Text Filtering in POESIA

A General Introduction

Mark Hepple & Neil Ireson
Sheffield University

Internet Pornography (2003)

Volume of material is immense:

- 4.2 million websites (12%)
- 372 million pages
- 68 million search engine requests/day (25%)
- 1.5 billion downloads/month (35%)

Pornographers Tricks

- ***Cyber-squatting***: buy legitimate sounding domain names for porn sites
  - whitehouse.com (c.f. whitehouse.org), civilwarbattles.com, tourdefrance.com
- ***Mis-spelling***: buy domain names that are mis-spellings of important sites for porn

  e.g. googlle.com

Pornographers Tricks (cont)

- ***Doorway scams***: non-porn pages designed for indexing by search engines
  - user accessing page is redirected to porn site
- ***Porn-napping***: buy lapsed domains for porn sites (sell back to ex-owner for large fee)

  e.g. moneyopolis.org: money management site for kids by Ernst & Young

Current Filtering Systems

- Blacklists
  - addresses of ‘unacceptable’ sites
- Keywords
  - block pages containing certain words/phrases

BUT …

- many words ambiguous, meaning depends on context
  - inclusion (resp. exclusion) of these terms as keywords results in over (resp. under) blocking
  - much offensive content in image form
- need for effective filtering based on *content*

Text filtering in POESIA

- POESIA: multiple filters, including filters for image & textual content
- Textual content filters use NLP methods to enhance recognition of categorisation relevant uses of terms
- Textual content filters are *language specific* for English, Italian, Spanish (+ limited for French)
- Approach requires a *language identification* component

Language Identifier

- Models probability distribution of 3-character n-grams
  - Parallel language text
    - 11 European languages (~560 Mbytes total)
  - Smoothing: Good-Turing Estimation
  - Term (Feature) selection: Information-Gain
  - Similarity metric: Entropy measure

Language Identifier Evaluation

- Non-porn pages = ~3% error
- Porn pages = ~2% error
- Pages with low amount of text
  - n-grams <30 (3.5% of pages): ~45% error
  - 30< n-grams <100 (6.5% of pages): ~12% error
  - 200< n-grams (80% of pages) ~0% error
- Problems
  - Imperfect & impure language use
  - Specific terminology
    - Domain specific identifier
  - Proper names

Language Specific Text Filters

- NLP based filtering quite computationally expensive, useful to provide both two filtering modes:
  - Light: gives fast accept/reject for simple cases
  - Heavy: applies more complex methods for difficult cases, i.e. cases not resolved by light filtering

- Light Filters
  - Statistical “bag-of-word” models
  - Stemming, term-selection, term-weighting…
- Heavy Filters
  - NLP Techniques
  - Machine Learning Techniques
  - Localised context

Data Collection for Text Filtering

- Both porn/non-porn data collected automatically by spidering from the Google directory
- Approach has advantages/disadvantages
  - Pros:
    - large dynamic sample of web
    - 72 specific language categories
  - Cons:
    - biased sample
    - misclassified pages