

Towards microaggregation of log files for Web usage mining in B2C e-commerce

Guillermo Navarro-Arribas, Vicenç Torra

IIIA - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research

`{guille,vtorra}@iia.csic.es`

NAFIPS-2009

Outline

- 1 Introduction
- 2 Microaggregation
- 3 Microaggregation of Web logs
- 4 Conclusion

Web usage mining

- Obtain useful usage information of a web site.
- Main information source: Web Server access log.

```
10.0.0.1 - - [11/Dec/2008:16:01:22 +0100]
"GET /guille/index.html HTTP/1.1" 200 958
"http://hacks-galore.org/guille/others.html"
"Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5;
en-US; rv:1.9.0.4) Gecko/2008102920 Firefox/3.0.4"
```

Web usage mining information

- Statistical analysis mainly based on means and counts.
- Pattern discovery base on identified sessions:
 - Share the same IP
 - Same user agent
 - Maximum timeout between clicks \sim 30 min.

Web logs and privacy

Why privacy?

```
10.0.0.1 - - [19/Dec/2008:12:35:25 +0100] "GET /guille/  
HTTP/1.0" 200 1070 "-" "Emacs-w3m/1.4.304 w3m/0.5.2"
```

```
10.0.0.3 - - [24/Nov/2008:17:27:02 +0100] "GET  
/jao/journal/images/mistyforest.jpg HTTP/1.1" 200 122577  
"http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile\  
&friendid=762XXXX\&MyToken=938a1b89-6012-XXX-bf2e"  
"Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5; en-US; rv:1.9.0.4)  
Gecko/2008102920 Firefox/3.0.4"
```

```
10.0.0.4 - - [24/Nov/2008:17:44:02 +0100] "GET  
/secret/login?name=guille\&password=letmein HTTP/1.1" 200  
122577 "-" "Mozilla/5.0"
```

Web log anonymisation

The good

- Logs can be safely stored for future analysis.
- Log analysis can be outsourced.

The bad

- Anonymisation involves losing information.

Towards Web log anonymisation

Our approach

- Use SDC techniques
 - SDC successfully used in PPDM
 - Keep (most) statistical information
 - Allow pattern discovery based on sessions
- There is always a trade off:

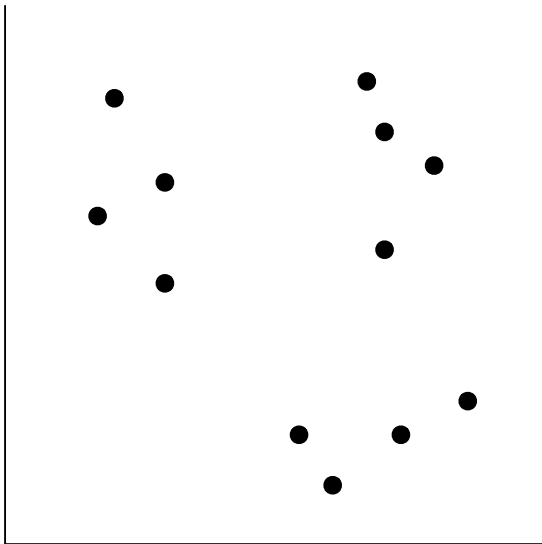
utility vs. privacy

Microaggregation

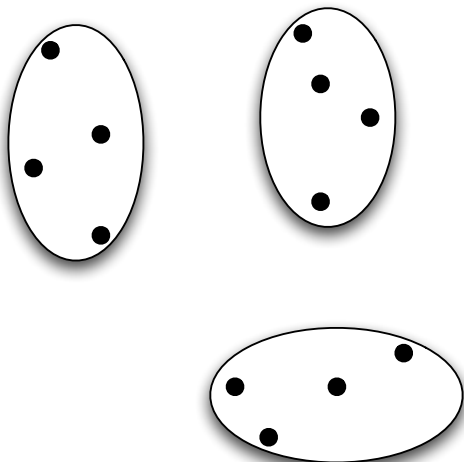
Microaggregation one of the most used SDC techniques

- Clustering with parameter k
 - k minimum number of element per cluster.
- Can provide **k-anonymity**
- Multivariate microaggregation is NP-hard
 - Need for heuristics \Rightarrow MDAV

Microaggregation explained

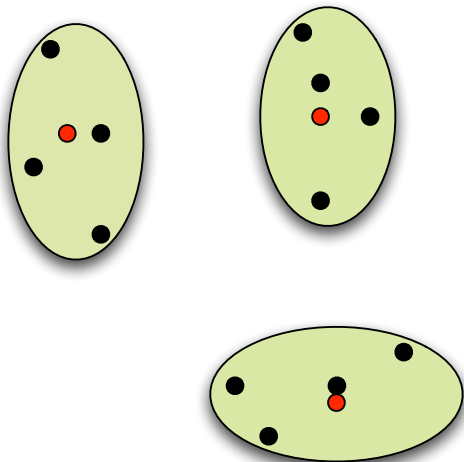


Microaggregation explained



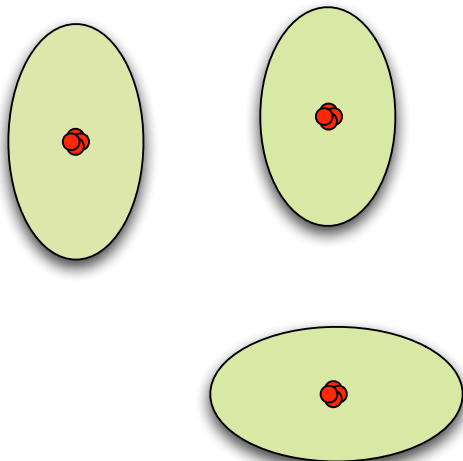
- 1 Make clusters \Rightarrow **distance** function.

Microaggregation explained



- 1 Make clusters \Rightarrow **distance** function.
- 2 Calculate centroid \Rightarrow **aggregator** operator.

Microaggregation explained



- 1 Make clusters \Rightarrow **distance** function.
- 2 Calculate centroid \Rightarrow **aggregator** operator.
- 3 Substitute points by their centroid.

Microaggregation of Web logs

What we do NOT want:

- logs from the same session in the same cluster
 - All belong to the same user
 - Lots of information lost for session (request)

Approach

- Distance function determines the cluster composition
- Provide distance such that clusters contain logs from different sessions.

Web log Microdata

Each web log data is parsed into a set of $20 + 3$ variables:

- Main variables:
ip, dns, authn, username, date, query-method, query-path,
query-querysting, query-protocol, status, size, referer-scheme, referer-host,
referer-port, referer-path, referer-querysting, ua-robot, ua-os, ua-os-version,
ua-browser, ua-browser-version, country
- Session based variables (for more accurate session identification):
basetime, reltime, sessionid
- IP is encrypted (hashed) or discarded.

Web log distance

- Most likely:
 - Same request \Rightarrow different session
 - Same referer \Rightarrow different session

\Rightarrow Use a **weighted mean**

- More weight in:
 - Request
 - Referer
- Distance of a log line as a weighed mean of distances of each field.

Distance example: DNS name

$$d_{dns}(X, Y) = \sum_{i=1}^n \omega_i \alpha_i \quad \text{where } \omega_i = \frac{2^{i-1}}{2^n - 1};$$

$$\alpha_i = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases}$$

$$d = 2/15 \approx 0.133$$

www.iiia.csic.es $\rightarrow x_1 = \text{'www'}, x_2 = \text{'iiia'}, x_3 = \text{'csic'}, x_4 = \text{'es'}$

www.eez.csic.es $\rightarrow x_1 = \text{'www'}, x_2 = \text{'eez'}, x_3 = \text{'csic'}, x_4 = \text{'es'}$

$$d = 7/15 \approx 0.466$$

www.iiia.csic.es $\rightarrow x_1 = \text{'www'}, x_2 = \text{'iiia'}, x_3 = \text{'csic'}, x_4 = \text{'es'}$

www.ugr.es $\rightarrow x_1 = \text{'-'}, x_2 = \text{'www'}, x_3 = \text{'ugr'}, x_4 = \text{'es'}$

Aggregation

Centroid composed by mainly means and majority rules.

Ex. DNS name

$\mathbb{C} = \text{csic.es}$

$\text{www.iiia.csic.es} \rightarrow x_1 = \text{'www'}, x_2 = \text{'iiia'}, x_3 = \text{'csic'}, x_4 = \text{'es'}$

$\text{www.eez.csic.es} \rightarrow x_1 = \text{'www'}, x_2 = \text{'eez'}, x_3 = \text{'csic'}, x_4 = \text{'es'}$

$\mathbb{C} = \text{es}$

$\text{www.iiia.csic.es} \rightarrow x_1 = \text{'www'}, x_2 = \text{'iiia'}, x_3 = \text{'csic'}, x_4 = \text{'es'}$

$\text{www.ugr.es} \rightarrow x_1 = \text{'-'}, x_2 = \text{'www'}, x_3 = \text{'ugr'}, x_4 = \text{'es'}$

Conclusions

- Use of microaggregation to anonymise Web log for Web usage mining.
- Goog results for statistical analysis.
- Not bad for pattern discovery.
- Currently:
 - good general results for small Web sites.
 - not so good for big sites.

- Future:
 - Improve parametrisation (mainly weights), by empirical analysis.

Towards microaggregation of log files for Web usage mining in B2C e-commerce

Guillermo Navarro-Arribas, Vicenç Torra

IIIA - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research

`{guille,vtorra}@iia.csic.es`

NAFIPS-2009