大阪市立大学
OSAKA CITY UNIVERSITY

Identification of User Behavior from Flow Statistics

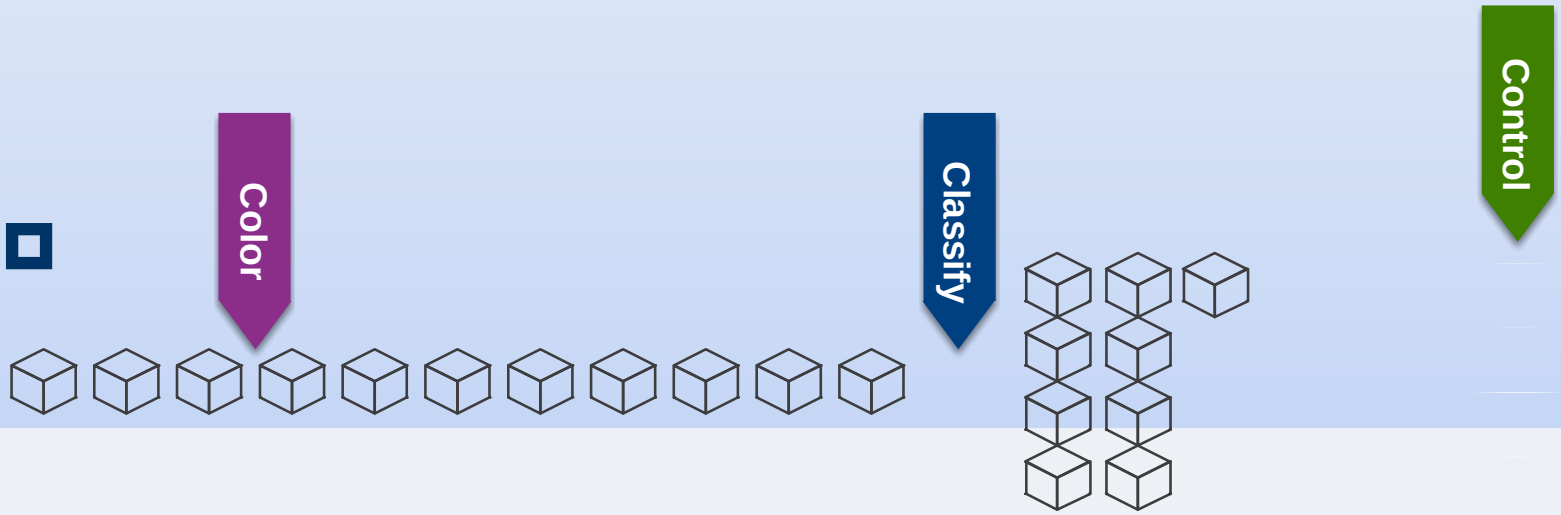
APNOMS 2017, September 2017

Shingo Ata

Graduate School of Engineering
Osaka City University

Coloring traffic and control

- Internet applications are diversified
  - Web browsing, file downloading, multimedia delivery, social networking services, cloud services
- Application-based traffic management is a key
  - To achieve QoE (Quality of Experience) to end users
  - To realize efficient management of networking resources



Problem statement

- Limitation of single flow identification
  - Recent applications use multiple flows (e.g., TCP connection)
    - Simultaneously in parallel
    - To improve the users' experience (e.g., latency, throughput, and response time)
  - Identification of single flow is insufficient

App/Svc	# flows (approx.)
Video Sharing	40
Portal Site	100
Social (micro-blogging)	20
Social (portal)	40
Web Shopping	80
Screen Sharing	1
Secure Shell	1

Problem statement (cont'd)

- Video is not always important
  - Importance of content depends on user's behavior
  - Web portal, timelines on social networking
    - Multiple content types in a single page

Introduction of “User Behavior”

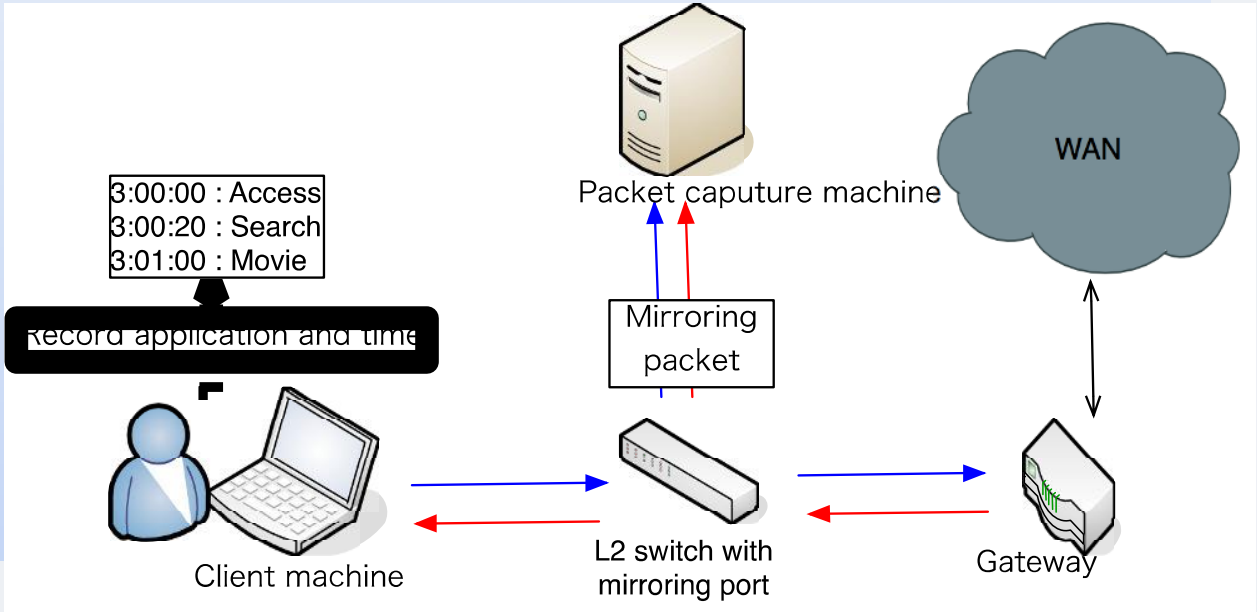
- Importance of flows are strongly related to real actions taken by users
  - Controlling the most important flow for the user can directly improve the user’s overall QoE
- User behavior
  - Detailed actions taken in the application (or service)
- Identification of user behavior
  - From statistics of multiple flows
  - Not focus on an individual flow but focus on multiple flows associated to application
- Extend application identification method to handle statistical relations among flows
  - Use functions to represent the relation of multiple flows
  - Use ML (Machine Learning) based algorithm for identification

Target behaviors (9 apps, 43 behaviors)

Application	URL	User behaviors
Youtube	www.youtube.com	Playing video, Search video, My channel, Authorization (login), Top page
Google	www.google.com	Top page, Search result, Image search result
Yahoo! Japan	www.yahoo.co.jp	Top page, Search, News (text only), News (with video)
Amazon	www.amazon.com	Top page, Login, Product search, Product details, View shopping carts
Facebook	www.facebook.com	Login, Timeline, Post (text), Post (with pictures), Profile
Gmail	www.gmail.com	Inbox, Send/Receive mails, Open mail
Skype	www.skype.com	Waiting, Calling, Video conference, Short message, File exchange
Dropbox	www.dropbox.com	Application initialization, Syncing, Upload, Delete, Name change, Folder creation
Twitter	www.twitter.com	Timeline, Posting tweets, Posting images, Top page

Measurement environment

- Behavior scenarios
  - Create a set of operations in every application/service
- Packet capture and flow analysis



Basic process of application identification

- Traffic is classified into flows (e.g., w/ 5-tuples).
- A set of traffic features is obtained for each flow.
  - Each flow has a multidimensional vector (f1, f2, …, fm).
- Supervised ML algorithm is applied to identify the application.
  - Training data is used as supervisor.

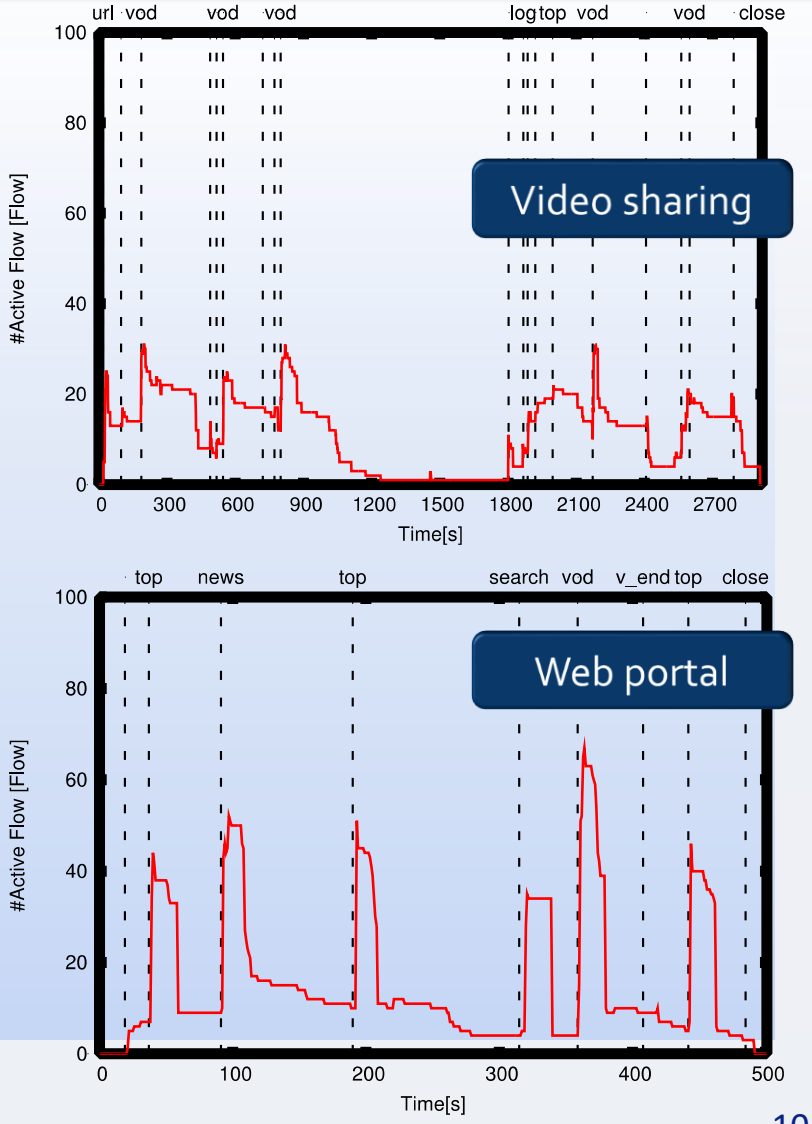
Traffic features used

- Packets -> flows
  - Classified by 5-tuples
- Calculate traffic features for every flow (48 features in evaluation)

Category	Direction	Traffic Features
Packet size	C->S, S->C, both	Min, Max, Med, Avg, Dev, 25%, 75%
Packet inter-arrival Time	C->S, S->C, both	Min, Max, Med, Avg, Dev, 25%, 75%
Avg. packet size in time window	C->S, S->C, both	win=10sec
# packets	C->S, S->C, both	Total
Transmission speed	S->C	bps
# bytes	C->S, S->C, both	Total
# active flows	S->C	Total
Duration	both	Total

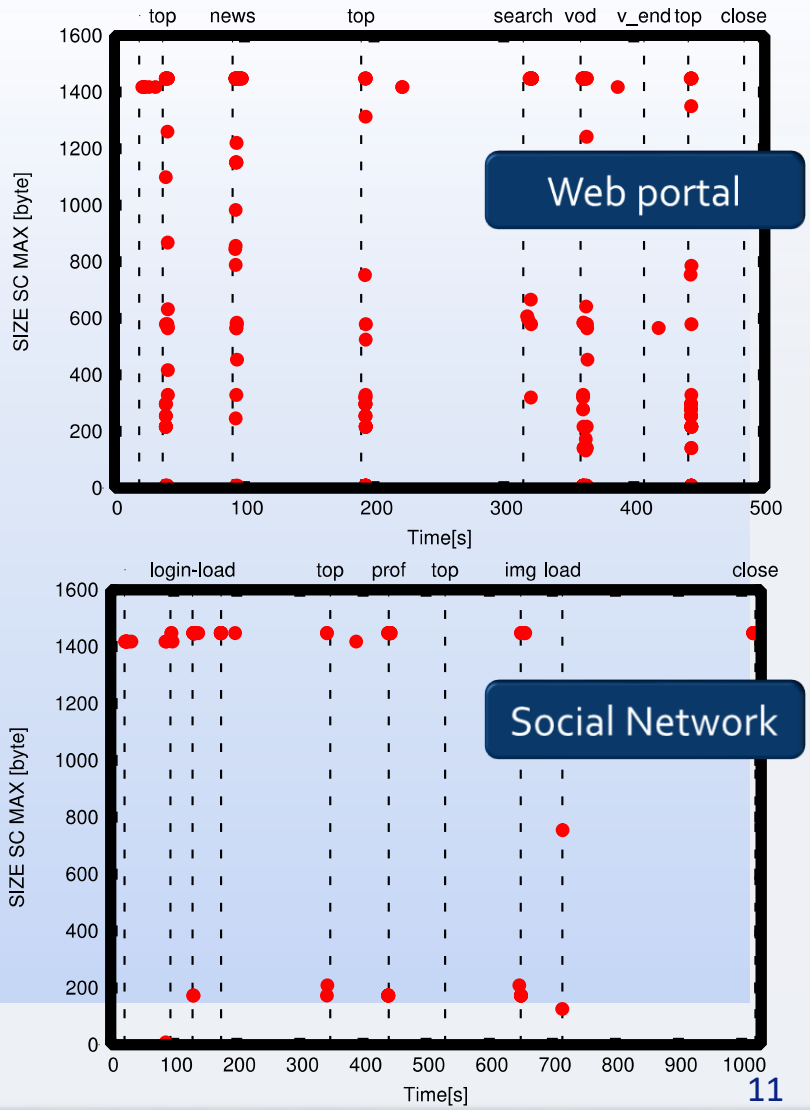
Preliminary example: # flows

- Main observations
  - # of active flows is significantly varied at the event of user behavior
    - Up to 70 in web portal
  - Video sharing is less sensitive than web portal
    - By SPDY and HTTP/2



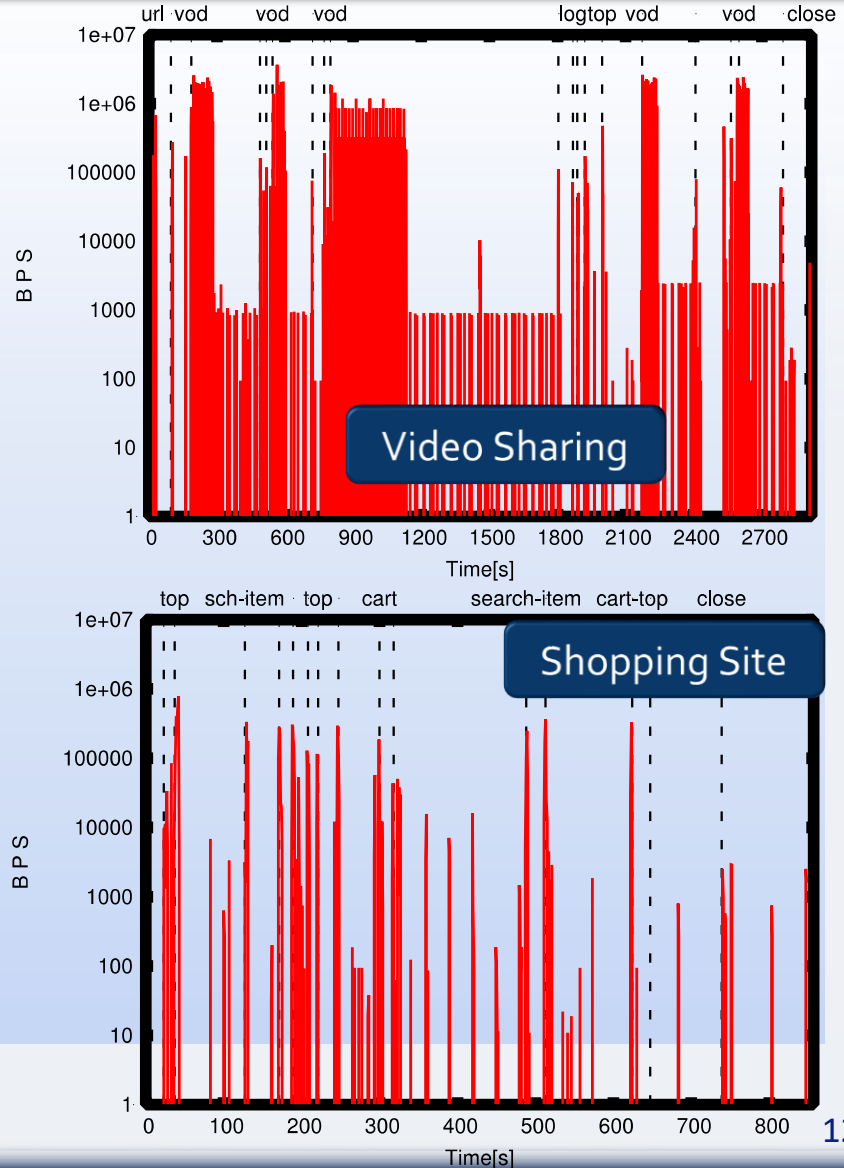
Preliminary example: Max pkt size

- Distribution of max packet size in flows
  - Widely distributed in web portal
    - Different contents from different sites
  - Almost two clusters in social network
    - By SPDY or HTTP/2.0
    - Reuse connections for different contents



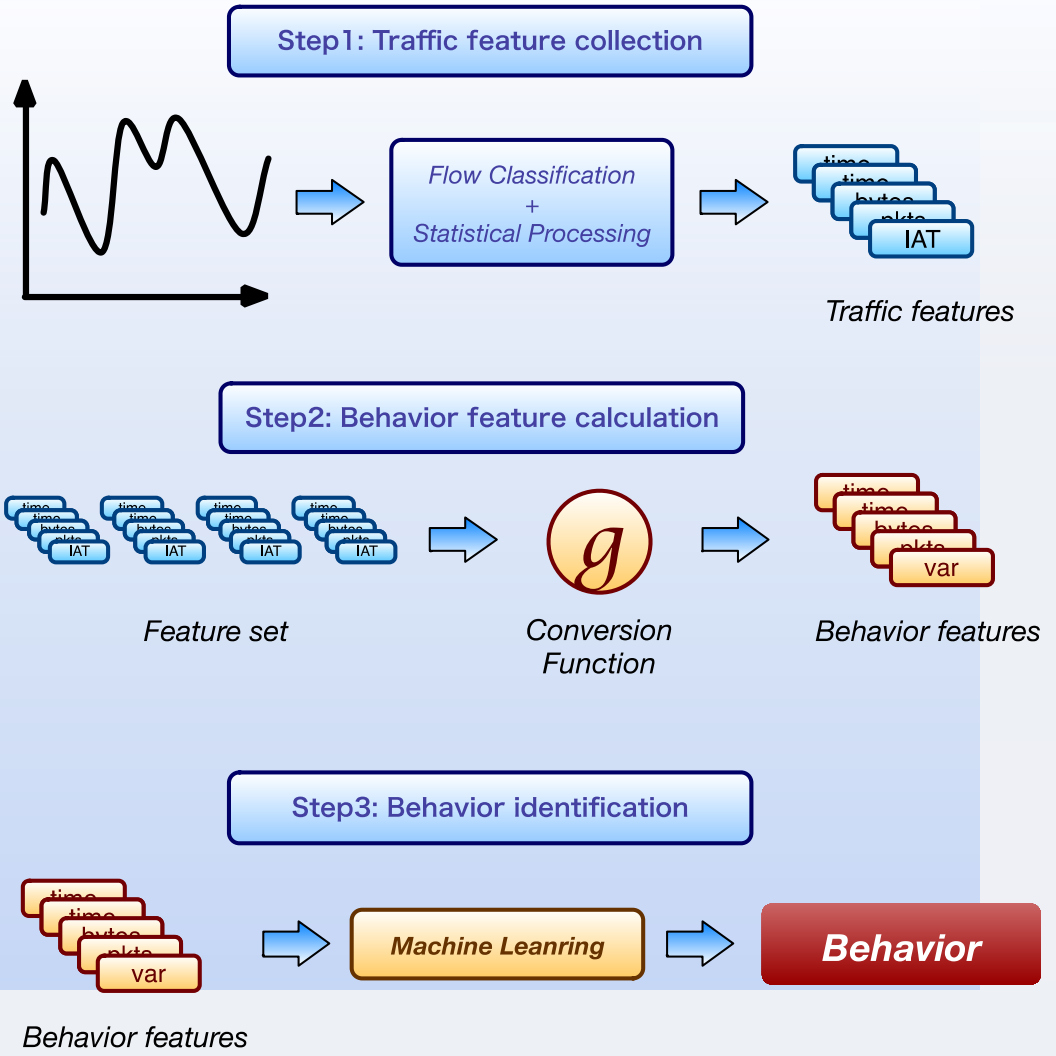
Preliminary example: transfer rate

- Different results between video and text or image
  - Video: constant and long
  - Text/image: varied and short



Outline of user behavior identification

- Step 1
  - Traffic feature collection
- Step 2
  - Behavior feature calculation
- Step 3
  - Behavior identification



Calculation of behavior features

- Application generates n flows for single behavior

$$W = \{w_1, w_2, \ldots, w_n\}$$

- Calculate traffic feature vector G for every flow

$$G = \{f_1, f_2, \ldots, f_k\}$$

- Group by traffic features

$$F_k = \{f_k^{w_1}, f_k^{w_2}, f_k^{w_3}, \ldots, f_k^{w_n}\}$$

- Calculate behavior feature

$f_k^{w_n}$ k-th traffic feature for flow w_n

$$b_k^j = R^j(F_k)$$

R^j : function to get j-th behavior feature from F_k

Functions for behavior features

14 functions

Average	Median Absolute Deviation
Standard Deviation	Variance Mean Ratio
Skewness	Geometric Mean
Kurtosis	Harmonic Mean
Bimodal Coefficient	Range
Coefficient of Variation	Trimmed Mean
Median	Interquartile Range

Implementation and evaluation

- 9 applications (43 behaviors) for identification
  - 1/3 of measured flows are used for training
  - 2/3 of flows are used for evaluations
- Machine learning algorithm
  - 48 x 14 = 672 features vector
  - SVM (support vector machine)
- Evaluation metric
  - Accuracy = (# of correctly identified behaviors) / (total # of behaviors)

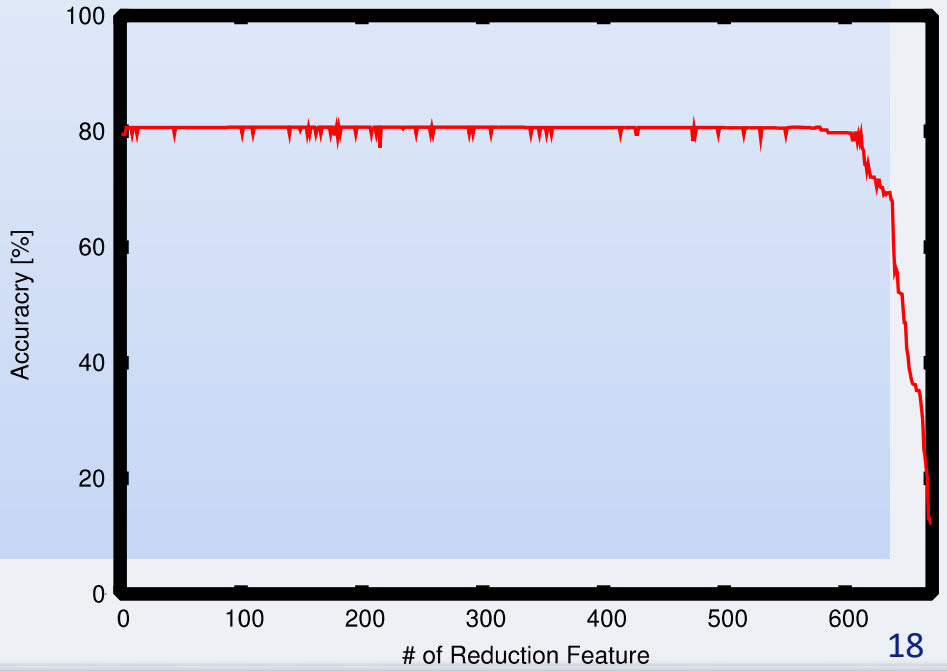
Identification results

- Application identification
  - Overall = 91%
- Behavior identification
  - Overall = 81%

App	Behavior	Accuracy (%)	App	Behavior	Accuracy (%)
Amazon (96%)	Buy	91	Gmail (90%)	Open	88
	Cart	80		Top	98
	Goods	91		Send	97
	Login	97	Google (92%)	ImgSearch	87
	Search	81		Search	55
	Top	92		Top	72
Dropbox (92%)	Start	93	Skype (92%)	Login	98
	Upload	90		Msg	98
	Sync	73		File	83
	Name	41		Video	64
	Folder	34		Voice	97
	Move	45	Twitter (84%)	Login	97
	Delete	33		Tweet	80
Face book (93%)	Load	82		Load	88
	Login	98		Image	81
	Image	82			
	Post	74			
	Profile	84			
	Top	98			

Impact to reduce # of features

- 672 features require huge computation
  - # of features should be as few as possble
- By applying SVM-RFE
  # of feature can
  be reduced 143



Summary and future works

- Identification of user behavior
  - Not to identify individual flows but identify real actions in application
  - Introduction of behavior features
  - Achieve over 80% of behavior identification
- Future topics
  - Increase applications and behaviors
  - Analyze impacts of contents or individual users