problem Definition**
- Proposed Method
- Implementation
- Experiments
- Conclusion

Motivation

- Personalization has become an important marketing tool.
- Personalization community must deal with an issue of **how to extract knowledge from the available data and store it in customer profiles**.

Data Model

- Two basic types of data
  - Factual and transactional
    - **Factual data** : who the customer is
    - **Transactional data** : what the customer does
    - e.g. Customer's purchasing history

Factual Data	Transactional Data
Name, gender, birth date, address, salary, social security number, etc.	Purchase date, product purchased, amount paid, coupon use, coupon value, discount applied, etc.

Factual

CustomerId	LastName	FirstName	BirthDate	Gender
0721134	Doe	John	11/17/1945	Male
0721168	Brown	Jane	05/20/1963	Female
0730021	Adams	Robert	06/02/1959	Male

Transactional

CustomerId	Date	Time	Store	Product	CouponUsed
0721134	07/09/1993	10:18am	GrandUnion	WheatBread	No
0721134	07/09/1993	10:18am	GrandUnion	AppleJuice	Yes
0721168	07/10/1993	10:29am	Edwards	SourCream	No
0721134	07/10/1993	07:02pm	RiteAid	LemonJuice	No
0730021	07/10/1993	08:34pm	Edwards	SkimMilk	No
0730021	07/10/1993	08:34pm	Edwards	AppleJuice	No
0721168	07/12/1993	01:13pm	GrandUnion	BabyDiapers	Yes
0730021	07/12/1993	01:13pm	GrandUnion	WheatBread	No

Figure 2. Fragments of data in a marketing application showing demographic information (factual data) and records of the customers' purchases (transactional data).

Profile Model

- Two parts of complete customer profile
  - Factual and behavioral
    - **Factual profile** : profile models customer
    - **Behavioral profile** : profile models customer behavior
    - e.g. Customer's purchasing history

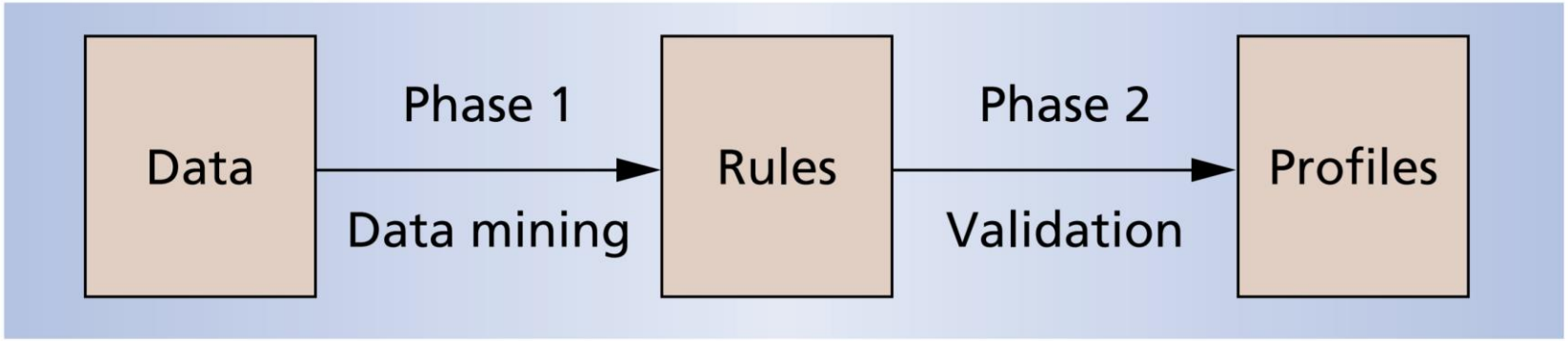
Factual Profile	Behavioral Profile
Name, gender, date of birth, "The customer's favorite beer is Heineken", "The customer's biggest purchase last monsth was for $237."	"When purchasing cereal, John Doe usually buys milk.", "On weekends, John Doe usually spends more than $100 on groceries."

Problem Definition

- Given: Large amount of customer data (factual and transactional)
- Do: *Discover rules and validate them*
- Output: Behavioral customer profile
- Such that: Under limited resource of experts who are able to validate rules

Two Main Phases of Profile Building Process

1. Rule Discovery



2. Rule Validation

1. Rule Discovery

- Individual customer behavior can be modeled with various types of **conjunctive rules**.

Figure 3. Association rules discovered in a marketing application help to describe the customer's behavior.

Advantage of Rule Discovery

- Rule discovery offers intuitive and descriptive way to represent behaviors.

- Conjunctive rule is a well-studied concept in many other fields such as
  - Data mining
  - Expert systems
  - Logic programming
  - Literature

Rule Discovery in Personalization

- For personalization applications, rule discovery methods are **applied individually** to every customer's data.
  - Various data mining algorithms can be used to discover rules that describe the behavior of individual customers.
    - Apriori for association rules
    - CART(Classification and Regression Trees) for classification rules
    - Moreover, the paper's profiling approach is not limited to any specific representation of data mining rules or discovery methods.

2. Rule Validation

- Datamining methods often generate large numbers of rules that are trivial or not relevant to the application at hand.
  - Therefore, validating the discovered rules is an important requirement.
  - e.g. Whenever John Doe takes a business trip to Los Angeles, he stays in an expensive hotel.
    - John went to Los Angeles seven times over the past two years, and five of those times he stayed in expensive hotels.
    - We must validate that the rule captures John's behavior rather than a spurious correlation and it is not simply correlevant to the application.

How to Validate Rules

- **Inspection by the experts by letting them decide how well they represent customers' actual behaviors**

- Acception and rejection of rules
  - Accepted rules form behavioral profile

Challenges

- Important issue in validation : ***SCALABILITY***
  - The number of customer is very large.
  - It is simply impossible for a human expert to validate all the rules one by one.

  - e.g. in a credit-card application, the number of customers can be measured in millions. If we discover 100 rules per customer on average, the total number of rules in that application would be hundreds of millions.

Outline

- Problem Definition
- **Proposed Method**
- Implementation
- Experiments
- Conclusion

Profile Building Overview

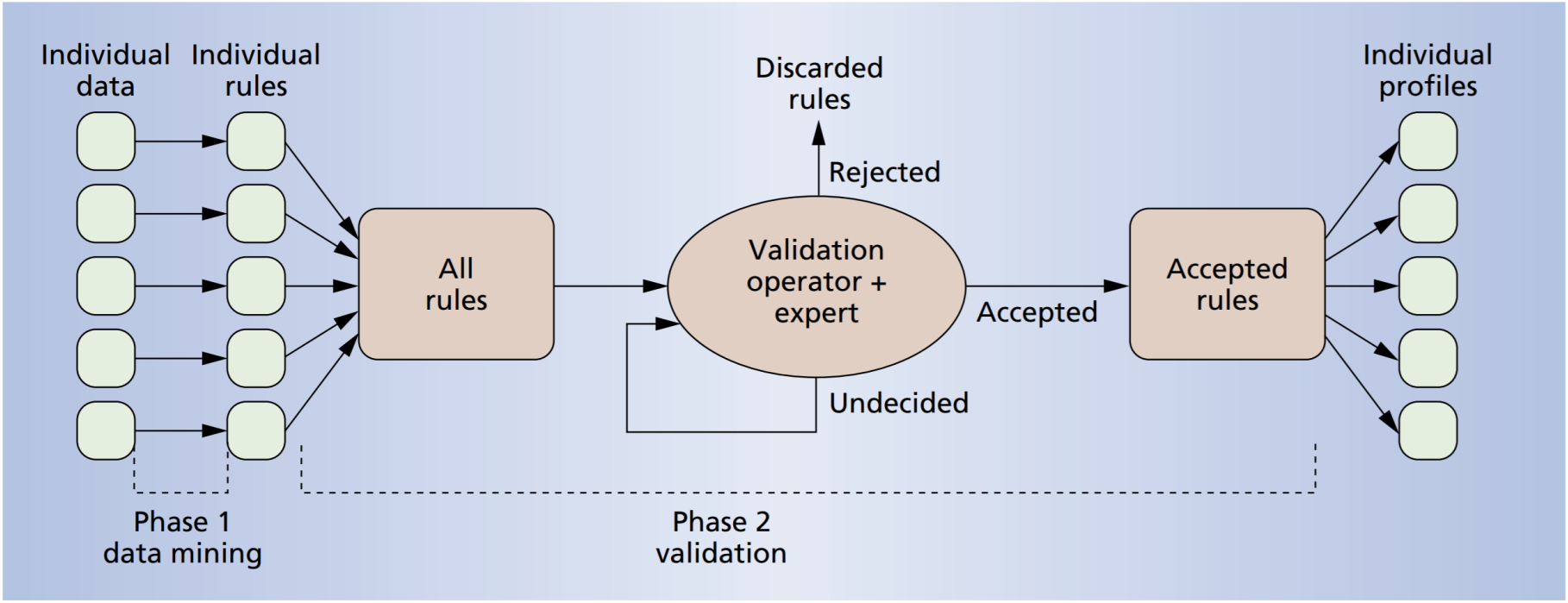


Figure 4. An expanded view of the profile-building process. Rule validation is an iterative process in which the expert applies various operators successively to validate rules.

Collective Rule Validation

- Rule validation is not a separate process for each customer, but takes place for all customers at once.

- As a result, the expert usually validates many similar rules for different customers.
  - e.g. “When buying cereal, John Doe also buys milk.”,”When shopping on weekends, John Doe usually spends more than $100 on groceries.”

Validation Mechanism

- The system collects rules from all the customers into one set and tags each rule with the Id of the customer to which it belongs.
  - After validation, the system places each accepted rule in that customer's profile.

Input: Set of all discovered rules R_{all} .
Output: Mutually disjoint sets of rules R_{acc} , R_{rej} , R_{unv} , such that $R_{all} = R_{acc} \cup R_{rej} \cup R_{unv}$.

(1) $R_{unv} := R_{all}$, $R_{acc} := \emptyset$, $R_{rej} := \emptyset$.
(2) **while** (**not** *TerminateValidationProcess()*) **begin**
(3) Expert picks a validation operator (say, O) from the set of available validation operators.
(4) O is applied to R_{unv} . Result: disjoint sets O_{acc} and O_{rej} .
(5) $R_{unv} := R_{unv} - O_{acc} - O_{rej}$, $R_{acc} := R_{acc} \cup O_{acc}$, $R_{rej} := R_{rej} \cup O_{rej}$.
(6) **end**

Validation Operators

- Validation operators allow human experts to validate large numbers of rules with **relatively little input** from the expert.
  - Similarity-based rule grouping
  - Template-based rule filtering
  - Redundant-rule elimination

Similarity-based Grouping

- This operator puts similar rules into groups according to expert-specified similarity criteria

Figure 3. Association rules discovered in a marketing application help to describe the customer's behavior.

  - Attribute structure similarity condition
    - (1), (2) both have *Product => Store*
    - (3) would not be grouped with (1) and (2) since *Product => CouponUsed*

Template-based Filtering

- This operator filters rules that match expert-specified rule templates.

Figure 3. Association rules discovered in a marketing application help to describe the customer's behavior.

  - Consider the following rule templates
    - *REJECT HEAD = {Store = RiteAid}*
    - *ACCEPT BODY ⊇ {Product} AND HEAD ⊂ {DayOfWeek, Quantity}*

Redundant-rule Elimination

- This operator eliminates rules that can be derived from other, usually more general, rules and facts.
  - Redundancy condition
    - *Product = AppleJuice => Store = GrandUnion (2%, 100%)*

Other Validation Operators

- Visualization operator
  - It lets the expert view subsetso f unvalidated rules in visual representations
  - Pie charts, histograms
- Statistical-analysis operator
  - It computes various statistical characteristics of unvalidated rules, thus providing the expert with important information to use during validation.
- Browsing operator
  - It allows the expert to inspect individual rules or groups of rules directly by viewing them on the screen.

Outline

- Problem Definition
- Proposed Method
- **Implementation**
- Experiments
- Conclusion

Implementation

- **1:1Pro system**
  - One-to-One Profiling
  - Input : factual and transactional data stored ia database or flat files
  - Do: generate a set of validated rules capturing individual customers' behavior.

  - DBMS and various data mining tools
    - It corporates many tools useful to validate rules.

1:1Pro Architecture

- Client-server model

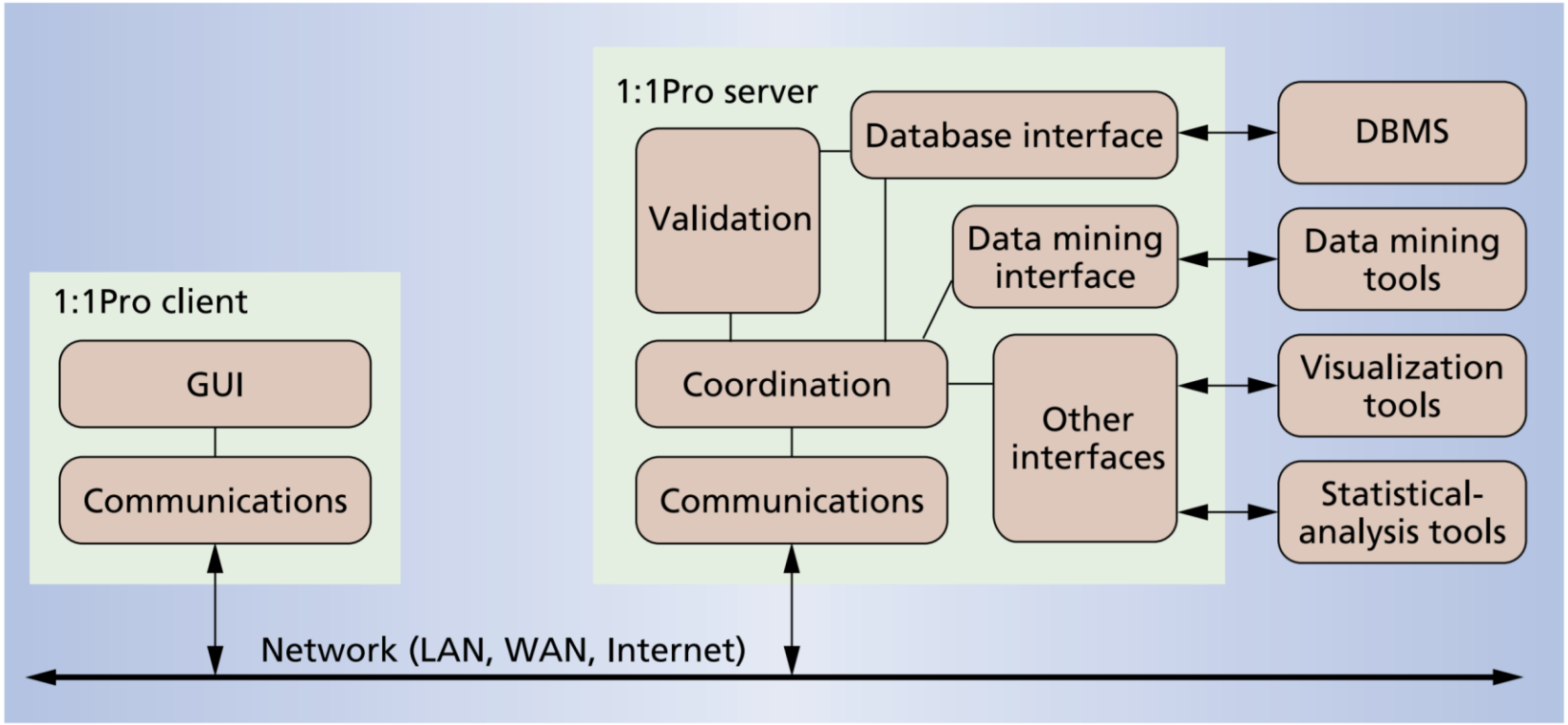


Figure 6. The 1:1Pro system architecture. 1:1Pro is an open system that incorporates a broad range of data sources as well as data mining, visualization, and statistical-analysis tools.

Server-side Components

- Coordination module
  - Profile construction
- Validation module
  - Rule validation
- Communications module
  - Communication handling with the client
- Interfaces to external modules
  - DBMS, datamining tools, visualization tools

Client-side Components

- Client communication module
- GUI

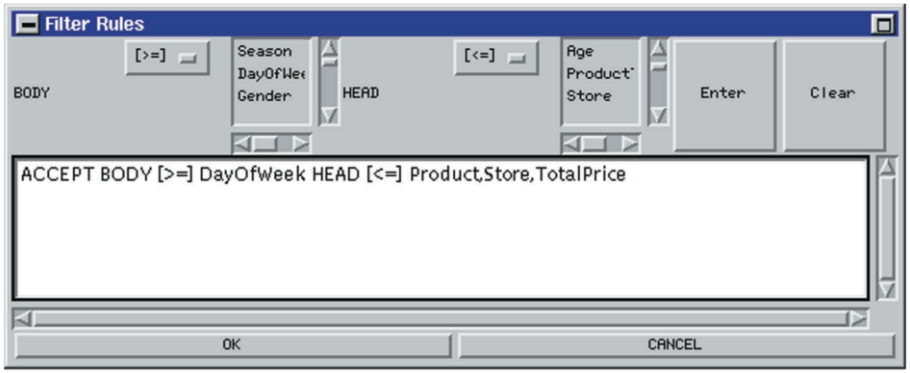


Figure 7. Graphical user interface window for a filtering operator. The expert uses the GUI to specify validation operations and view the results of the iterative validation process.

- Log file
  - ResultId, Operator, SourceId, Data/Time, Notes

ResultId	Operator	SourceId	Date/Time	Notes
...	...	...	...	...
6	Filter	5	11/23/1998 5:26pm	Rejecting: demogr. in the body
7	Group	3	11/23/1998 5:37pm	Used attribute-level setting here
8	Browse	7	11/23/1998 5:51pm	Accepted: 7 groups, rejected: 11
9	Filter	3	11/23/1998 6:28pm	Rejecting: 'age' in the head
...	...	...	...	...

Figure 8. A 1:1Pro system log file fragment. The log file captures the entire validation process, allowing the expert to keep track of all validation activities.

Outline

- Problem Definition
- Proposed Method
- Implementation
- **Experiments**
- Conclusion

Experiments

- Testing on a real-world marketing application
  - Included data on 1,903 households purchasing various non-alcoholic beverages over a one-year period
- Seasonality analysis
  - The experiment constructed customer profiles containing individual rules describing season-related customer behaviors

- Generated 1,022,812 association rules (about 537 rules per household)
  - Most rules pertain to a very small number of households
  - Many discovered rules capture truly idiosyncratic behavior of individual households.
  - Contrast with traditional segmentation based approaches
  - Several discovered rules are applicable to a significant portion of the households.
    - *"DayOfWeek = Monday => Shopper = Female"* (859 households)
  - The system can validate a significant number of rules in medium-size personalization application.
  - The average customer profile size was reduced from 537 unvalidated rules to 21 accepted rules.

Table 1. A validation process for the seasonality analysis of market research data.

Validation operator	Accepted rules	Rejected rules	Unvalidated rules
Redundancy elimination	0	186,727	836,085
Filtering	0	285,528	550,557
Filtering	0	424,214	126,343
Filtering	0	48,682	77,661
Filtering	10,052	0	67,609
Grouping (652 groups)	23,417	6,822	37,370
Grouping (4,765 groups)	7,181	1,533	28,656
Total	40,650	953,506	1,724,281

Limitations

- The accepted rules, although valid and relevant to the expert, may not be effective.

- The system generates many irrelevant rules that are subsequently rejected during the validation process.
  - Solution : Specifying constraints on the type of rules of interest

Outline

- Problem Definition
- Proposed Method
- Implementation
- Experiments
- **Conclusion**

Conclusion

- To build accurate and representative customer profile, two-phase process which consists of rule discovery and validation stage is effective.
- When validating rules, validation operators may help experts deal with relatively few rules out of enormous amount of customer data.
- The implementation of this process, 1:1Pro showed better performance than

Thank you !