# Ús d'un element segur per protegir perfils d'usuari generats per Motors de Cerca Web

David Pàmies-Estrems  Jordi Castellà-Roca

MISTIC
Universitat Oberta de Catalunya

Treball Final de Màster

UOC
Universitat Oberta de Catalunya

1. Introducció
2. Estat de l'art
3. Requisits
4. Sistema de privadesa
   - Classificador
   - Anonimitzador i creació del perfil
   - Desanonimitzadors
5. Avaluació
6. Conclusions i treball futur

UOC
Universitat Oberta de Catalunya

# Motors de Cerca Web (MCW)

Punt d'entrada a la Web

Guarden registres, *query logs*:

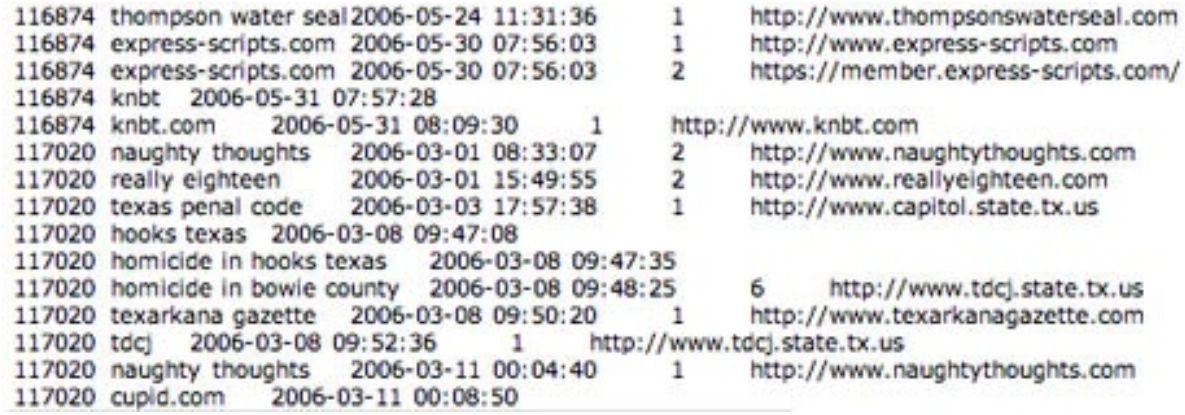
- Personalitzar resultats
- Millorar cerca
- Marketing
- Recerca

UOC
Universitat Oberta de Catalunya

# Privadesa en risc

Registre:

- Usuari
- Consulta
- Moment
- Destí



```
116874 thompson water seal 2006-05-24 11:31:36 1 http://www.thompsonswaterseal.com
116874 express-scripts.com 2006-05-30 07:56:03 1 http://www.express-scripts.com
116874 express-scripts.com 2006-05-30 07:56:03 2 https://member.express-scripts.com/
116874 knbt 2006-05-31 07:57:28
116874 knbt.com 2006-05-31 08:09:30 1 http://www.knbt.com
117020 naughty thoughts 2006-03-01 08:33:07 2 http://www.naughtythoughts.com
117020 really eighteen 2006-03-01 15:49:55 2 http://www.reallyeighteen.com
117020 texas penal code 2006-03-03 17:57:38 1 http://www.capitol.state.tx.us
117020 hooks texas 2006-03-08 09:47:08
117020 homicide in hooks texas 2006-03-08 09:47:35
117020 homicide in bowie county 2006-03-08 09:48:25 6 http://www.tdcj.state.tx.us
117020 texarkana gazette 2006-03-08 09:50:20 1 http://www.texarkanagazette.com
117020 tdcj 2006-03-08 09:52:36 1 http://www.tdcj.state.tx.us
117020 naughty thoughts 2006-03-11 00:04:40 1 http://www.naughtythoughts.com
117020 cupid.com 2006-03-11 00:08:50
```

Identificadors $\rightarrow$ **Persona** $\leftarrow$ Quasi-identificadors

Modificacions no garanteixen protecció

UOC Universitat Oberta de Catalunya

# Estat de l'art

- Batch
  - Privadesa diferencial
  - Heurístiques
  - Micro-agregació
- Temps real
  - Tractament immediat
  - Volums infinits
  - Dades estructurades i numèriques

**Cap proposta per protegir logs en temps real**

Universitat Oberta de Catalunya

# Requisits

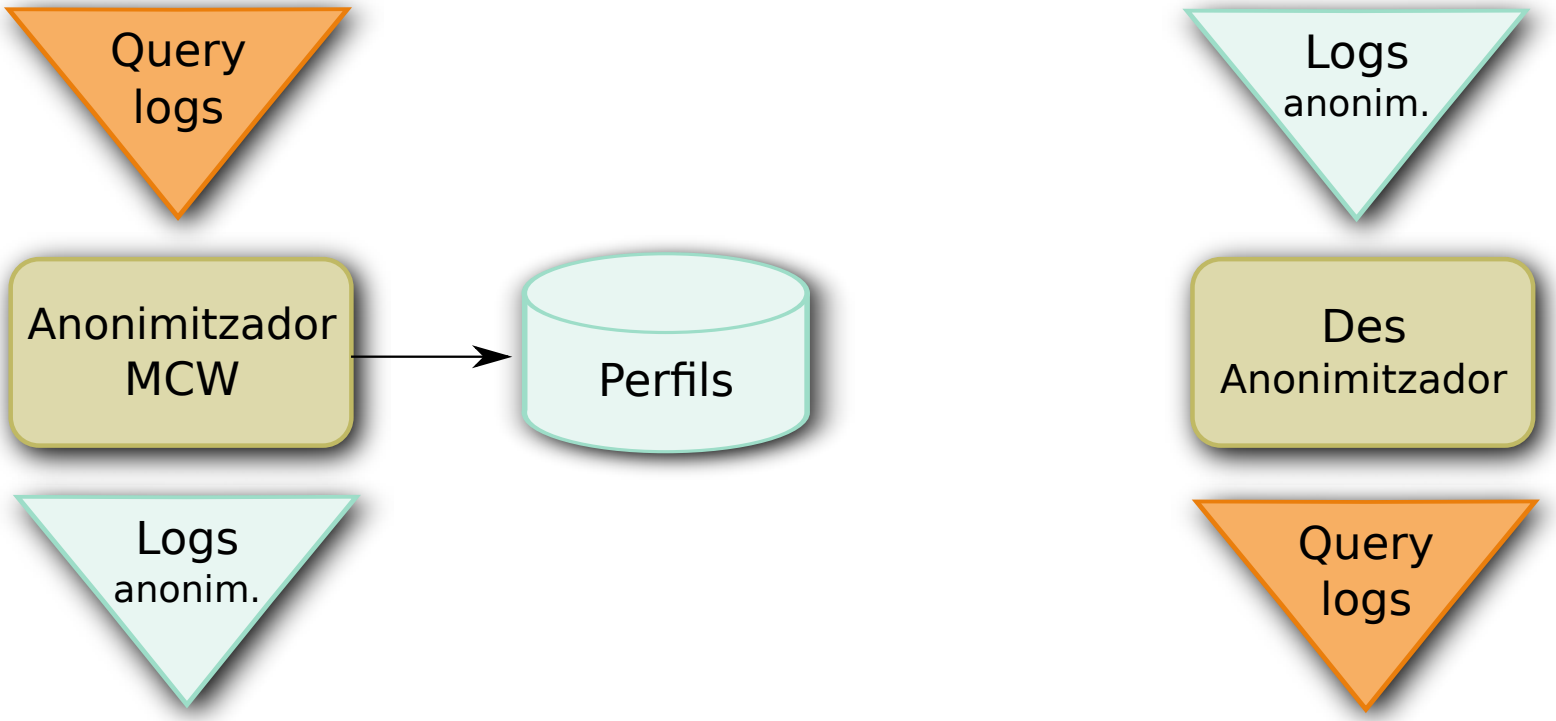
**Monetitzar = Privadesa + Utilitat**

Baix consum de recursos

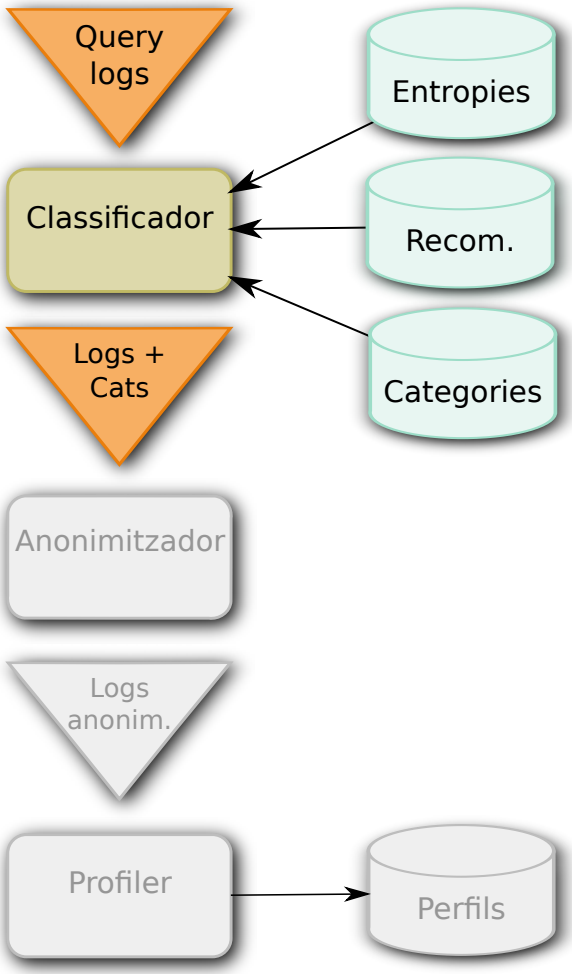
Consultes d'altres usuaris, mateixos interessos

MCW pot guardar logs originals o protegits

# Proposta


Query logs
Anonimitzador MCW
Perfils
Logs anonim.
Logs anonim.
Des Anonimitzador
Query logs

# Classificador




- Llenguatge natural
- Recomanacions
- Entropies
- Categorització

# Classificador

Tasques addicionals:

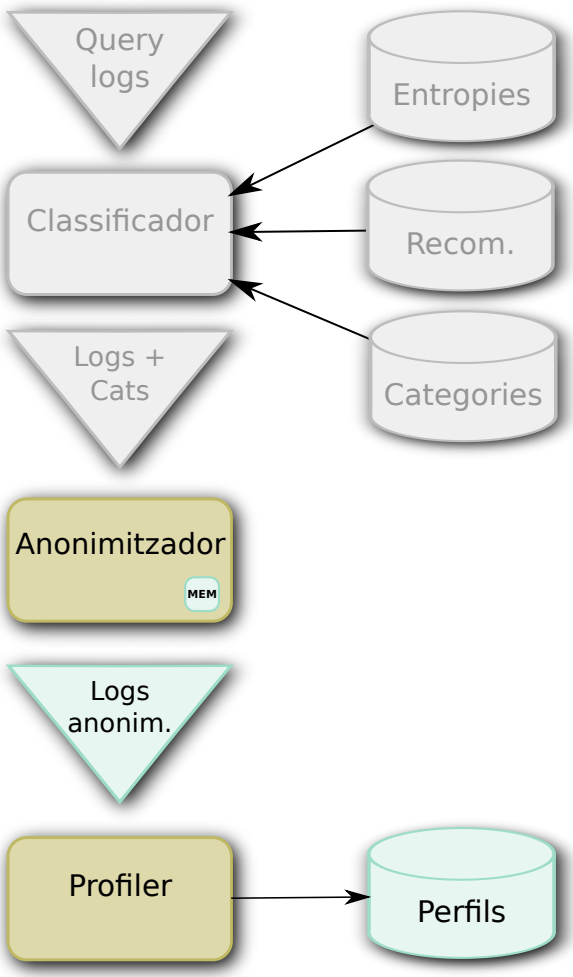
- Recomanacions
- Entropies
- Categories

TOR + Privoxy

Versió ràpida NLTK


UOC
Universitat Oberta de Catalunya

# Anonimitzador i creació del perfil




- Categories de primer nivell
- Separem usuaris i consultes
- Unim aleatòriament

**Restricció: No repetim combinació original**

- Profiler actualitza perfils

# $D(x)$: Desanonimitzadors

Reconstruïr logs originals
Consulta aleatòria dins la categoria

Usuari:

1. Aleatori, amb restricció
2. Més vegades en categoria
3. Perfil més consultes
4. Perfil més consultes i més vegades

UOC
Universitat Oberta de Catalunya

# $D(x)+T$: Desanonimitzadors amb Temps

Efecte camp temps, en temps real?

Camp temps sense utilitzar
Ordre de publicació
$\Downarrow$
Logs ordenats per camp temps
S'han processat tots?

UOC
Universitat Oberta de Catalunya

# Heurístiques

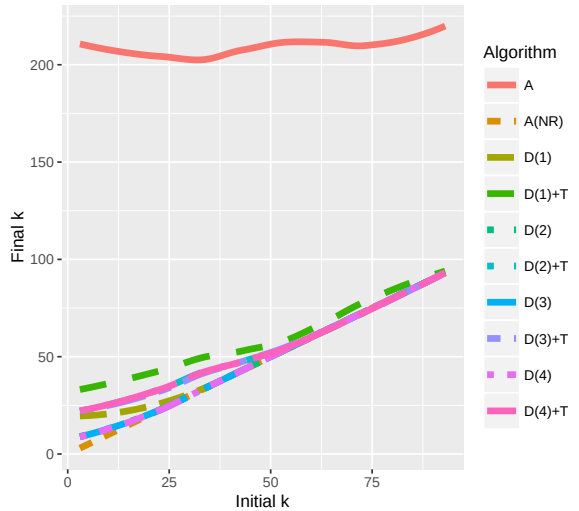
**Sense restricció**

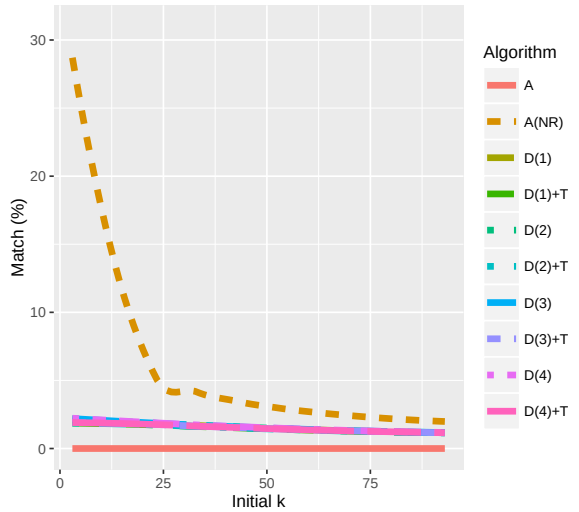
- $k$ constant
- % sense anonimitzar

**Amb restricció**

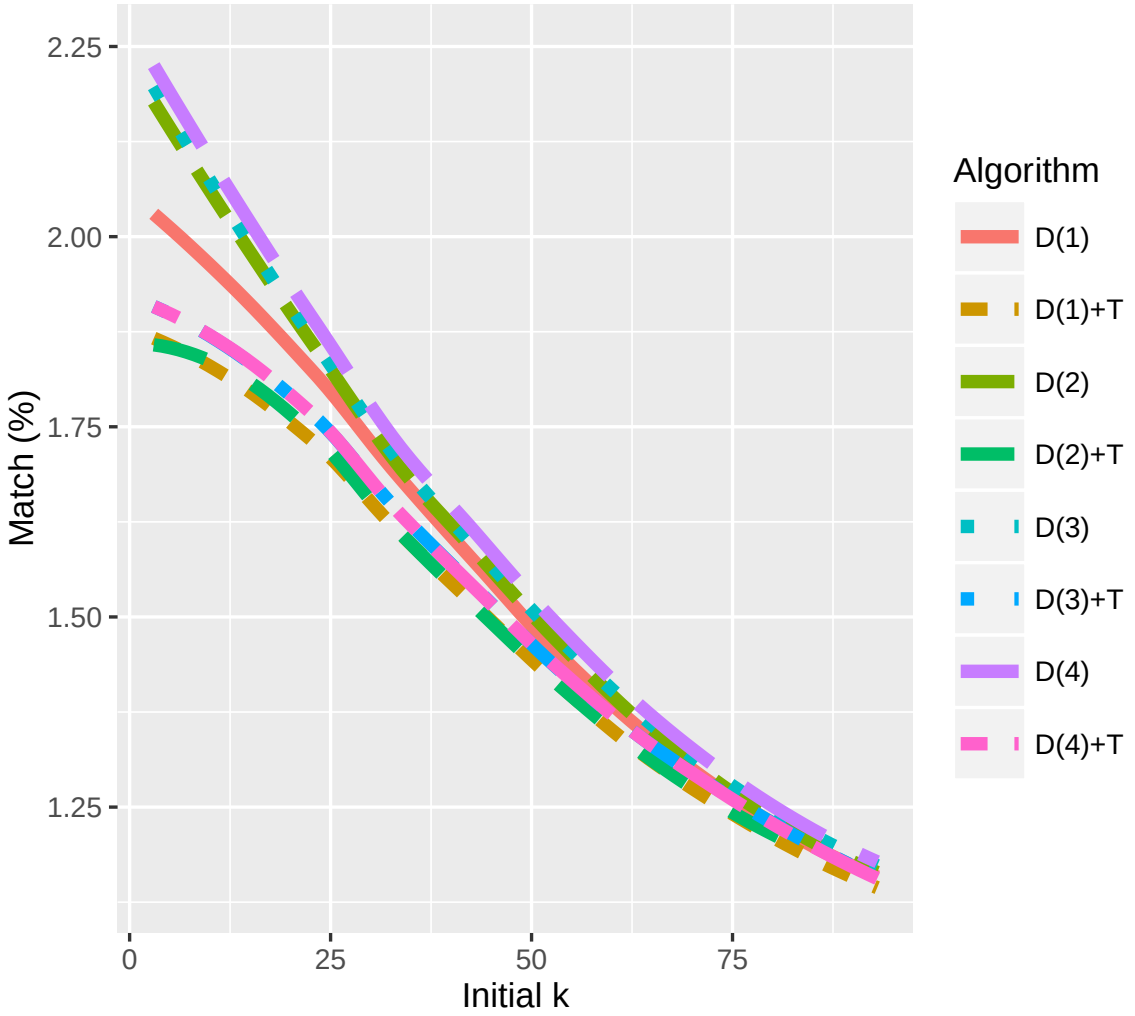
- $k > 200$
- Tots anonimitzats

$\delta \rightarrow$ Consum de memòria



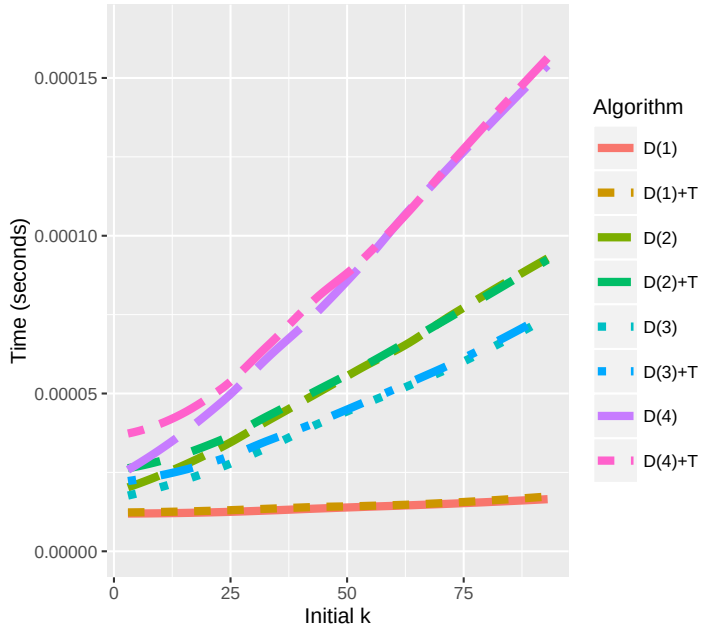
# Privadesa


2.25
2.00
1.75
1.50
1.25
Match (%)
0
25
50
75
Initial k
Algorithm
D(1)
D(1)+T
D(2)
D(2)+T
D(3)
D(3)+T
D(4)
D(4)+T

# Requisits funcionals

| | Temps/log($\mu$s) | Consultes/segon |
|---|---|---|
| Classificador | **1.503** | 665 |
| Anonimitzador | 22 | **45.454** |
| Profiler | 267 | 3.745 |

| | Mem.(MiB) |
|---|---|
| Classificador | 112 |
| Anonimitzador | 10 |
| Profiler | 12 |



# Utilitat

| | | | |
|---|---|---|---|
| Arts | 57% | Recreation | 55% |
| Business | 77% | Reference | 69% |
| Computers | 65% | Regional | 69% |
| Games | 56% | Science | 47% |
| Health | **88%** | Shopping | 70% |
| Home | 81% | Society | 57% |
| Kids and Teens | 38% | Sports | 40% |
| News | **17%** | World | 31% |

Classificades 74% $\rightarrow$ Recomanacions **85%**

**59%** Correctes

# Conclusions

**Sistema per protegir logs en temps real als MCW**

Proposat, desenvolupat i avaluat

**Preserva la utilitat dels logs → Distribució a tercers**

Permet monetització

**Protecció ràpida amb alt nivell de privadesa**

Utilitat bona, però millorable

# Treball futur

Millorar categorització, en velocitat i utilitat

Diferents nivells d'agregació per categories

Anonimitzador per blocs de temps

UOC
Universitat Oberta de Catalunya

# Moltes gràcies