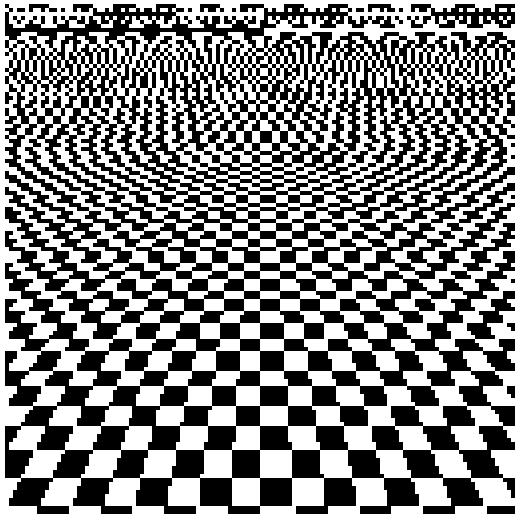
Anti-Aliasing on the Web

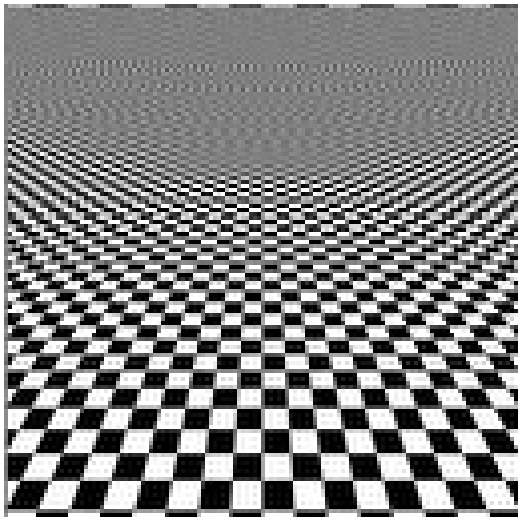
Jasmine Novak
Prabhakar Raghavan
Andrew Tomkins

in WWW 2004

presented by Chad Hogg
2005/12/01

No, Not Graphics!





Anti-Aliasing means finding multiple usernames that belong to the same person.

Why Multiple Aliases ?

- Different sites have different username constraints.
- Artificially increase weight of opinions.
- Express socially unacceptable opinions
- Argue from different perspectives

Similarity Between Aliases

- Writing style features
  - Bag-of-words
  - Misspelled words
  - Punctuation usage
  - Emoticon usage :)
  - Function words
    - Content-free, occur in every domain
    - and, but, which, that, might, this, very, however, ...
    - Used in existing stylometric studies

Accuracy of Different Features

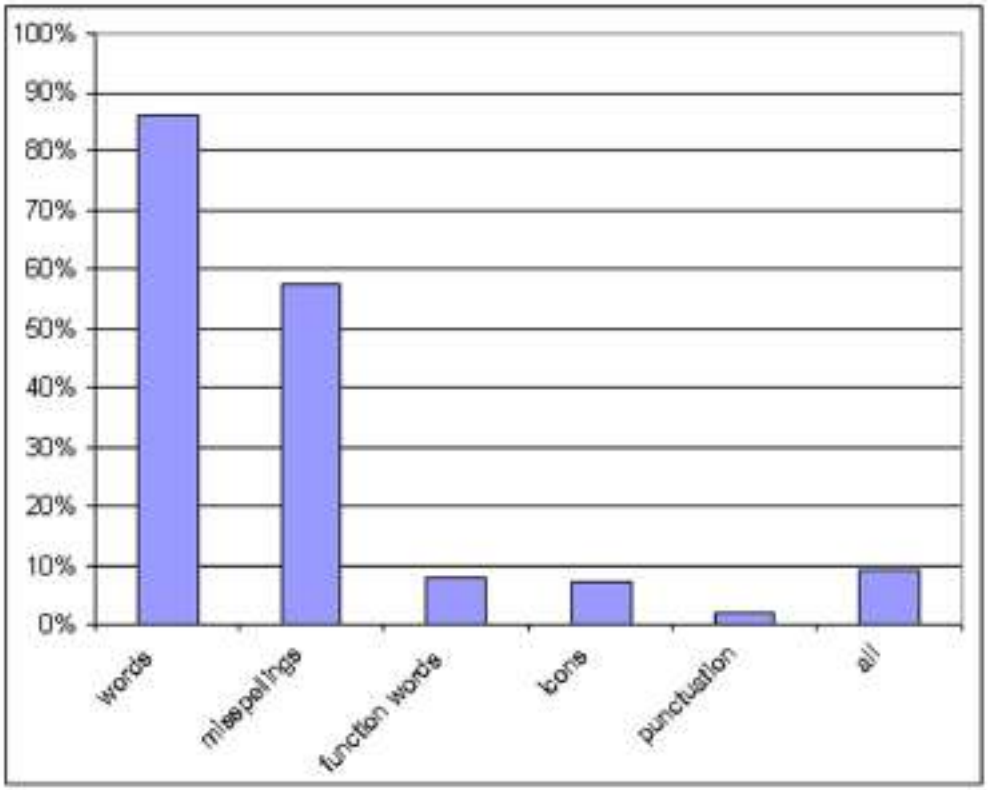


Figure 1: Evaluation of Different Feature Sets.

- Words outperforms all others

Similarity Between Aliases (2)

- Similarity Measures
  - TF/IDF
  - KL Divergence
    - Uses complicated information theory results
    - Intuitively, how efficiently an encoding for one represents the other
  - Probabilistic KL Divergence
    - Not explained well
    - Produces same ranking as KL Divergence

Smoothing

- Naive algorithm fails if first alias uses a word that second alias does not
- General solution: mix small probability of background noise in
- Remarkable find: best accuracy if result is 98% background, 2% alias
  - Because of Zipf's Law: only rare words matter

Measuring Clustering Success

- Variation of Information
  - Traditional approach, also complicated
- Precision / Recall / F-measure
  - Aliases of the same person should be in the same cluster (recall)
  - Aliases of different people should not be in the same cluster (precision)
  - The paper seems to be confused about this?

Clustering Algorithm

- Iteratively combine the two clusters that have the highest pairwise similarity between elements

Let $\mathcal{C} = \{\{a\} | a \in A\}$ be the "current clustering"
Until stopping_condition(C):
    Pick $C_1, C_2 \in \mathcal{C}$ to minimize cohesion($C_1 \cup C_2$)
    Replace C_1 and C_2 in $\mathcal{C}$ with $C_1 \cup C_2$

$$cohesion(C') = \frac{\sum_{a,b \in C'} r(a,b)}{|C'|(|C'|-1)}$$

Data and Results

- Postings from 100 authors on www.courttv.com, each author split into two aliases
- Same thing, but using authors that posted on multiple unrelated topics
- Postings from 400 aliases on the same message boards, not introducing synthetic splits
- Not clear what results were found with any dataset - claim 90% accuracy in conclusion