ISDS 2012 Conference Abstracts

Towards Linking Anonymous Authorship in Casual Sexual Encounter Ads

Jason A. Fries*[1], Alberto M. Segre[1] and Philip M. Polgreen[2]

[1]Computer Science, The University of Iowa, Iowa City, IA, USA; [2]The University of Iowa - Department of Internal Medicine, Iowa CIty, IA, USA

Objective

This paper constructs an authorship-linked collection or corpus of anonymous, sex-seeking ads found on the classifieds website Craigslist. This corpus is then used to validate an authorship attribution approach based on identifying near duplicate text in ad clusters, providing insight into how often anonymous individuals post sex-seeking ads and where they meet for encounters.

Introduction

The increasing use of the Internet to arrange sexual encounters presents challenges to public health agencies formulating STD interventions, particularly in the context of anonymous encounters. These encounters complicate or break traditional interventions. In previous work [1], we examined a corpus of anonymous personal ads seeking sexual encounters from the classifieds website Craigslist and presented a way of linking multiple ads posted across time to a single author. The key observation of our approach is that some ads are simply reposts of older ads, often updated with only minor textual changes. Under the presumption that these ads, when not spam, originate from the same author, we can use efficient near-duplicate detection techniques to cluster ads within some threshold similarity. Linking ads in this way allows us to preserve the anonymity of authors while still extracting useful information on the frequency with which authors post ads, as well as the geographic regions in which they seek encounters.

While this process detects many clusters, the lack of a true corpus of authorship-linked ads makes it difficult to validate and tune the parameters of our system. Fortunately, many ad authors provide an obfuscated telephone number in ad text (e.g., 867-5309 becomes 8sixseven5three oh nine) to bypass Craigslist filters, which prohibit including phone numbers in personal ads. By matching phone numbers of this type across all ads, we can create a corpus of ad clusters known to be written by a single author. This authorship corpus can then be used to evaluate and tune our existing near-duplicate detection system, and in the future identify features for more robust authorship attribution techniques.

Methods

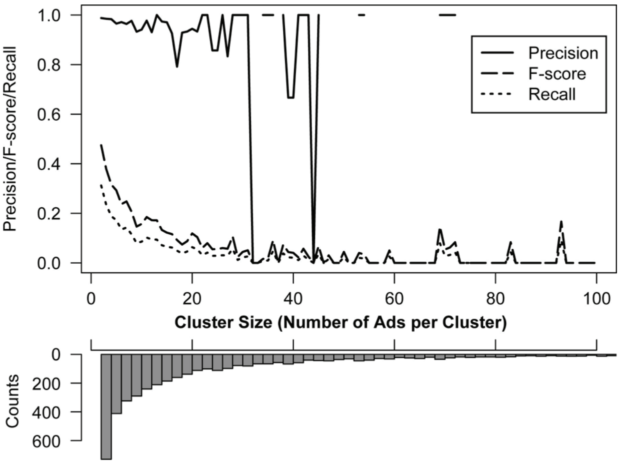
From 7-1-2009 until 7-1-2011, RSS feeds were collected daily for 8 personal ad categories from 414 sites across the United States, for a total of 67 million ads. To create an anonymous, author-linked corpus, we used a regular expression to identify obfuscated phone numbers in ad text. We measure the ability of near-duplicate detection to link clusters in two ways: 1) detecting all ads in a cluster; and 2) correctly detecting a subset of ads within a single cluster. Ads incorrectly assigned to more than 1 cluster are considered false positives. All results are reported in terms of precision, recall, and F-scores (common information retrieval metrics) across cluster size, expressed as number of ads.

Results

652,014 ads contained phone numbers, producing a total of 46,079 authorship-linked ad clusters. For detecting all ads within a cluster, precision ranged from 0.05 to 0.0 and recall from 0.02 to 0.0 for all cluster sizes. For detecting partial clusters, see Figure 1.

Conclusions

We find that near-duplicate detection alone is insufficient to detect all ads within a cluster. However, we do find that the process can, with high precision and low recall, detect a subset of ads associated with a single author. This follows the intuition that an author's total set of ads is itself comprised of multiple self-similar subsets. While a near-duplicate detection approach can correctly identify subsets of ads linked to a single author, this process alone cannot attribute multiple clusters to a single author. Future work will explore leveraging additional linguistic features to improve author attribution.



(Top) Evaluations for partial cluster detection using the near-duplicate identification approach to linking anonymous authorship in Craigslist ads and (bottom) the distribution of ad cluster sizes.

Keywords

Surveillance; Public Health; STDs; Authorship Attribution; Computer Science

References

[1] JA Fries, AM Segre, PM Polgreen .Using Online Classified Ads to Identify the Geographic Footprints of Anonymous, Casual Sex-seeking Individuals. ASE/IEEE International Conference on Social Computing 2012.

*Jason A. Fries
E-mail: jason-fries@uiowa.edu