Detecting Users from Website Sessions: A Simulation Study

Corné de Ruijt
Sandjai Bhulai

DATA ANALYTICS 2020

Vrije Universiteit Amsterdam
Contact: c.a.m.de.ruijt@vu.nl

Introduction

- PhD student at the Vrije universiteit Amsterdam
- Research is mainly about job recommender systems and recruitment websites

Introduction

- In this study we wish to address the problem of recognizing unique users from session data.
- Recognizing users from session data is typically an interplay between cookies and log-in.
- 20% of the internet users may delete their cookie at least once a week [1].
- We study this problem by: 1) proposing a click simulation model, 2) studying how effective (H)DBSCAN*-type algorithms are to this problem.

[1] Dasgupta et al. Overcoming browser cookie churn with clustering. Proceedings of the fifth ACM international conference on Websearch and data mining, pages 83–92. ACM, 2012

Outline

- ~~Introduction~~
- Literature
- Click simulation model
- MS-(H)DBSCAN* algorithms for session clustering
- Results and conclusion

Literature

- Specific type of entity/identity resolution problem [2]
- ICDM[3] and CIKM [4] cross-device matching competitions
- Ambiguity in the literature:
  - Problem is addressed under different names (user stitching, visitor stitching, automatic identity linkage)
  - Problem is addresses from single and multiple website perspective

[2] Di et al. ”node2bits: Compact Time-and Attribute-aware Node Representations”, 2019
[3] ICDM 2015: Drawbridge Cross-Device Connections, 2015. (https://www.kaggle.com/c/icdm-2015-drawbridge-cross-device-connections)
[4] CIKM Cup 2016 Track 1: Cross-Device Entity Linking Challenge, 2016. (https://competitions.codalab.org/competitions/11171)

Research objective

We assume:

- Single website (a search engine)
- Homogeneous queries, but with different utility functions

Interested in:

- What simulation model should we use to obtain realistic simulated datasets, without making the simulation model overly complex?
- How effective are (H)DBSCAN*-type algorithms on clustering sessions compared to using cookies, in terms of multiple website statistics?
- Sensitivity of (H)DBSCAN*-type algorithms on the underlying dataset.

Simulation

Why simulation?

- Have an actual ground truth (problem of unary classification).
- Sensitivity of clustering algorithm under different simulation settings.
- Quite some literature on models explaining website behavior on search engines [5].
- Most public data sets originate from large search engines/large advertisers. How generalizable are these for other websites?

[5] Chuklin et al. "Click models for web search." Synthesis lectures on information concepts, retrieval, and services 7.3 (2015): 1-115.

Outline

- ~~Introduction~~
- ~~Literature~~
- Click simulation model
- MS-(H)DBSCAN* algorithms for session clustering
- Results and conclusion

Simulation

- **Users' item preferences**: Fleder-Hosanagar model[6].
- **Interaction with search engine**: Simplified Dynamic Bayesian Network click model[7].
- **Cookie churn**: 1) modeling cookie lifetime, 2) allowing users to use multiple devices (with different cookie-IDs).

[6] Fleder et al. "Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. Management science", 55(5):697–712, 2009

[7] Chapelle et al. "A dynamic bayesian network click model for web search ranking." Proceedings of the 18th international conference on World wide web. 2009.

Simulation summary

Draw items $\mathcal{V}$ and users $\mathcal{U}$ and compute each user's attraction and satisfaction over all items using Fleder's model;

for $u \in \mathcal{U}$ **do**

  Draw initial device, cookie lifetime;

  Set t and t_{start} to arrival time of u, draw user lifetime T_u^{user};

  $i \leftarrow 1$;

  while $t \leq t_{start} + T_u^{user}$ **do**

    Simulate clicks for query session i according to SDBN, items are presented according to a popularity recommender;

    Draw $T_{u,i}^{\text{abs}}$ and update t;

    Update device and cookies;

    $i \leftarrow i + 1$

  end

end

Simulation, output

User (u) 1	Query-session (i)	Cookie (o)	Device (d)	List position	Item (v)	t	Click/ skip
1	1	1	1	1	12	12	1
1	1	1	1	2	13	12	0
1	1	1	1	3	44	12	0
...	...	...	...	...	...	...	...
1	1	1	1	10	84	12	0
2	1	1	2	1	5	12	0
...	...	...	...	...	...	...	...
312	32	2	2	1	871	16	0

Outline

- ~~Introduction~~
- ~~Literature~~
- ~~Click simulation model~~
- ~~MS-(H)DBSCAN* algorithms for session clustering~~
- Results and conclusion

Results

Model	Dataset	APE unique users	KL-div. session count	KL-div conversion	**ARI**	New user accuracy
MS-DBSCAN	train	15	**0.55**	0.13	0.0012	**0.56**
MS-DBSCAN$_P$	train	77	0.74	**0.092**	**0.14**	0.5
DBSCAN-RAND	train	**0.011**	1	0.096	0.0002	0.42
MS-HDBSCAN*$^+$	train	10	0.75	0.15	0.00079	0.52
MS-HDBSCAN*$^-$	train	10	0.75	0.15	0.00079	0.52
MS-HDBSCAN*^+_P	train	**0.011**	0.9	0.11	0.092	0.46
MS-HDBSCAN*^-_P	train	**0.011**	0.9	0.11	0.1	0.46
OBS	*train*	*15*	*0.017*	*0.0032*	*0.91*	*0.95*
MS-DBSCAN	valid	60	**0.11**	**0.0026**	**0.0022**	**0.56**
MS-DBSCAN$_P$	valid	**6.8**	1.4	0.13	0.0015	0.4
DBSCAN-RAND	valid	40	0.32	0.015	0.00046	0.5
MS-HDBSCAN*$^+$	valid	53	0.16	0.0042	0.002	0.55
MS-HDBSCAN*$^-$	valid	53	0.16	0.0042	0.002	0.55
MS-HDBSCAN*^+_P	valid	7.2	1.4	0.13	0.0015	0.4
MS-HDBSCAN*^-_P	valid	7.2	1.4	0.13	0.0015	0.4
OBS	*valid*	*51*	*0.1*	*0.0076*	*0.91*	*0.95*

Conclusion

- We presented a click simulation model with cookie censoring and illustrated its usage and advantages on the problem of uncovering users from their web sessions.
- Usage of cookie-IDS as clusters outperforms MS-(H)DBSCAN*-type algorithms on the homogeneous query case;
- Of the MS-(H)DBSCAN*-type algorithms, MS-DBSCAN significantly outperformed other MS-(H)DBSCAN* type algorithms in terms of ARI, including DBSCAN-RAND;
- Increasing ARI or new user accuracy also seems to increase the unique user average percentage error;
- Increasing the signal (e.g.,by increasing the number of clicks) improved the results, but slightly.

Further research

- What happens if we move to a multi-query case?
- Adjust clustering approaches using supervised methods s.t. they are less prone to overfitting.
- Holistic research on cookie (censoring) models.