

# Analýza chování návštěvníků na webu (Web Usage Mining)

4IZ470, LS 2020, cvičení

# Případová studie

- Analýza webového logu cestovní kanceláře
- Zcela reálná data
  - Za období o něco větší než 1 měsíc
- Více než 10 let stará
  - Technologický způsob strukturování a pojmenovávání webů se změnil
  - Způsob interpretování výsledků v kontextu online marketingu se částečně změnil (viz přednáška Ing. Tichého)
  - *Základní principy navigace návštěvníka na webu (tj. i složení návštěv ze zobrazení stránek), ani dataminingové postupy použitelné pro analýzu „clickstreamů“, se prakticky nezměnily*

# Clicks.csv

- \* **LocalID** - interní identifikátor události
- \* **PageID** - identifikátor zobrazené stránky
- \* **VisitID** - identifikátor session
- \* **PageName** - relativní URI navštívené stránky
- \* **CatName, CatID** - typ stránky (Navigace), obecnější granularita
- \* **ExtCat, ExtCatID** - typ stránky (Obsah), konkrétnější granularita
- \* **TopicName, TopicID** - téma stránky (VHT = vysokohorská turistika)
- \* **TimeOnPage** - čas na stránce v sekundách ( $t$ ); údaj je zaokrouhlen na půlminutové bloky, na poslední stránce session se předpokládá 30s
- \* **PageScore** - váha stránky odvozená od času  $t$  stráveného na stránce a pořadí  $o$  stránky v clickstreamu, podle heuristiky
$$\text{PageScore} = (\ln(o)+1)*t$$
- \* **SequenceNumber** - pořadí stránky v clickstreamu ( $o$ )

# Kategorizace typů stránek

(stránky s podtrženým typem: mají i téma)

- Zájezd
- Info
  - Homepage
  - Kdo jsme
  - Slevy
  - Pojištění
  - Zaslát katalog
  - Posílat novinky
- Hledání
  - Katalog
  - Dle KW (... klíčových slov)
  - Dle země
  - Rozšířené
- Přihláška
  - Přihláška

# Visitors.csv

- \* **VisitID** - identifikátor session
- \* **Referrer** - anonymizované označení odkazující domény
- \* **Den** - den započetí návštěvy
- \* **Hodina** - hodina započetí návštěvy
- \* **Delka\_sekundy** - délka návštěvy v sekundách  
(součet hodnot *TimeOnPage* v řádcích se stejným *VisitID* v *clicks.csv*)
- \* **Delka\_pocetstranek** - počet navštívených stránek  
během návštěvy (počet řádků se stejným *VisitID* v *clicks.csv*)

# Search\_engine\_map.csv

- \* **Referrer** - anonymizované označení odkazující domény
- \* **Typ\_Odkazovace** - typ odkazující domény

# Clickstreams.csv

- Předzpracovaná data
  - Vybrané jen návštěvy s >2 pageviews  
(asi 4,5 tisíce z původních více než 15 tisíc)
- Jeden řádek na návštěvu
- Agregovaná data
  - **NejTema** je TopicName s nejvyšším součtem PageScore stránek, které toto téma mají

# Základní statistiky



# Kvíz 1

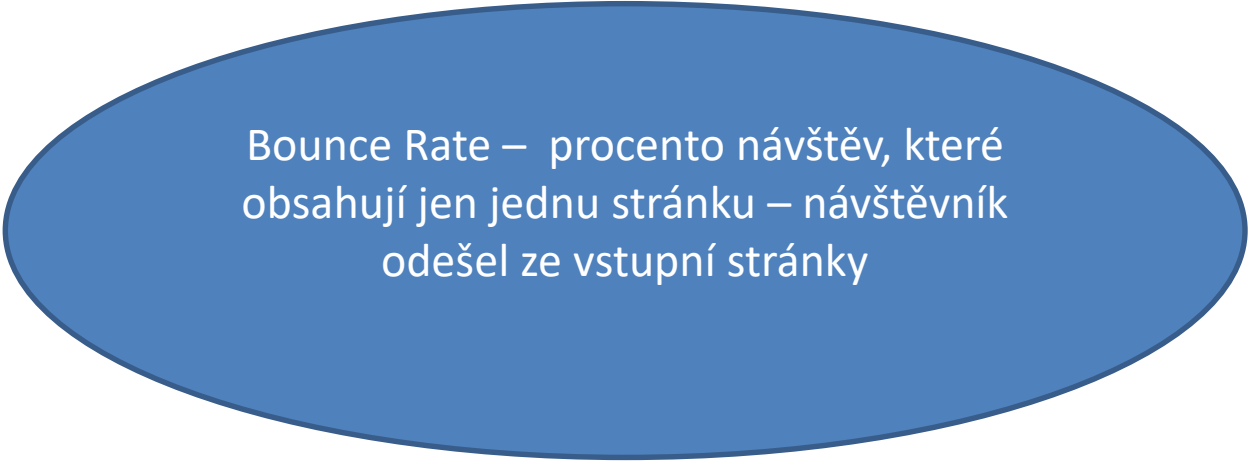
- Považujte navštívení stránky “n\_katalog.asp” za konverzi.
  - Jaký je konverzní poměr?
  - Lze spočítat dvěma způsoby ze dvou různých vstupních souborů
- *Řešení je na další straně*

# Kvíz 1

- Považujte navštívení stránky “n\_katalog.asp” za konverzi.
  - Jaký je konverzní poměr?
- Řešení:
  - V souboru *clicks.csv* spočítejte relace (unikátní *VisitID*) ve kterých se vyskytuje stránka *n\_katalog.asp* a vydělte celkovým počtem relací (počet unikátních *VisitID*)
  - V souboru *Clickstreams.csv* vydělte počet řádků s nenulovou hodnotou v poli *KonverzeStrankaScore* celkovým počtem řádků
    - Vyjde jiný výsledek – kvůli nezahrnutí „krátkých návštěv“

# Kvíz 2

- Jak zjistit následující informace?
  - Jaká je „bounce rate“?
  - Jaké jsou nejčastější “bounce page”?
  - Odkud přicházejí návštěvníci, kteří se “odrazí”?
- *Řešení je na další straně*



Bounce Rate – procento návštěv, které obsahují jen jednu stránku – návštěvník odešel ze vstupní stránky

# Kvíz 2

- Jak zjistit následující informace?
  - Jaká je „bounce rate“?
  - Jaké jsou nejčastější “bounce page”?
  - Odkud přicházejí návštěvníci, kteří se “odrazí”?
- Řešení:
  - Bounce rate získáme jako podíl řádků ve *visitors.csv*, které mají ve sloupci *Delka\_pocetstranek* hodnotu 1, vůči všem řádkům
  - Nejčastější „bounce page“ pak získáme přes *VisitID* ze souboru *clicks.csv*, jako hodnotu *PageName*
  - Konkrétní zdroje návštěvníků, kteří se „odrazí“, zjistíme už z *visitors.csv*; přes *search\_engine\_map.csv* je pak můžeme rozdělit do typů, a sledovat i opačně, které typy zdrojů (referrerů) vedou na vysokou „bounce rate“ a které na nízkou

# Individuální domácí úkol

- Termín 27. 3., do odevzdávárny („WUM - Individuální úkol“), hodnocení max. 2 body
- Zpracovat nad daty tyto úkoly (zhruba odpovídající předešlým kvízům)
  - Konverzní poměr, bounce rate
  - Četnosti stránek v roli bounce page
  - Korelaci typu referreru s výskytem konverze a bounce (jako podmíněné četnosti, vypočtené oběma směry)
- Lze buď v Excelu (některé části mohou být pracné), nebo v programovacím jazyce
  - V obou případech je nutno dodat jak *výpočet* (tabulku se vzorci / kód programu), tak i zřetelně vyznačené dosažené *výsledky*