

tracking web spam with hidden style similarity

AIRWEB 2006

tanguy urvoy, thomas lavergne, pascal filoche

france telecom R&D

august 10 2006

an example of dynamic web site

www.guardian.co.uk



www.guardian.co.uk/science

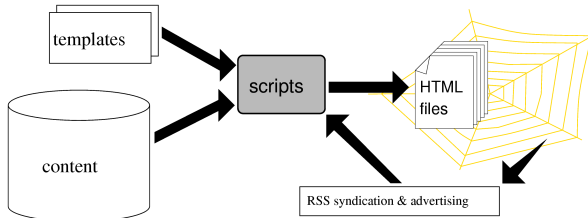
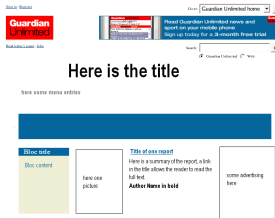


html pages from the same dynamic web site :

- ▶ common visual features : templates
- ▶ common hidden features : web design tools & scripts



dynamic web sites : the three tier architecture



this three tier architecture is also heavily used by spammers to generate large sets of pages



a typical example from "La Sapiensa" datasets

get4me.co.uk

Search: Search

Finance

- Banks
- Building Society
- Credit Cards
- Credit
- Debt
- Consolidation
- Endowments
- Finance
- Financial Calculators
- Financial Consultants
- Financial Guides
- Financial Service

Books

- buy book
- phone book
- amazon book
- library book mark
- sport book
- book shop
- book mark
- book worm
- guinness book of record

Fashion & Clothes

- car insurance
- computer part
- car hire
- car accessory
- car part
- car audio
- car wash
- new car
- car rental

Computers

- computer game
- computer part
- car hire
- computer hardware
- new computer
- computer shop
- mesh computer
- cheap computer
- laptop computer

Credit Cards

- credit card
- card credit
- credit card balance transfer
- uk credit card
- egg credit card
- hollis credit card
- bad credit credit card
- capital one credit card
- gold fish credit card
- 0 credit card

Electronics

- lg electronics
- electronics maplin
- electronics walford
- consumer electronics
- cheap electronics
- hologram electronics
- electronics job
- electronics maplin
- sharp electronics

Estate

- estate
- estate agent
- estate agent london
- halfax estate agent
- real estate
- estate agent kent
- agent bingley bradford estate
- agent estate main road
- agent andrew estate
- coronelle estate agent

Fashion

- man fashion
- fashion world
- lady fashion
- fashion designer
- women's fashion
- fashion model
- fashion jewelry
- fashion underwear

Finance

- finance
- business and finance
- intelligent finance
- yahoo finance
- personal finance
- insoo finance
- finance job
- wellcome finance
- black horse finance

Games

- online game

Gifts

- wedding gift

Health & Beauty

- sexual health

Holiday

- last minute holiday

goto4me.co.uk

www.goto4me.co.uk

Home Links Party Games Mail Savings Loans Insurance Finance Credit

Search: Google Search

Best Results for financial calculators

- [solutions4me Solutions Plus Financial Calculator Financial Calculators Finance Calculator](#)
Solutions4me.com is for Sale! Associated Office Financial Calculator Financial Calculators Finance Calculator Payment Calculator Life Insurance Calculator Loan Repayment Calculator Loan Calculator Loan Calculators Mortgage Calculator Home Mortgage Calculator Mortgage
- [on-line financial calculators - www.calculate-uk.com](#)
Search Results for "on-line financial calculators" Home Financing Calculators Here are some easy to use on-line calculators to help you solve some common financial problems. If you find these calculators useful, be sure to bookmark this page or ... Here are some easy to use on-line calculators ...

www.injuryclaim-uk.com

injuryclaim-uk.com

Home Links Car Road Personal Claims Whistleblow News

Search: Google Search

Best Results

You are searching here: UK

Select search region: Channel Islands - England - Isle of Man - Northern Ireland - Scotland - Wales

Ads by Google

Victims of Truck Accident
Experienced Top Lawyer Group California, Arizona and Texas Cases
Bleakwaterapi.com
Attorney Jobs - New York
Search Attorney job listings in New York today
www.lawjobs.com
Accident Injury Claims
Free information on Accident Injury Claims

4u-insurance.co.uk

4u-insurance.com

Home Car Boat Life Holiday Claims

Search: Google Search

Best Results

You are searching here: UK

Select search region: Channel Islands - England - Isle of Man - Northern Ireland - Scotland - Wales



objectives

main objective of our algorithm : detect if two (or more) html pages were generated by the same tools and with the same templates

applications :

- ▶ detect new instances of already encountered spam pages across different url domains ;
- ▶ point out large sets of generated pages spanning accross different domains (links farm symptom) ;
- ▶ detect web sites boundaries.

main principle :

1. modelize the "html style" of web sites ;
2. compute an "html style similarity" ;
3. use *hash-sampling* to build hss-clusters on a large scale.



[illegible]

The screenshot shows a web browser window with the address bar displaying 'www.google.com'. The page content is a Google search interface. At the top, there are navigation links: 'Home', 'Images', 'Videos', 'Maps', 'Shopping', 'Gmail', 'News', 'Books', 'Flights', 'More', 'Gmail', 'Maps', 'Shopping', 'News', 'Books', 'Flights', 'More'. Below these links is a search bar with the text 'Google Search' and a magnifying glass icon. Below the search bar is a button labeled 'Google Search'. Below the button is a link labeled 'Google Search'.

[illegible]

The screenshot shows the Google search results page. At the top, there is a navigation bar with links for "Google", "Gmail", "Images", "Videos", "Maps", "Books", "More", and "Web". Below this is a search bar with the text "Google Search" and a search button. The main content area displays the search results for the query "Google Search". The first result is "Google Search" with a link to "http://www.google.com/search". The second result is "Google Search" with a link to "http://www.google.com/search". The third result is "Google Search" with a link to "http://www.google.com/search". The fourth result is "Google Search" with a link to "http://www.google.com/search". The fifth result is "Google Search" with a link to "http://www.google.com/search". The sixth result is "Google Search" with a link to "http://www.google.com/search". The seventh result is "Google Search" with a link to "http://www.google.com/search". The eighth result is "Google Search" with a link to "http://www.google.com/search". The ninth result is "Google Search" with a link to "http://www.google.com/search". The tenth result is "Google Search" with a link to "http://www.google.com/search".

Same menu

Same google interface

Same margin

Same "style" of ads

JavaScript partiellement autorisé (script=14) [J+P+O]

"html style" of web sites : hidden features

get4me.co.uk (index)



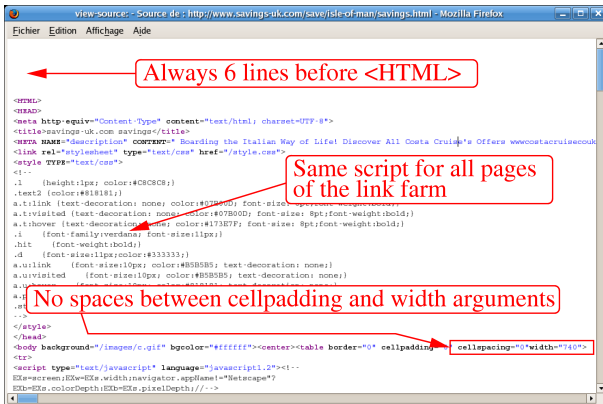
www.goto4me.co.uk



www.injuryclaim-uk.com



4u-insurance.co.uk



"html style" of web sites : focus on html "noise"

<><>

```
< -= " = " / ; = - "><> - . < / >
```

```
< = " " = " ! ' / ' , ! , ! // ! ! ! / , ' , / & ; %  
, ! $ " >
```

```
< = " " = " / " = " _ / . ">
```

```
< = " / ">
```

```
< ! --
```

```
. { : ; ; # ; }
```

```
. { : # ; }
```

```
.. { - : ; : # ; - : ; - : ; }
```

```
.. { - : ; : # ; - : ; - : ; }
```

```
.. { - : ; : # ; - : ; - : ; }
```

```
. { - : ; - : ; }
```

```
. { - : ; }
```

```
. { - : ; : # ; }
```

```
.. { - : ; : # ; - : ; }
```

```
.. { - : ; : # ; - : ; }
```

```
.. { - : ; : # ; - : ; }
```

```
. { - : ; : # ; - : ; }
```

```
. { - : ; }
```

```
-->
```

straightforward filtering : **sed s/[a-zA-Z0-9]//g**

similarity measure ?



HS similarity : franao example

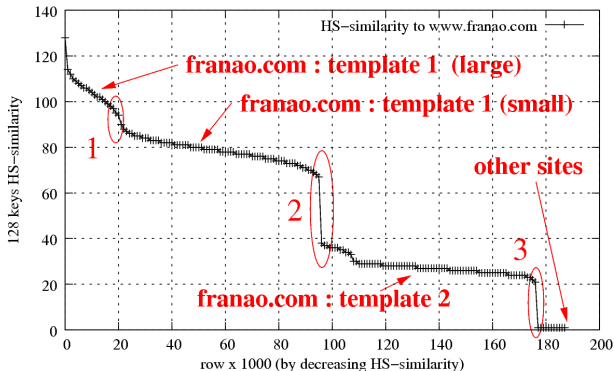
franao template 1 (large)



franao template 1 (small)



franao template 2 (index)



HS similarity : get4u example

get4me.co.uk (index)



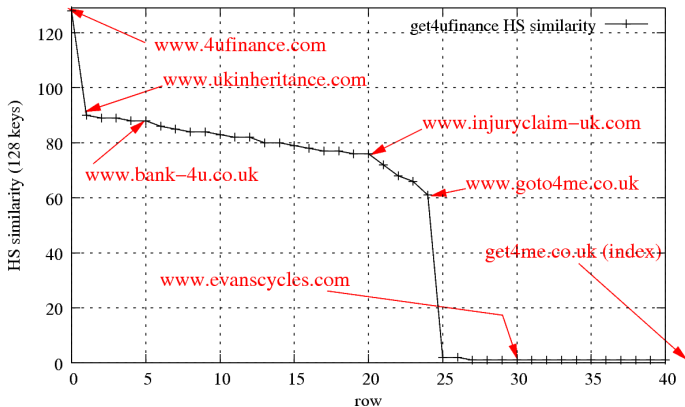
www.goto4me.co.uk



www.injuryclaim-uk.com



4u-insurance.co.uk



building hs-similarity clusters on a large scale :
some recalls about *hash-sampling*



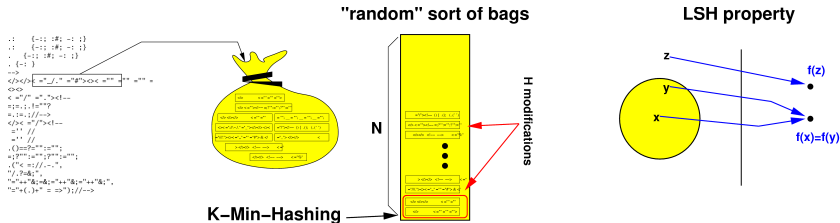
min-sampling, LSH property and clustering

large scale clustering :

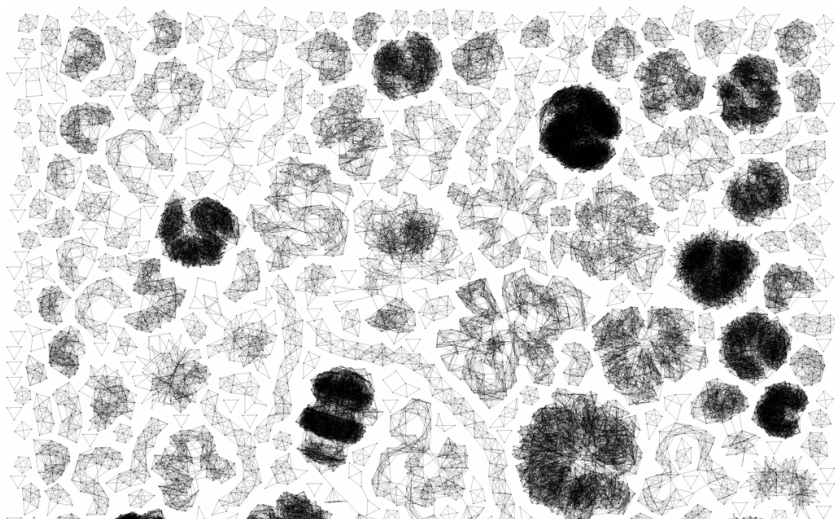
computing HS-similarity for each pair of web pages $\Rightarrow$ quadratic explosion ;

- ▶ min-hashing;
- ▶ quasi-transitivity.

min-hashing :



quasi-transitivity of HS-similarity relation



experimentation on 5M html files



clusters with highest mean similarity and domain count

Urls	Domains	Prototypical member (centroid)	
268	231	www.9eleven.com/index.html	Copy/Paste
93148	313	www.les7laux.com/hiver/forum/phpBB2/membe...	Template (forums)
3495	255	www.orpha.net/static/index.html	Template (Apache)
966	174	www.asliguruney.com/result.php?Keywords=m...	Link farm
122	91	anus.fistingfisting.com/index.htm	Copy/Paste
1148	173	www.basketmag.com/result.php?Keywords=gif...	Link farm
19834	164	www.series-tele.fr/index.html?mo=serie.t...	Template
122	55	www.ie.gnu.org/philosophy/index.html	Mirror
139	101	www.reha-care.net/home_buying.htm?r=p	Link farm
218	195	chat.porno-star.it/index.html	Copy/Paste
177	60	www.ie.gnu.org/home.html	Mirror
2288	44	www.cash4you.com/insuranceproviders/index...	Link farm
626900	70	animalworld.petparty.com/automotivecenter...	Link farm



conclusion



conclusion

- ▶ HS-similarity is good to track templates ;
- ▶ hash-sampling allows a large scale use of HS-similarity ;

perspectives :

- ▶ combination of HSS-clustering and term based clustering
- ▶ clustering refinement by low-pass filtering
- ▶ more testing with uk2006 dataset

The next adversarial step : filtering shuffled html ?



thanks for your attention

