Web Spam Challenges

Carlos Castillo
Yahoo! Research
chato@yahoo-inc.com

WEBSPAM\-UK200[67]

Got a crawl from Boldi/Vigna/Santini in 2006Q3

Wrote to 20-30 people to ask for volunteers

- Most said yes, and most of them didn't defect

Created an interface for labeling

- 3-4 days of work to get it right

Labeled a few thousand elements together

Then, did basically the same again in 2007Q3

Why is it **good** to do collaborative labeling?

The labeling reflects some degree of consensus
- After-the-fact methodological discussions can be very distracting, and actually there was none of it

Webmasters do not harass you
- Responsibility is shared and furthermore you tell search engines not to use these labels

Labellers get insights about the problem

Why is it **bad** to do collaborative labeling?

In this particular problem, it is very expensive

- You get more labels for less money if you just pay for the labels to MTs

Lessons (learned?)

Would do WEBSPAM-UK2006 waiting for the 2nd or 3rd crawl instead of using the 1st one

Would try to raise money for WEBSPAM-UK2007 and do it with MTs

- If the money were enough, try to go for a larger collection

Web Spam Challenges

The good

- Saved a lot of processing to participants, thus ...
- Got several submissions with diverse approaches
- Baseline was strong but not too much
- Side-effect: good dataset for learning on graphs

The bad

- Train/test splits at host level (I, fixed in III)
- Snowball sampling (II, fixed in III)

Lessons (learned?)

Would do mostly the same

Avoid the mistakes

Promote much more the competition

- Try to appeal to a wider audience
- Get sponsorship for a prize

What is the point of all this?

Remove a roadblock for researchers working on a topic

Encourage multiple approaches to a certain problem – in parallel

Keep web data flowing into universities

Allow repeatability of the results

So, if a new dataset+challenge appears

It has to be a good problem: novel, difficult and with a range of potential applications

Why? Because if we are going to encourage many information-retrieval researchers to work on this problem, there has to be **a large treasure chest to split** at the end

Good signals to look for

"The dataset for X removed a roadblock towards a complex information-retrieval problem for which no other dataset existed"

"Research about X was only done inside companies before"

"Problem X was increasingly threatening Web-based work/search/collaboration/etc."

Ideas (1/4)

- Disruptive or non-cooperative behaviour in
  - peer-production sites
    - Examples: review/opinion/tag/tag-as-vote/vote spam
    - Adversary: wants to promote his agenda/business
  - social networks
    - Examples: find fake users
    - Adversary: wants to be seen as multiple independent people
    - Examples: find users that are too aggressive on promoting their own stuff? most social networks have norms against it (wikipedia/kuro5hin/digg/etc.)

Ideas (2/4)

- Plagiarism or missing attribution
  - Web-scale automatic identification of sources for the statements on a document
  - Adversary: wants to make his posting look original

Ideas (3/4)

Checking/joining facts on the Web

- "The capital of Spain is **Toledo** (Wikipedia: Madrid)"
- "The oil spill from the tanker has killed **500 seals** (BBC: 541 seals FOX: 2 anchovies)"

Adversary: wants you to believe something wrong

Related problem: revealing networks of mutually-reinforcing sites pushing a certain agenda

Aspect of **credibility** on the Web (there is already a workshop on that)

Ideas (4/4)

- Simpler problem: validating citations
  - This citation to page P validates the claim it is cited about? where specifically in P?
  - Adversary: wants to convince you of something that is not supported by the pages he is linking

    E.g.: someone wants to convince you that Einstein believed in a personal God by quoting him selectively – but you have access to all his books/letters/etc.

Summary of proposals

Non-cooperative behaviour in peer-production networks

Disruptive usage of social networking sites

Distortions or falsehoods on the web

Citations: missing attribution (plagiarism)

Citations: distorted attribution (invalid citation)