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

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

[Diagram showing the three-tier architecture with templates, scripts, HTML files, and RSS syndication & advertising]
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

www.injuryclaim-uk.com

4u-insurance.co.uk
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 across 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.
"html style" of web sites: visual features

get4me.co.uk (index)  www.goto4me.co.uk  www.injuryclaim-uk.com  4u-insurance.co.uk

Same google interface
Same margin
Same "style" of ads
"html style" of web sites: hidden features

get4me.co.uk (index)  www.goto4me.co.uk  www.injuryclaim-uk.com  4u-insurance.co.uk

Always 6 lines before <HTML>

Same script for all pages of the link farm

No spaces between cellpadding and width arguments
"html style" of web sites: focus on html "noise"

straightforward filtering: `sed s/[a-zA-Z0-9]/\&/g`

similarity measure?
usual term based similarity measures

\[ \text{Sim}(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \]

 bagging

pre-processing

jean smith guardian syria story family blasted road safety lebanese family targeted because driving minivan

Guardian Unlimited &nbsp; &nbsp; Jul 23 4:15 PM

family &nbsp; (;) family ordered to flee were hit because they were driving &nbsp; minivan

http://www.guardian.co.uk/%2Fsyria%2Fstory%2F0%2C2%2C1827422.00.html

lebanese family ordered flee family ordered flee because driving minivan family blasted road safety guardian
html noise pre-processing $\Rightarrow$ HS similarity
HS similarity: franao example

franao template 1 (large)  franao template 1 (small)  franao template 2 (index)

1. franao.com: template 1 (large)
2. franao.com: template 1 (small)
3. other sites

128 keys HS-similarity vs. row x 1000 (by decreasing HS-similarity)
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 ⇒ quadratic explosion;

- min-hashing;
- quasi-transitivity.

min-hashing:

"random" sort of bags

K-Min-Hashing

LSH property

f(x)=f(y)
quasi-transitivity of HS-similarity relation
experimentation on 5M html files
clusters with highest mean similarity and domain count

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