Text Filtering for English

Mark Hepple & Neil Ireson
Sheffield University

English Text Filtering

- Light Filter
  - Statistical Modelling
    - general term distribution
- Heavy Filter
  - Keyword context
    - Contextual Modelling of Common Misclassified Terms

Statistical Modelling

- Strip HTML from porn & non-porn pages
- Remove “stoplist” words
- Stem words (Porter)
- Selection to 300 words according to document frequency
- Use “out-of-place” frequency ranking to determine likely class of unknown page

Statistical Model Evaluation

Predicted / Actual	Harmful	Harmless	Unknown	Total
Harmful	4769	269	48	5086
Harmless	154	4455	233	4842
Total	4923	4724	281	9928
Precision	0.969	0.943		
Recall	0.938	0.920		
F-Measure	0.953	0.931		

- Effectiveness = 93.8%
- Over-blocking = 3.2%

Contextual Modelling of Common Misclassified Terms

- Determine terms which are frequently observed in misclassified porn pages
  - E.g. porn, casino, adult, explicit, eighteen, depict
- Create general contextual patterns of term usage for each category
- Compare use of term in unknown page predicted as non-porn with contextual patterns

Contextual Patterns

$$W_{-3}\ ,\ W_{-2}\ ,\ W_{-1}\ ,\ W_{(explicit)}\ ,\ W_{1}\ ,\ W_{2}\ ,\ W_{3}$$

- Keyword window is ±3 words
- Can consider word, POS & Named Entities
- Create general patterns using “wildcards”
- Matching words based on loci, sets, lists

Light & Heavy Filter Evaluation

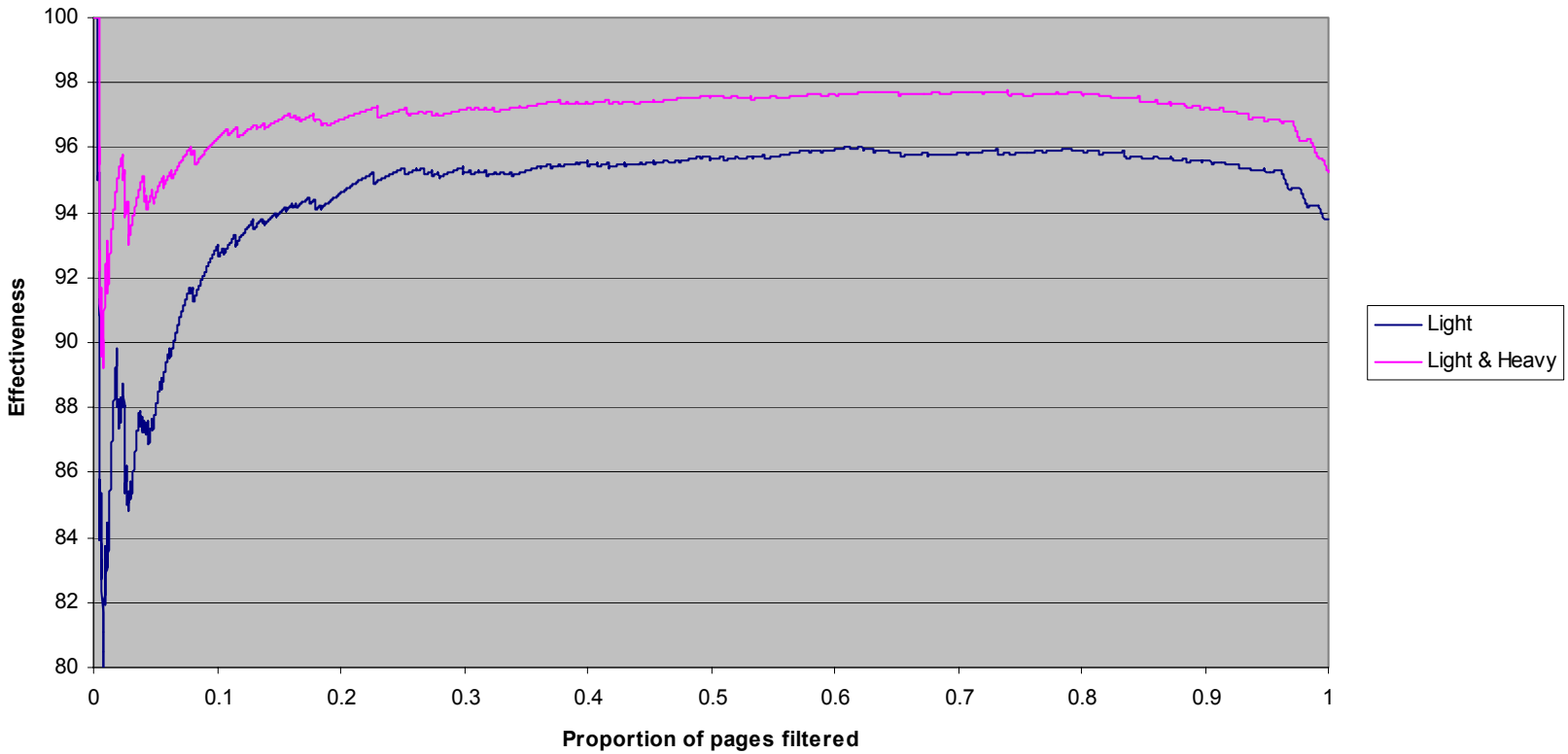
Predicted / Actual	Harmful	Harmless	Unknown	Total
Harmful	4843	195	48	5086
Harmless	163	4446	233	4842
Total	5006	4641	281	9928
Precision	0.967	0.958		
Recall	0.952	0.918		
F-Measure	0.960	0.938		

- Effectiveness = 95.2% (1.4% increase)
- Over-blocking = 3.4% (0.2% increase)

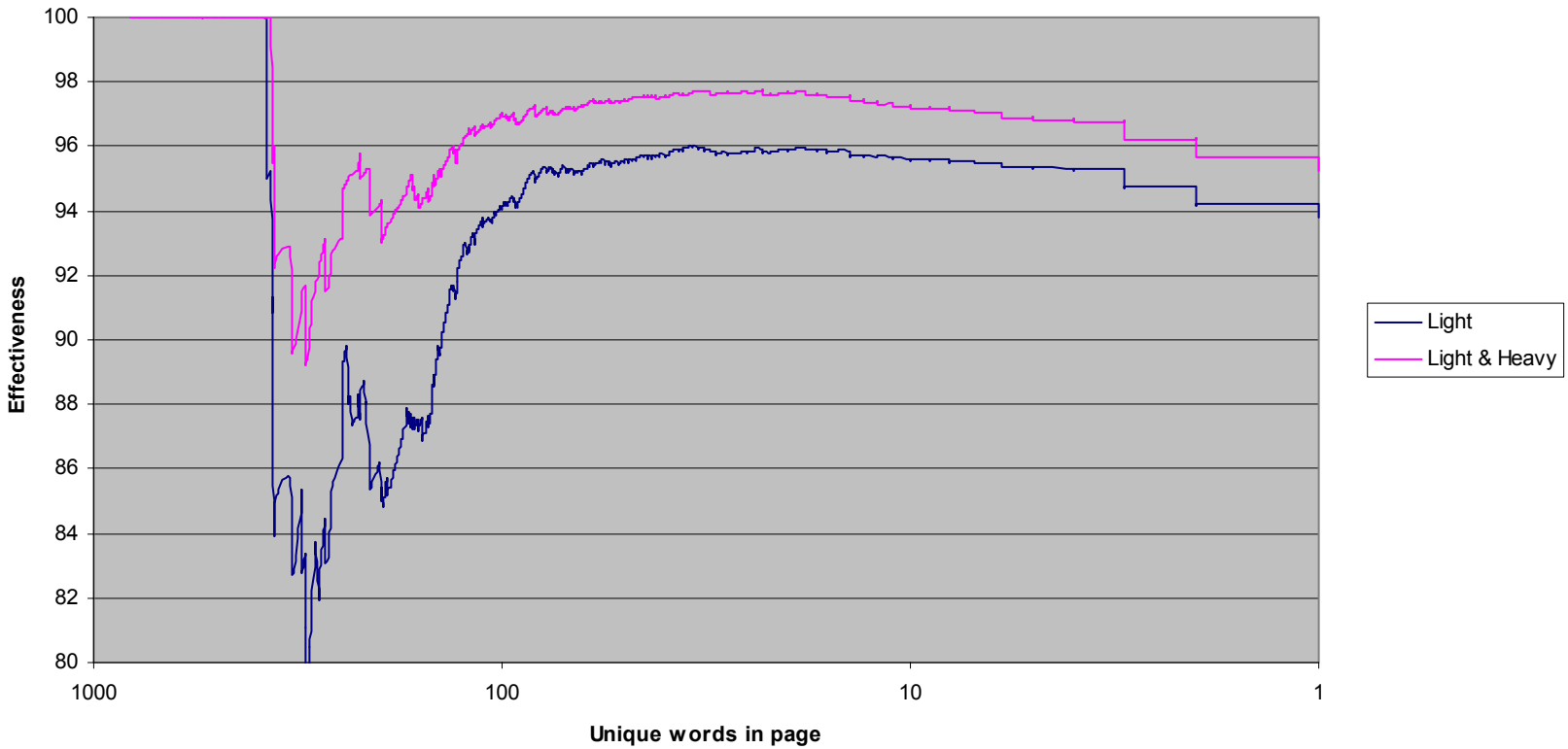
Filter Evaluation

- Pages > 20 Unique Words (78%)
  - Effectiveness: L = 95.9% L&H = 97.7%
  - Over-blocking: L = 2.1% L&H = 2.3%
- Pages <= 20 Unique Words (22%)
  - Effectiveness: L = 86.9% L&H = 87.2%
  - Over-blocking: L = 7.2% L&H = 7.2%

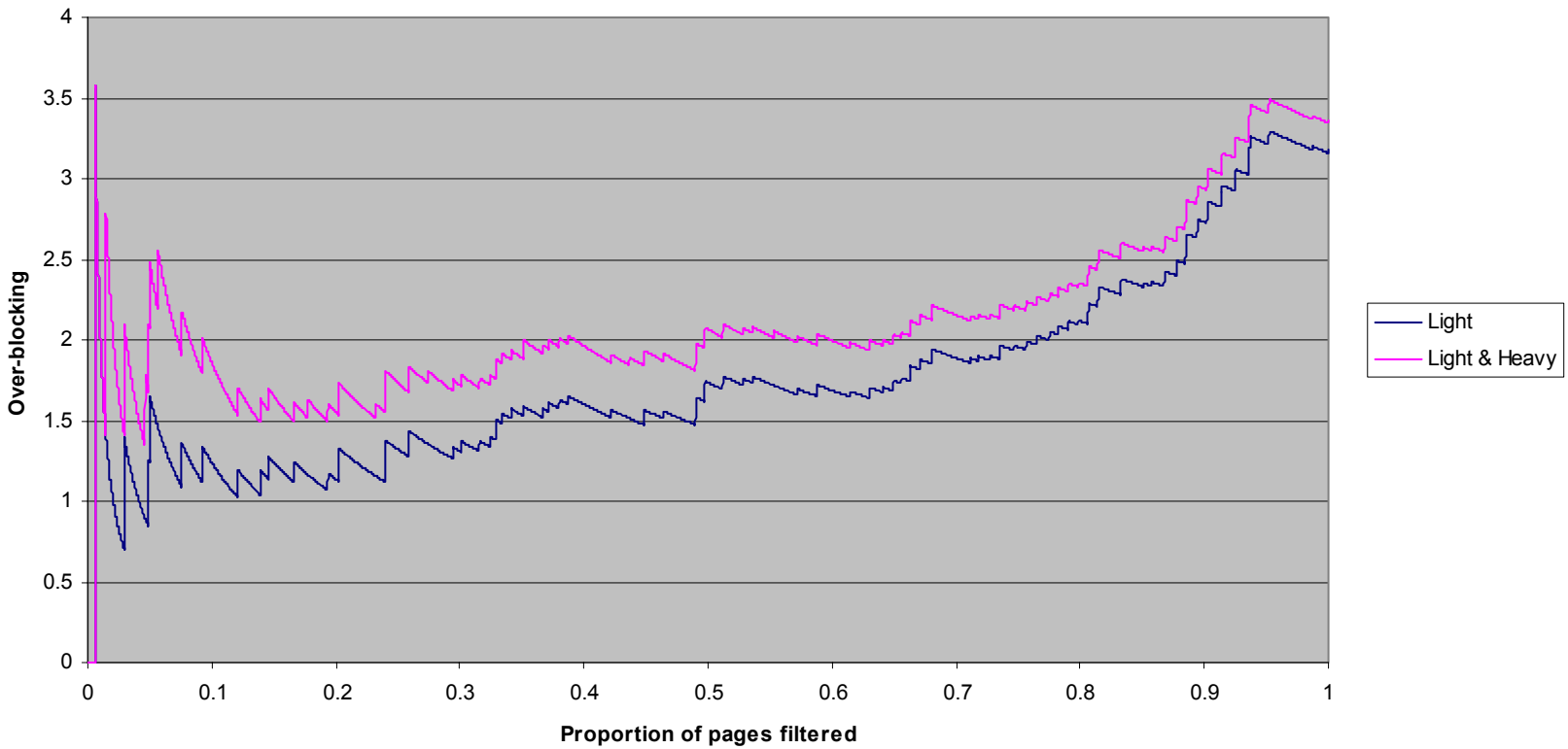
Filter Effectiveness
(Pages Ranked in Descending Size)



Filter Effectiveness
(Pages Ranked in Descending Size)



Filter Over-blocking
(Pages Ranked in Descending Size)



Filter Over-blocking
(Pages Ranked in Descending Size)