Michael G. Noll, PhD cand.

Web Page Classification: An Exploratory Study of the Usage of Internet Content Rating Systems

Friday, 14-Oct-2005

Table of Contents

- Introduction to Internet content rating
  - Overview, ideas, pros & cons
- Benchmarking requirements
  - Data sets, experimental setup
- Results
  - Availability, trustworthiness, performance
- Conclusions

Introduction to content rating

- Ideas:
  - Allow error-proof classification of web content
  - Focus in practice: pornography, violence, etc.
- Concepts:
  - Special metadata describes web content
  - Content providers self-rate (*label*) their content such as web pages on a voluntary basis
  - End users are empowered to decide themselves which content to see based on these content labels
- Internet “standards”: PICS, ICRA
- Available end user software:
  - MS Internet Explorer (Content Advisor), ICRAplus, …

Content rating with ICRA (1/2)

- ICRA: Internet Content Rating Association
  - Non-profit organization, established in 1999
  - Supported by the European Commission's Safer Internet Action Plan
  - Most prominent content rating system in the Internet today (successor of RSAC)
- ICRA's rating system
  - Rating vocabulary from Dec '00: 45 elements (new vocabulary since July 2005: 49)
  - Covers nudity and sexual content, violence, language, chat, drugs etc.

Content rating with ICRA (2/2)

- Example: www.liasit.lu
- Using the label generator on www.icra.org:
  - No elements listed in “Nudity and sexual material”
  - No elements listed in “Violence”, “Language”, etc.
- Generated label is embedded into LIASIT’s web pages where appropriate

```
<HTML><HEAD>
<META http-equiv=“pics-label”
content=‘(pics-1.1
“http://www.icra.org/ratingsv02.html” 1
gen true for “http://www.liasit.lu/” r (nz
1 vz 1 lz 1 oz 1 cz 1))’>
```

Pros and cons (1/2)

- Pros
  - Manual classification of Internet content should provide better results in terms of performance than automated tools based on heuristics
    - See discussion about **[adv]** in email marketing
  - In particular: IF web content is correctly labeled, than the classification performance will be perfect
    - No false positives (→ www.userfriendly.com blocked)
    - No false negatives (→ www.porn.com allowed)
  - Technically easy to implement and also works for "difficult" content types (videos, Flash, Java)
  - Transparency and simplicity for the average Joe

Pros and cons (2/2)

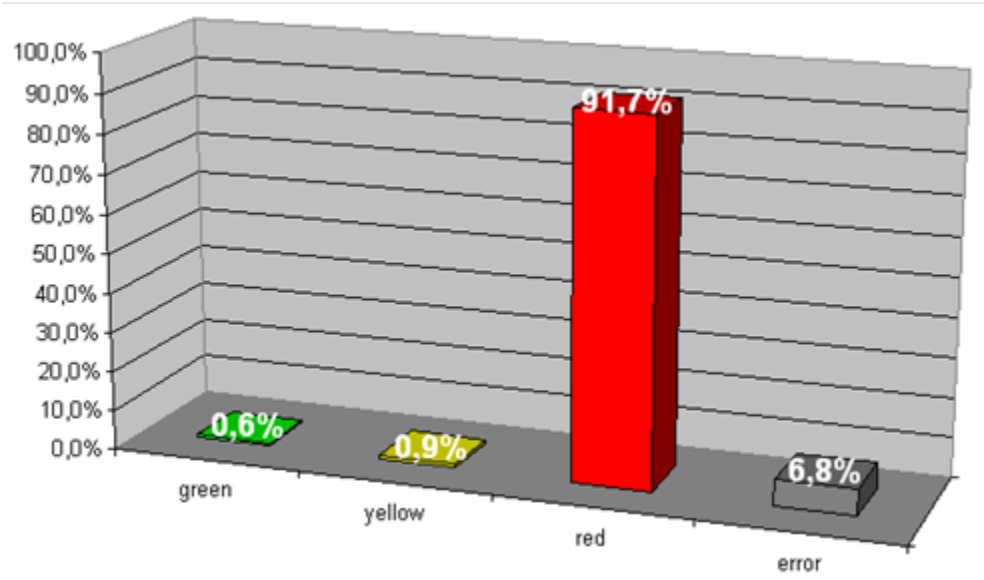
- Cons
  - Voluntariness and trust issues (→ unrated content)
    - Awareness, critical mass, verification of content labels
  - Conflict of interest for content providers
    - Censorship
    - See discussion about .XXX top level domain
  - Criminals don't label – e.g. does not help to fight child pornography
  - Subjectivity/individual vs. objectivity/all
    - Cultural differences, interpretation of content descriptors, handling of fetishes, etc.
  - Different roles in the content creation and publishing process (incl. syndication, RSS feeds)
  - Manual classification is not 100% correct, too!

Benchmarking requirements

- Data sets: unbiased, representative sample of websites
  - Here: data sets based on an anonymized collection of more than 8 million WWW requests of several thousand users collected over a 1-month period
  - TOTAL corpus: 152,617 websites (not categorized)
  - RANDOM5000 corpus: 5,000 websites (categorized)
- Experimental setup: valid and reliable test for content labels
  - Here: development of an automated software tool, which queries the official ICRA label tester web application in “strict rules” mode

Results: availability (1/2)

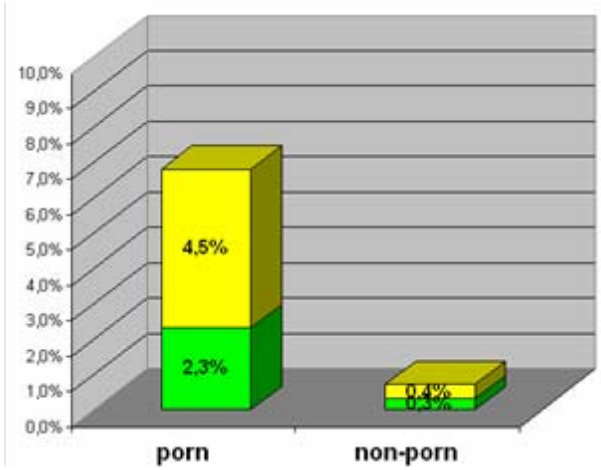
- TOTAL corpus
  - Only 0.6% of analyzed websites are fully, correctly labeled (1.5% if we include partial labels)



- Almost identical results for RANDOM5000

Results: availability (2/2)

- Comparing label availability for pornographic and non-pornographic websites



- Pornographic websites are much more likely to be labeled (6.8%) than non-pornographic sites (0.7%)
  - Possible explanation: higher awareness & “we are in charge”

Results: trustworthiness

- Trustworthiness of content labels: qualitative measurement based on semantics
  - Indicator for the trustworthiness of the corresponding content rating system
  - Directly affects system acceptance of end users
- Trustworthiness := $l_c / (l_c + l_f)$
  - l_c : number of semantically correctly labeled websites (= label matches content)
  - l_f : number of semantically incorrectly …
- Investigated trustworthiness: 81.5%

Results: performance (1/2)

- Assessing the (theoretical) performance of a rating-dependent content filter
  - two possible options to deal with unrated content: 1) allow unrated content, 2) block unrated content
- Performance criteria
  - Recall
    - R := #{correct positive predictions} / #{positive data}
    - “how many porn sites are blocked?”
  - Precision
    - P:= #{correct positive predictions} / #{positive predictions}
    - “how many blocked sites are in fact pornographic?”
  - F1 score
    - F1 := 2*recall*precision / (recall + precision)
    - harmonic average of recall & precision

Results: performance (2/2)

Unrated content will be	allowed	blocked
Recall	1.7%	100.0%
Precision	100.0%	18.9%
F1	3.3%	31.8%
False positive rate	0.0%	99.8%
False negative rate	98.3%	0.0%

- Option 1 (allow unrated) aka “The Reluctant”
  - Pro: will only block websites with a valid pornographic label
  - Con: it almost never catches a porn site
  - *Only marginally better than not using a web content filter at all*
- Option 2 (block unrated) aka “The Merciless”
  - Pro: will catch every single porn site out there
  - Con: blocks almost all non-porn sites, too
  - *Almost equivalent than pulling the network cable (no Internet access)*

Conclusions

- Usage of Internet content rating systems is only marginal today
  - In relative comparison, pornographic websites are much more likely to contain content rating information
- Content labels are *not* 100% trustworthy themselves
  - Basic assumption of content rating systems is false in practice
- The classification performance of rating-dependent content filters is very poor and thus not recommended in practice
  - Based on the situation, users should rather rely on automated filtering tools or classic whitelisting/blacklisting approaches
  - Not every problem is best served with a technical solution: e.g., parents should educate and actively support their children when using the Internet