



# Can We Find Common Rules of Browsing Behavior?

**Ganesan Velayathan**, Seiji Yamada  
The Graduate University for  
Advanced Studies  
National Institute of Informatics  
Tokyo, Japan



## Today, we will discuss:

- How to personalize
- Our approach to personalization
- Experiments conducted
- Some early conclusions
- Ongoing and future work



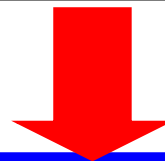
# Introduction

---



**Too much information** but **too little time**  
to evaluate all the pages we look at!

We need a **system to assist**  
in our daily **web tasks**



**Web Personalization**



However...



**When focusing on Web Personalization**

**Often, we see**

**We spend **more time telling** the system to learn rather than letting the system help us**

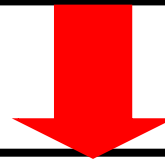
**Don't you agree with me...?**



# Why?



It's often hard to know which web pages  
a user “likes” or “dislikes”



Users need to teach the system  
their individual preference information

This is called **building a profile**



So, what do we need to do?



**We need to automatically evaluate  
the sites we visit**

**This will save time, user burden, & stress**

**Don't you agree...?**



## But how to do it automatically?



When we browse, we display certain habits  
(whether we “**like**” or “**dislike**” the webpage)

So, we need to...

*Determine, log & study these user habits*

to find:

Rules of behavior when the user shows  
“**interest**” or “**non-interest**”



# What is User Behavior?

---



# What is User Behavior?



Behavior is **all the things we do**  
when we use the browser

Some behaviors are performed  
intentionally, and some we never realize...



**Let's take a look!**

---



# Demonstration of User Behavior

---



# Examples of user behavior



Print Page

Bookmark Page

Copy Text



# How are these performed?



Print Page

Ctrl + P, Click Print Button, File → Print

Bookmark Page

Ctrl + D, Click Bookmark, Favorite → Add

Copy Text

Click, Select, Ctrl + C



# Some Important Terms



## **Navigation Actions:**

*The individual components of user behavior*

## **User Behavior:**

*The intended results of performing these navigation actions*



# Why User Behavior?



Habits do not change over the short term



Evaluation can be performed automatically



Content-independent



Language-independent





**How can we collect these user behaviors?**

---



# 3 ways



**Server Side**



**Proxy Side**



**Client Side**





**We chose...**

---



**...the client side**

---



**Why?**

---



# Why use the client side?



Easier to collect high-granularity data

A survey by **Shahabi and Chen** [2003] pointed out that web usage data from the server side is not reliable



**So let's go!**

---



# This study involves...



**1) Logging the user's habits (user behavior)**



**2) At the client side**



**3) Learning these behaviors**



**4) Seeing common patterns / rules**

---



**So, what do we need?**

---



# GINIS Framework



**Browser**



**Logger Database**



**Learner**



**Predictor**





- Browser → Client interface, similar to IE
  - Logger → To log database
  - Analyzer → To clean log & perform machine learning
  - Predictor → To predict user interest based on pre-defined behavior
-



# Building the Browser

---



# Building the Browser



**A client-side logging tool**



**Mimics Internet Explorer 6, with extra features**



**Ajax & form input plug-in**



**Tab browsable, right click extension**

---



# Logging Behavior

---



# Logging Behavior



**70 navigation actions**



**40 user behaviors**



**Real-time user interaction logger**





# Machine Learning

---



# Machine Learning



## C4.5 ML Algorithm





# Why C4.5 ?



**Most widely used CL in practice**



**Fastest main-memory CL [Lim 00]**



**Rule-generating capability**



**Rules are “human-readable”**



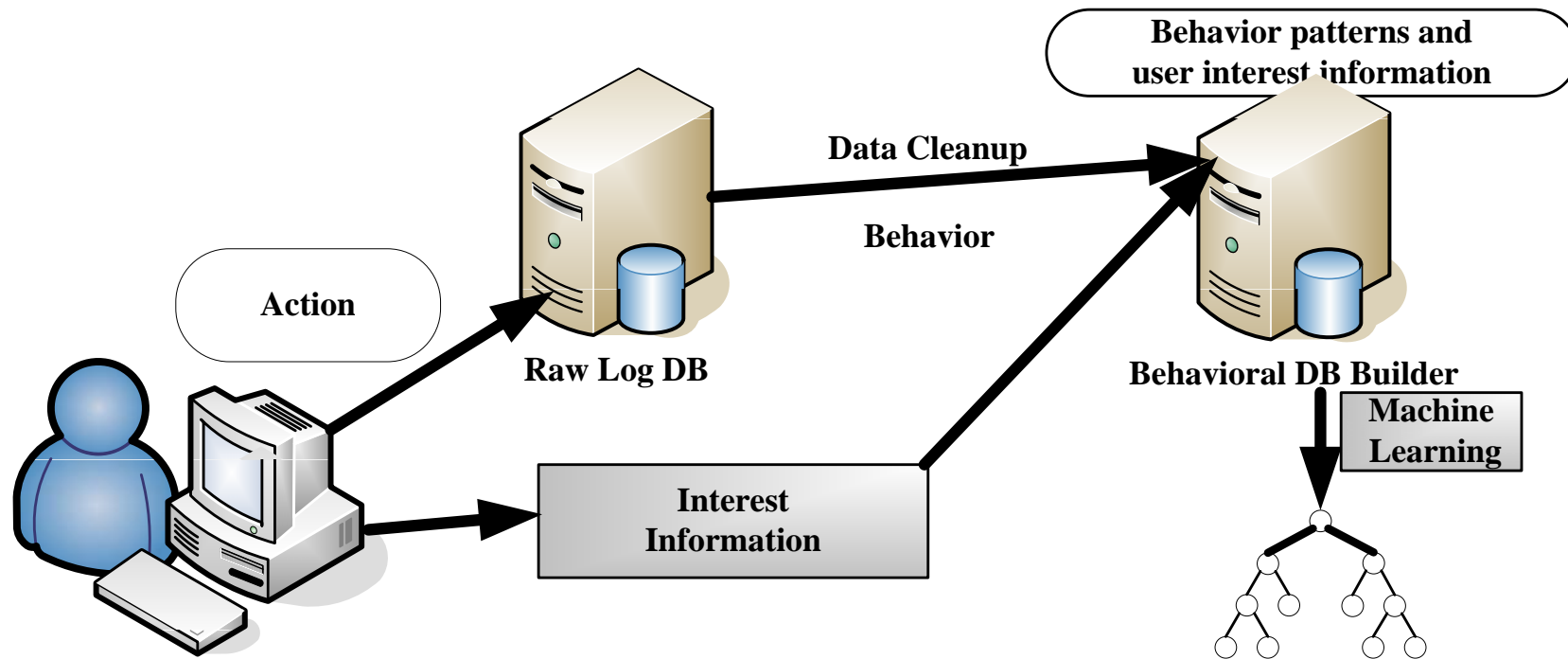


# System Overview





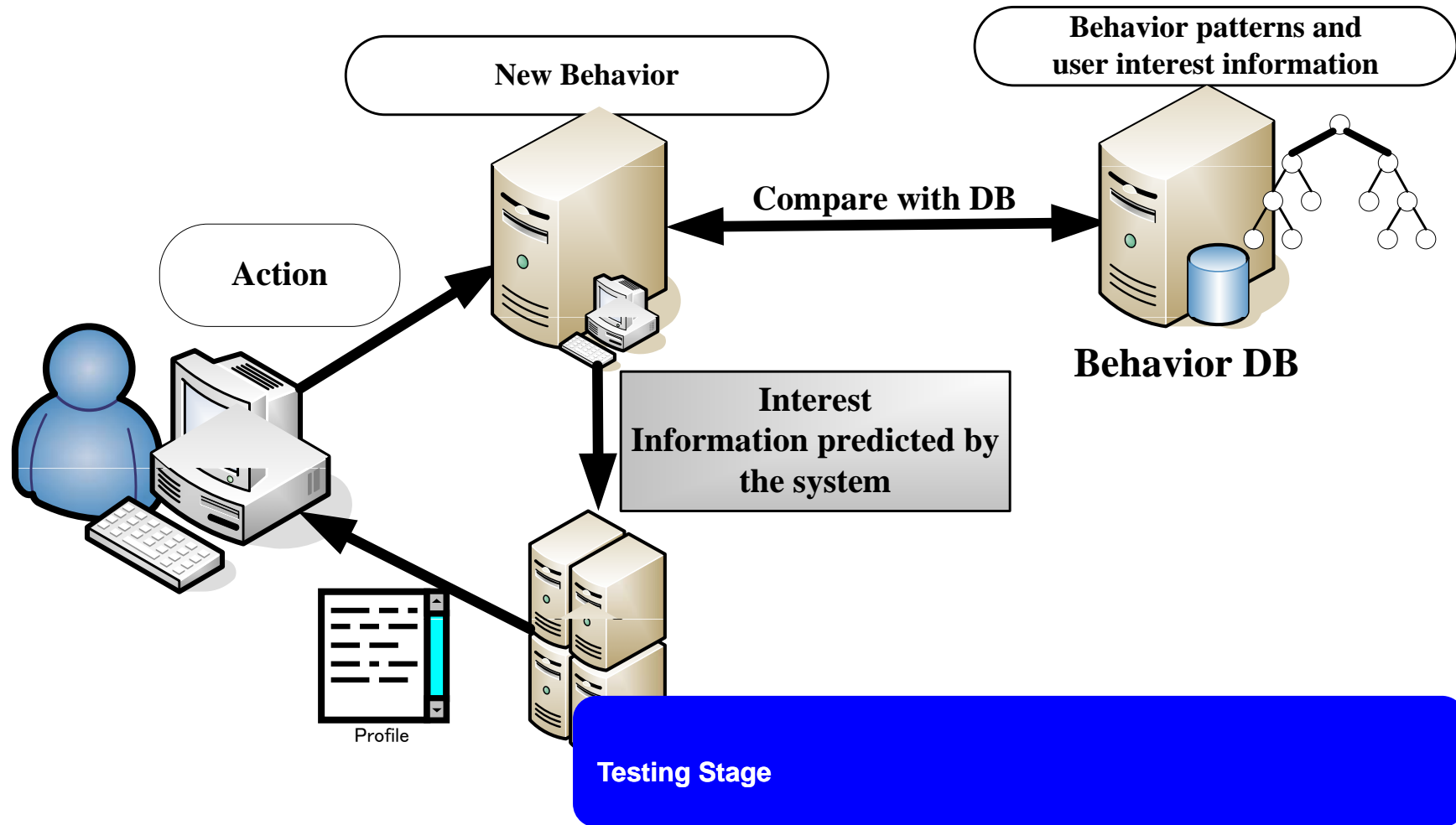
# System Overview (1)



Learning Stage



# System Overview (2)





**So let's go on to...**

---



# Experimental Stage

---



# Setup



Performed in an open environment  
(not a controlled lab environment)

Why?

- ✓ Habits might change if controlled
- ✓ Daily environment is important



# Gathering Data

---



# Statistical Information



**Participants:** 6 male, 4 female (10 persons)

**Age:** 21 – 38 years old (mean: 29.1 years)

**Experience:** 4 - 12 years (mean: 8.3 years)

**Duration:** 31 days (mean duration: 22 days)



# Pre-Processing



**Data :** 460,000 lines of raw log data

**Clean-Up :** 65% of data cleaned up

**Cleaned-Up:** Mouse locus info, copied text, etc.

**Remaining:** 65,000 lines

---



# Results





# Summary of Data



## Most Frequent 5

Scroll	19091	times
Key Input	14188	times
Form Input	9329	times
Copy Text	1685	times
Search Text	1284	times

## Least Frequent 5

Go Forward	126	times
Stop Loading	88	times
Add to Favorite	79	times
Print	64	times
Save As	2	times

**Average**

**5.2 behaviors per page**



# User Evaluation



**Total Pages**

**2,856 pages**

**“Of Interest”**

**1,997 pages**

**“Not of Interest”**

**859 pages**



**C4.5 CL**

---



# C4.5 CL



**Confidence Factor**

**0.25**

**Validation**

**10-fold cross validation**

**Correctly Classified**

**2027 pages (70.97%)**

**Incorrectly Classified**

**829 pages (29.03%)**



# CL Results



User Evaluation	Classified as "Of Interest"	Classified as "Not of Interest"	Total
"Of Interest"	1852	145	1997
"Not of Interest"	684	175	859



# Rules Generated

---



# Rules of "Interest"



**Rule 1 [86.1%]**

**Scroll > 14 ^ Reload <= 0**

**Rule 2 [85.4%]**

**Stay Time <= 14s ^ Search Text > 0 ^  
Form Input <=0 ^ Key Input <=0**

**Rule 3 [83.5%]**

**Reload <=0 ^ Form Input > 2 ^  
Navigate <=0 ^ Go Back > 0 ^ Go Back < 1**

**Rule 4 [73.8%]**

**Scroll < 3 ^ Reload <= 0 ^ Search Text <= 0 ^  
Go Back <= 0**

**Rule 5 [73.0%]**

**Search Text > 0 ^ Key Input > 0**



# Rules of "Non-Interest"



**Rule 6 [82.0%]**

**Form Input > 2 & Go Back > 1**

**Rule 7 [72.2%]**

**Stay Time > 5 ^ Scroll > 3 ^ Scroll <= 7 ^  
Search Text <= 0 ^ Form Input > 1 ^ Go Back <= 0**

**Rule 8 [72.0%]**

**Stay Time > 11 ^ Scroll <= 0 ^ Search Text <= 0 ^  
Form Input > 0 ^ Form Input <= 1 ^ Go Back <= 0 ^  
Copy <= 0 ^ Paste Text <= 0**

**Rule 9 [67.5%]**

**Stay Time > 4 ^ Scroll > 3 ^ Form Input > 1 ^  
Key Input <= 0 ^ Go Back <= 0**

**Rule 10 [59.8%]**

**Reload > 0 ^ Key Input <= 1 ^ Copy <= 0**



# Discussion



**Classification of “Non-Interest” is not good yet**



**Same behavior, but different evaluation**





# Further Cleanups

---



# Clean up Again



**Total Pages**

**2,249 pages**

**“Of Interest”**

**1,885 pages**

**“Not of Interest”**

**364 pages**



# C4.5 CL(2)



**Confidence Factor**

0.25

**Validation**

10-fold cross validation

**Correctly Classified**

2005 pages (89.15%)

**Incorrectly Classified**

224 pages (10.85%)



## CL Results (Better?)



User Evaluation	Classified as "Of Interest"	Classified as "Not of Interest"	Total
"Of Interest"	1808	77	1885
"Not of Interest"	167	197	364



# Future Work



**Same behavior, but different evaluation  
What can we do about this?**



**What are the individual rules  
for each participant?**





# Summary & Conclusions



We built a client side logging browser and considered connections between user interest and user behavior

We proposed and evaluated a better, more straightforward method of evaluating web pages