Web Data Mining

Lecture 9: Web Usage Mining/Web Analytics

Jaroslav Kuchař & Milan Dojčinovski
jaroslav.kuchar@fit.cvut.cz, milan.dojchinovski@fit.cvut.cz

Czech Technical University in Prague - Faculty of Information Technologies - Software and Web Engineering

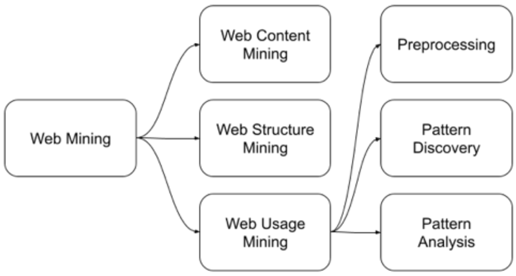
Summer semester 2019/2020
Humla v0.3

Overview

- Web Usage Mining
- Collecting and Preprocessing
- Pattern Discovery
- Web Analytics

Web Usage Mining (Recall)

- Motivation
  - *Huge amount of clickstream, transaction data, and user profile data*
- Main ideas
  - *Discover usage patterns from Web data to understand and better serve the needs of web-based applications.*
  - *Extracting useful information from server logs.*
  - *Process of finding out what users are looking on Internet.*
- Views
  - *Web Usage Mining*
    → research field
  - *Web Analytics*
    → practical application of the analysis of clickstream data

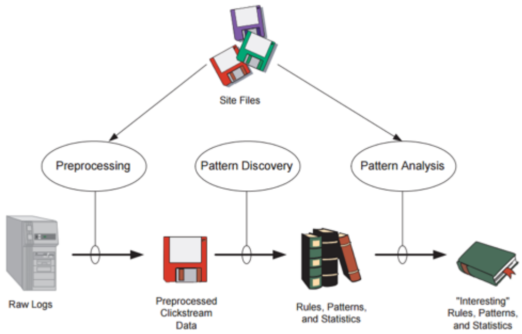


Web Usage Mining (cont.)

- Categories
  - *General Access Pattern Tracking*
    → Analysis of user patterns and general trends to get overview about the overall behavior.
  - *Customized Usage Tracking*
    → Analyses individual trends where the goal is to customize pages to individual users.
- Applications
  - *Usage characterization*
    → web analytics, …
  - *System improvement and site modifications*
    → web sites optimization, …
  - *Personalization*
    → user and customer behavior modeling
  - *Business intelligence*
    → web marketing/advertising
    → recommender systems

WUM Process

- Three inter-dependent stages
  - *data collection and pre-processing*
    - → *cleaning, transactions identifications*
    - → *enhancements - site structure, semantics, …*
  - *pattern discovery*
    - → *detection of hidden patterns using statistical, database, machine learning operations*
    - → *summary statistics resources, sessions, users*
  - *pattern analysis*
    - → *filtering, aggregations*
    - → *validations, interpretations*



Srivastava et al. (2000). Web usage mining: Discovery and applications of usage patterns from web data.

WUM Input Data

- Usage Data
  - *Primary source*
  - *Web and application server logs*
  - *Main information*
    - → *time, (IP address), resource + parameters, status, HTTP method, User-Agent, cookies (optionally)*
  - *Aggregations*
    - → *pageview*
      - → *collection of resources representing one user-action e.g. reading article, viewing a product page etc.*
    - → *session*
      - → *sequence of pageviews during one user visit*
  - *Apache Access Log Example:*

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] \
"GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.
"Mozilla/4.08 [en] (Win98; I ;Nav)"
```

WUM Input Data (cont.)

- Content Data
  - *Collection of objects and relationships*
  - *Usually textual data and multimedia (e.g. HTML, images)*
  - *Semantic and structural medatada (e.g. keywords, HTTP variables/headers)*
  - *Domain ontology (e.g. page categories)*
- Structure Data
  - *Content organization within the site*
  - *Hyperlink structure*
    - → *links connecting various resources as a graph*
    - → *intra-page links forming a structure of information within one resource*
- User Data
  - *User profile information*
    - → *usually from registration forms etc.*
    - → *demographics, user ratings, historic data (e.g. purchases), explicit/implicit user interests.*

Overview

- Web Usage Mining
- Collecting and Preprocessing
- Pattern Discovery
- Web Analytics

Data Collecting

- Data Collecting
  - *Explicitly*
    - → *The simplest method*
      - → *Usually based on filling forms, questionnaires, providing ratings.*
    - → *Potentially high quality*
    - → *Issues*
      - → *Users dislike spending time submitting any data.*
      - → *Privacy concerns with providing any personal data.*
  - *Implicitly*
    - → *Non-invasive way*
      - → *Does not require an intervention of any user.*
    - → *Inferring information from user interactions.*
      - → *Many channels: clicks, gestures, posture, eye tracking, language and choice of words, …*
    - → *Issues*
      - → *Difficulties with interpretations of available interactions.*
      - → *Privacy issues with monitoring of users.*

Implicit Data Collecting

- Web/Search logs
  - *Original source of usage data*
  - *Limited information*
- Proxy servers
  - *Complex data from multiple sources, web sites*
  - *Requires using the proxy server*
- TCP/IP packet sniffers
  - *Too limited application*
- Browser/Desktop agents
  - *For advanced analysis of multiple implicit data channels*
    - → *includes specific hardware e.g. eye-trackers*
  - *Requires installation and usage of the agent*
- Client-side JavaScript trackers
  - *Most popular technique*
  - *Complex approaches allowing collecting many relevant data without any requirements on installation of additional SW.*

Client-side JavaScript trackers

- Collected data
  - *Domain name*
  - *Random user identifier*
    - → *Usually created at the time of the first visit*
    - → *Often includes timestamp*
  - *Refferal*
    - → *Information about the first time visit relation*
    - → *domain name, search vs campaign etc.*
    - → *search term*
  - *Referrer*
    - → *Information about the current relation*
  - *User + User-Agent data*
  - *Custom variables*
- Cookies
  - *Using cookies (first/third-party) to store data across many pageviews*
  - *Usually combinations of many cookies*
    - → *session cookie*
    - → *request rate cookies*
    - → *"permanent" user id cookie*
- Data transfer
  - *usually as a part of the query string*
  - *returning "invisible" image deals with JS blocking etc. (1x1 transparent gif), no cache, random identifier*

Client-side JavaScript trackers (cont.)

- Piwik.org
  - *Leading open source web analytics platform*

```
<script type="text/javascript">
  var _paq = _paq || [];
  _paq.push(['trackPageView']);
  _paq.push(['enableLinkTracking']);
  (function() {
    var u="//{$PIWIK_URL}/";
    _paq.push(['setTrackerUrl', u+'piwik.php']);
    _paq.push(['setSiteId', {$IDSITE}]);
    var d=document, g=d.createElement('script'), s=d.getElementsByTagName('script')[0];
    g.type='text/javascript'; g.async=true; g.defer=true; g.src=u+'piwik.js'; s.parentNode.ins
  })();
</script>
```

```
http://piwik-server/piwik.php?_cvar=
{"1":["OS","iphone 5.0"],"2":["Piwik Mobile Version","1.6.2"],
"3":["Locale","en::en"],"4":["Num Accounts","2"]}
&action_name=View settings
&url=http://mobileapp.piwik.org/window/settings
&idsite=8876&rand=351459&h=18&m=13&s=3 &rec=1&apiv=1
&cookie=1&urlref=http://iphone.mobileapp.piwik.org
&_id=af344a398df83874 &_idvc=19&res=320×480&
```

Client-side CSS tracking

- Two features of CSS
  - *the ability to inject content into HTML elements*
  - *the ability to change the style after a user performs an action.*

```
// general idea
#link:active::after {
    content: url("https://example.com/track?action=link_clicked");
}

// Browser detection
@supports (-webkit-appearance:none) and (not (-ms-ime-align:auto)){
    #chrome_detect::after {
        content: url("https://example.com/track?action=browser_chrome");
    }
}

// OS detection
@font-face {
    font-family: Font1;
    src: url("https://evil.com/track?action=font1");
}

#font_detection {
    font-family: Calibri, Font1;
}
```

- https://github.com/jbtronics/CrookedStyleSheets

Preprocessing

- Data cleaning
  - *removing of irrelevant items, log entries produced by spiders and crawlers or error log entries.*
    - → *references to image, css, multimedia or script files*
    - → *specific user-agents/IP addresses including lists of well-known bots*
      - → *heuristic methods to identify bots not well identified*
    - → *entries with status codes not conforming to 2XX*
- Pageview identification
  - *heavily dependent on the way of collecting data and domain specific definition*
    - → *access log entries vs JS trackers*
    - → *viewing a specific page vs product view or purchase*

Preprocessing (cont.)

- User identification
  - *assigning unique user identifier to all entries coming from one user*
  - *Approaches*
    - → *using unique client-side cookies*
    - → *IP address - generally not sufficient (e.g. proxy, NAT, ...)*
    - → *IP address + User-Agent*
    - → *fingerprints*
      - → *HTTP Headers, Plugins, Fonts, Canvas, ...*
    - → *application level identifiers*
    - → *combinations*

```
for (plugin of navigator.plugins) { console.log(plugin.name); }
console.log(navigator.userAgent);
console.log(screen.width + "x" +screen.height)
```

```
Widevine Content Decryption Module
Chrome PDF Viewer
Native Client

Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_4) ...

1680x1050
```


Preprocessing (cont.)

- Session identification
  - *grouping of log entries to sequences related to one user visit*
  - *using time-based or navigation based heuristics*
    - → *Total duration of one session is no longer than a threshold e.g. 30 minutes*
    - → *Total time spent on the page cannot be longer then a threshold e.g. 10 minutes*
    - → *The referrer of the currently visited page should be already part of the session, otherwise start a new session.*
- Path completion
  - *automatic detection of missing entry*
  - *can be caused by caching, proxy servers or any corrupted communication*
    - → *Site structure can help to fill missing entries*
    - → *Many candidate solutions can be available*
      - → *Using the fewest number of "back" references heuristic*
      - → *Identify patterns from existing user patterns*

Overview

- Web Usage Mining
- Collecting and Preprocessing
- Pattern Discovery
- Web Analytics

Data Modeling

- The results of preprocessing
  - *Pageviews*
    - → $P = \{p_1, p_2, \ldots, p_n\}$
  - *Weights:*
    - → *binary*
      - → *represents existence (1) or non-existence (0)*
    - → *function of the duration*
      - → *reflects the time spent on the page (not available for the last pageview - mean value)*
    - → *order of the pageview*
      - → *higher is better*
    - → *combinations or heuristics*
      - → e.g. $(\ln(o) + 1) \times t$
  - *Transactions*
    - → $T = \{t_1, t_2, \ldots, t_m\}$
      - → *where each transaction is a set of pageviews*
      - → $t = \langle (p_1, w(p_1)), (p_2, w(p_2)), \ldots \rangle$

Data Modeling (cont.)

- User-pageview matrix or transaction matrix (UPM)
  - $t = (w_{p_1}, w_{p_2}, \ldots, w_{p_n})$
  - *Simple representation*
  - *For situations when the order of pageviews is not relevant*
  - *Columns usually represent pageviews - all unique page identifiers*
    - → *Can be extended about events within one page*
    - → *Issues with dimensionality*

	A.html	B.html	C.html	D.html	E.html	F.html
t_1	20	9	0	0	0	168
t_2	0	0	32	4	0	0
t_3	21	0	0	46	114	0
t_4	0	0	21	11	0	0

Semantic Information Integration

- Description of each pageview
  - *Issues with granularity*
    - → *URI identifier does not provide any information and is too "fine grained"*
  - *Features*
    - → *Semantic descriptions*
      - → *extracted keywords/data from the page*
      - → *taxonomies*
      - → *e.g. product price, category, ...*
    - → *Classifications*
      - → *e.g. product detail, navigation, general, ...*
- Each feature is assigned during the data collecting phase or linked from an internal knowledge-base

Semantic Information Integration (cont.)

- Pageview-feature matrix (PFM)
  - $p = (fw(f_1), fw(f_2), \ldots)$
  - $fw(f_i)$ *is the weight of the feature in the pageview*

	f_1	f_2	f_3	f_4	f_5	f_6
A.html	1	1	0	0	0	1
B.html	0	0	1	1	0	0
C.html	1	0	0	1	1	0
D.html	0	0	1	1	0	0

- Content-enhanced transaction matrix or Transaction-feature matrix (TFM)
  - $TFM = UPM \times PFM$

	f_1	f_2	f_3	f_4	f_5	f_6
t_1	20	12	12	5	0	0
t_2	0	0	32	4	0	0
...	...	...	...	...	...	...

Data Modeling - Example

- User transactions/visits/sessions

transaction	order	URL	Duration
1	1	Norway.html	60
1	2	AlpTrip.html	120
1	3	Ski.html	240
2	1	Belgium.html	30
2	2	Norway.html	2
2	3	Belgium.html	240

- User-pageview matrix (UPM)

	AlpTrip.html	Belgium.html	Ski.html	Norway.html
t_1	120	0	240	60
t_2	0	270	0	2

Data Modeling - Example (cont.)

- Pageview-feature matrix (PFM)

	Adventure	Leisure	Europe	USA	Norway	Alps	Belgium
AlpTrip.html	1	0	1	0	0	1	0
Belgium.html	0	1	1	0	0	0	1
Ski.html	1	0	1	0	0	1	0
Norway.html	0	1	1	0	1	0	0

- Content-Enhanced transaction matrix or Transaction-feature matrix (TFM)

	Adventure	Leisure	Europe	USA	Norway	Alps	Belgium
t_1	360	60	420	0	60	360	0
t_2	0	281	281	0	2	0	279

Pattern Discovery - Clustering

- Views
  - *User clusters*
  - *Page clusters*
- Algorithms
  - *standard clustering algorithm such as k-means*
    - → *similarities in clusters are maximized and similarities between clusters are minimized*
    - → *using cosine similarity etc.*
- Clustering of users
  - *the most commonly used task*
  - *the goal is to find clusters of users that exhibiting similar browsing patterns*
  - *application in market segmentations, personalizations, user communities*

Clustering Example

- Clustering using using pageviews

	A.html	B.html	C.html	D.html	E.html	F.html
$user_1$	0	0	1	1	0	0
$user_4$	0	0	1	1	0	0
$user_7$	0	0	1	1	0	0
—	——	——	——	——	——	——
$user_0$	1	1	0	0	0	1
$user_3$	1	1	0	0	0	1
$user_6$	1	1	0	0	0	1
$user_9$	0	1	1	0	0	1
—	——	——	——	——	——	——
$user_2$	1	0	0	1	1	0
$user_5$	1	0	0	1	1	0
$user_8$	1	0	0	1	1	0

Clustering Example (cont.)

- Clustering using semantic features
  - *semantic description of each page*
  - *page classifications*

Cluster	#Transactions	Descriptions
1	788	Hiking
2	1398	Bulgaria, Montenegro, Corsica, Last Minute, Search
3	779	Mountaineering, Climbing school, Alpine hiking, Rafts
4	2084	Package holiday, Tour details
5	596	Expeditions, Exotic holidays
…	…	…

Pattern Discovery - Association Analysis

- Association Analysis
  - *Can find groups of items or pages that are commonly accessed (purchased) together.*
  - *Typically uses well known Apriori*
  - *Application in web structure optimizations or recommendations*
- Example
  - *Association rule*
    - → /special-offers/ & /products/software/ → /shopping-cart/
  - *"promotional campaign on software products is positively affecting online sales"*

Association Rule Mining

- Mining of association rules is a fundamental data mining task
  - *Its objective is to find all co-occurrence relationships, called associations*
  - *First introduced in 1993 by Agrawal*
    - → *Apriori algorithm*
- Well-known application
  - *Market basket data analysis*
    - → {Cheese} → {Beer}
      - → *support = 10%, confidence = 80%*
    - → *10% customers buy Cheese and Beer together*
    - → *those who buy Cheese also buy Beer 80% of the time*

Association Rules

- Association Rule
  - *An implication expression X → Y*
    - → *X is an antecedent*
    - → *Y is a consequent*
- Metrics
  - *Support*
    - → *Fraction of transactions that contain both X and Y*
  - *Confidence*
    - → *Measures how often items in Y appears in transactions that contain X*
- Example
  - *{Milk, Diaper} → {Beer}*
    - → $support = \frac{|Milk, Diaper, Beer|}{|T|} = 2/5 = 0.4$
    - → $confidence = \frac{|Milk, Diaper, Beer|}{|Milk, Diaper|} = 2/3 = 0.67$

Transaction	Items
1	{Bread, Milk}
2	{Bread, Diaper, Beer, Eggs}
3	{Milk, Diaper, Beer, Coke}
4	{Bread, Milk, Diaper, Beer}
5	{Bread, Milk, Diaper, Coke}

Apriori Algorithm

- Apriori Algorithm
  - *Introduced in 1993 by Agrawal*
  - *Parameters*
    - → *Minsup - minimum support of rules*
    - → *Minconf - minimum confidence of rules*
  - *Two steps*
    - → *Generate all frequent itemsets*
      - → *A frequent itemset is an itemset that has transaction support above minsup*
    - → *Generate all confident association rules from the frequent itemsets*
      - → *A confident association rule is a rule with confidence above minconf*

Apriori Algorithm (cont.)

- Step 1

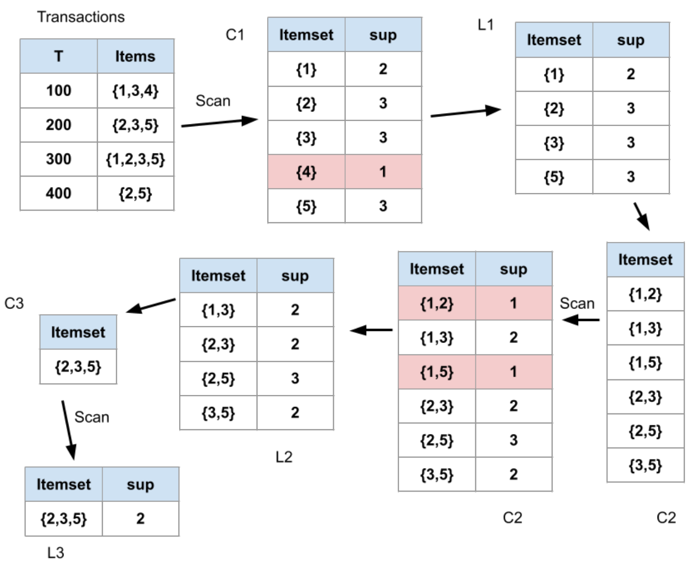
```
Ck =  Candidate itemset of size k
Lk =  frequent itemset of size k

L1 = {frequent items}
for (k=1; Lk.size() != 0; k++):
    Ck+1 = candidates generated from Lk
    for each transaction t in database do:
        increment the count  of all candidates in Ck+1
        that are contained in t
    Lk+1 = candidates in Ck+1 with min_support
return all Lk
```

- Step 2

```
For each frequent itemset l, generate all nonempty subsets of l.
For every nonempty subset s of l, output the rule s -> (l-s) if
support(l) / support(s) >= minconf
```


Apriori Algorithm Example

- Min support = 50%



https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput499/slides/Lect10/sld054.htm

Apriori Algorithm Example 2

- Apriori is not able to work with numeric values
  - *There is a requirement to properly preprocess the data*
- Preprocessing
  - *Discretization to binary values*
    - *→ loss of information*
  - *Binarization*
    - *→ increased number of dimensions*
- Example
  - *Content-Enhanced transaction matrix or Transaction-feature matrix (TFM)*

	Adventure	Leisure	Europe	USA	Norway	Alps	Belgium
t_1	360	60	420	0	60	360	0
t_2	0	281	281	0	2	0	279

	Adventure	Leisure	Europe	USA	Norway	Alps	Belgium
t_1	1	1	1	0	1	1	0
t_2	0	1	1	0	1	0	1

	AdventureH	AdventureM	AdventureL(<50)	LeisureH(>200)	LeisureM	LeisureL(<50)	…
t_1	1	0	0	0	1	0	…
t_2	0	0	1	1	0	0	…

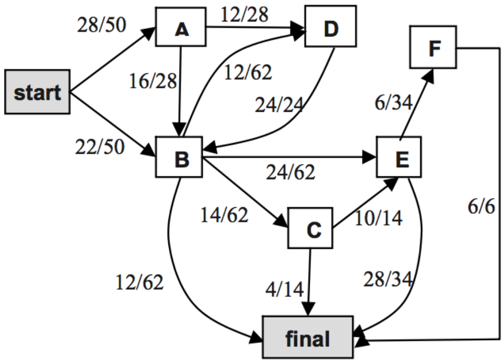
Apriori Algorithm Example 2 (cont.)

- Example of task
  - *Format of rules*
    - *→ Antecedent*
      - *→ temporal information, referral information*
    - *→ Consequent*
      - *→ Purchase flag*
  - *Settings*
    - *→ min support = 2% (e.g. 50 transactions)*
    - *→ min confidence = 70%*
  - *Results*
    - *→ {Referral=GoogleSearch, Hour=Morning} → {Purchase=True}*
    - *→ support= 3%, confidence = 75%*

Sequential and Navigational Patterns

- Sequential and Navigational Patterns
  – *Similar to association rules*
    → take into account temporal information = order of pageviews
- Markov models can be used as the underlying concept for the sequential modeling.

Transaction	Frequency
A, B, E	10
B, D, B, C	4
B, C, E	10
A, B, E, F	6
A, D, B	12
B, D, B, E	8



Liu, B. "Web Data Mining", Springer-Verlag Berlin Heidelberg, 2011. ISBN 978-3-642-19459-7.

Overview

- Web Usage Mining
- Collecting and Preprocessing
- Pattern Discovery
- Web Analytics

Web Analytics

- Web Analytics
  - *General Access Pattern Tracking*
    - → *Analysis of user patterns and general trends to get overview about the overall behavior.*
  - *Collecting data, analysis and reporting*
- Two main categories
  - *Traffic analysis*
    - → *pageviews*
    - → *sessions*
    - → *visitors*
    - → *time on page*
    - → *bounce rate*
  - *E-commerce analysis*
    - → *Conversion, Conversion rate*
    - → *Revenue*
    - → *Campaigns*
    - → *Impressions*
    - → *CTR - click through rate, CPC - cost per click*

Web Analytics

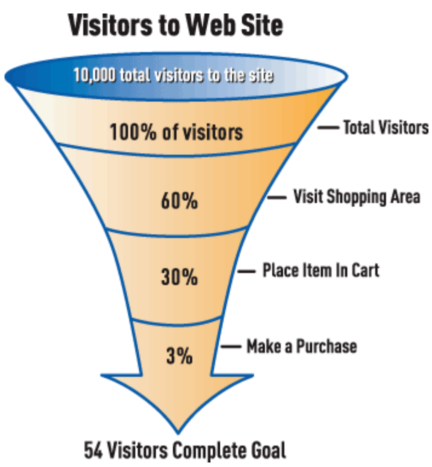
- Main overview
  - *Number of pageviews*
  - *Number of sessions*
  - *Number of visitors/unique*
    - → *returning visitors*
- Time on Page and Time on Site
  - *Time on the last page issue*
    - → *the average value from all previous pages*
    - → *specific hacks - JavaScript* onbeforeunload *event*
  - *Tabbed browsing*
    - → *identification and represent as separate sessions*
    - → *normalize as one session*
    - → *specific hacks - JavaScript* visibilitychange *event*
- Bounce rate
  - *"I came, I puked, I left."*
  - *the percentage of sessions on your website with only one page view*
  - *Specific use cases*
    - → *Blogs*

Web Analytics (cont.)

- Conversion rates
  - *Outcomes divided by Unique Visitors (or Visits)*
- Conversions - desired outcome
  - *Macro conversion*
    - → *Limited amount*
      - → *Submitted order*
      - → *Registration to a newsletter*
  - *Micro conversions*
    - → *Higher amount*
    - → *More complex solutions and issues with the interpretation*
      - → *product ratings*
      - → *video consumption*
      - → *shopping basket operations*
      - → *dynamic content interactions*
      - → *...*

Web Analytics (cont.)

- Conversion funnel
  - *describe the journey a consumer takes through an e-commerce website and finally converting to a sale*



King, B.A. "Website Optimization: Speed, Search Engine Conversion Rate Secrets", O'Reilly Media, 2008. ISBN 978-0596515089.

Web Analytics (cont.)

- Campaigns
  - *Sources*
    - → *Medium*
      - → *organic, cpc, banner, email, referral, none*
    - → *Source*
      - → *google, seznam, facebook, direct*
    - → *Search terms*
    - → *Example*
      - → *http://www.example.com/?utm_medium=cpc&utm_source=google*
- Landing and exit pages
  - *First and last pages of user sessions.*