Digital Detecting: New Resources for Determining Internet Authorship

Lynne Edwards
Media and Communication Studies
Ursinus College
Collegeville, PA 19426 USA
ledwards@ursinus.edu

April Kontostathis
Mathematics and Computer Science
Ursinus College
Collegeville, PA 19426 USA
akontostathis@ursinus.edu

WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.
ACM 978-1-4503-1228-8.

Abstract

Current research on youths and cyber aggression falls into three general areas: rates of occurrence, effects, and prevention of cyberbullying and cyberpredation [6, 10]. One area that has received less scholarly attention is the detection, identification and tracking of cyberbullies and cyberpredators. Cyber aggressors frequently shield themselves through anonymity or the use of creative screen names. Some aggressors employ multiple screen names, allowing for multiple attacks against one or more victims. Authorship identification is a well established field for books and manuscripts published via traditional channels; however authorship detection across social networking sites requires different approaches and techniques. This abstract describes the development of a new dataset for testing automated approaches for identifying authorship across multiple websites.

Keywords

Cyberbullying, cyberpredation, cyber aggression, detection, tracking, identity, screen names.

ACM Classification Keywords

H5.m. Information interfaces and presentation: Miscellaneous. H3.4. Social Networking.

General Terms

Social Networks, Dataset Development, Authorship

Introduction

Interest in cyber crime, particularly cyberbullying and Internet sexual predation has grown in recent years [11]. Automated detection of cyber crime is being established as a cross-disciplinary research area, and machine learning techniques, informed by research in communication theory are being leveraged to develop algorithms for automatic detection of predation and bullying in electronic media. This interdisciplinary subfield is still in its infancy, but it is of great social importance because more children have loosely supervised access to the Internet, and children are participating in online communities at younger ages [2, 9].

Research thus far has been limited to identifying individual posts or threads of conversation that contain aggressive content (either bullying or predatory) [4]. We posit that a broader approach is needed. If an aggressor is using one forum to harass victims, it is likely that other Internet channels are also being utilized, and for similar purposes. Therefore, if we identify a cyber aggressor, we'd like to be able to identify other screen names, and other content that was authored by this individual.

In this abstract we describe our initial approach to the development of a dataset, to be released to the research community, for authorship detection. We also describe some of the obstacles we expect to face in completing the dataset, as well as some of the research opportunities such a resource will provide.

Cyber Aggression and Language Pattern Analysis

We are currently engaged in a multi-pronged study analyzing the communicative approaches employed by cyberbullies and by sexual offenders to approach youths online for sex; we are developing machine learning technologies, Chatcoder, to aid in the recognition and detection of these communication patterns [4].

Olson, et al. established a luring communication theoretical model (LCT) that defines four phases of predation: gaining access, grooming, isolation, and approach [7]. We revised the model as it pertains to online predation. For example, minors often enter chat rooms and engage in conversations with strangers; therefore, gaining access is trivial. Furthermore, the three key aspects that make the encounter predatory are the ages of the victim and predator, the use of grooming language, and the approach (the attempt to actually meet). Therefore, we condensed and simplified the model into three classes: personal information exchange, grooming, and approach [5]. We were able to correctly identify lines containing these categories (as well as unlabeled lines) with 68% accuracy using both a rule-based, and a decision tree learner [4, 5].

Detecting identities of Cyber Aggressors

Cyberbullies and cyberpredators take advantage of the relative anonymity provided by screen names used on the Internet, in general, and on social networking sites in particular. To improve Chatcoder's accuracy and to improve the ability of parents and law enforcement to protect youths, we are beginning a new phase of our research that will allow us to identify authorship based on text-based activities on multiple online social networking sites. These online authorship techniques, once mature, will allow us to develop technology to track cyber aggressors even when they change their ids or move to a different platform.

Other disciplines use machine technologies and statistical processes to identify and define authorship, from cluster-based stylistic analyses [1] to analyzing non-native English speakers' use of Computer mediated Communication (CMC)-specific language to mask their ethnicity [8]. Practitioners in the field of forensic authorship identification have also made extensive use of text-analysis software to identify message authors [3].

Work in Progress

In order to facilitate research on Internet authorship, we are preparing a new dataset, which will be made available to the research community. We are currently manually collecting websites that present information about a particular online user. The userids that we are using to seed our collection activities have been collected randomly from a variety of social networking sites. Our research assistants are using search technologies to identify other userids and social networking sites used by these individuals. They are collecting three different categories of links:

1. Positive links: sites that are definitely this person
2. Possible links: sites that may or may not be this person
3. Negative links: sites that are definitely not the targeted individual, but which may be mistaken for that individual at first glance.

We have collected information on 85 users thus far, with more in progress. A small sample of the data collected appears in Figure 1.

```
<ID>alexpiink</ID>
<Crawl_Date>January 2012</Crawl_Date>
<Personal_Information>
<Name>Alex Cornwell</Name>
<Email>alexpiink69@yahoo.com</Email>
<Yahoo IM>alexpiink69</Yahoo IM>
<Education>Villa Park High School in Villa Park,
California </Education>
</Personal_Information>
<Positive_ID>
<Forum>Twitter</Forum>
<Link>https://twitter.com/#!/alexpiink</Link>
</Positive_ID>
<Positive_ID>
<Forum>Facebook</Forum>
<Link>https://www.facebook.com/profile.php?id=10
0001305437009</Link>
</Positive_ID>
<Possible_ID>
<Forum>Wiki Answers</Forum>
<Link>http://wiki.answers.com/Q/Special:Contributi
ons&target=User:Alexpiink (no information given,
username matches)</Link>
<Explanation>(no information given, username
matches)</Explanation>
</Possible_ID>
<Negative_ID>
<Forum></Forum>
<Link>https://www.facebook.com/people/Alex-
Cornwell/150400885</Link>
</Negative_ID>
```

figure 1: Sample data collection.

Collection process

The collection process begins with a screen name. An undergraduate research assistant is provided with a screen name and is tasked with finding all other screen names that are used by the same individual on a variety of social networking sites. The screen names that are used to seed the process were collected from a variety of sources, including Formspring.me, Twitter, YouTube, and Perverted-Justice.com (screen names for known sexual predators). The research assistants are not given the origin for the screen name. For example, one research assistant was tasked with the id: Likabby. This id was pulled from a recent crawl of pages from Formspring.me.

The research assistants are told to find identifying information about the person behind the screen name, if possible. They look for attributes such as: name (first and last), location, birth_year, age, description, address, phone number, etc. Some of the information will not be available for all screen names. For example, Likabby's real name is Malika, and she is from Orlando Fl.

The research assistants are also told to find additional screen names that are the same individual. Likabby also uses the screen name: yika100 (Myspace). The research assistants use global search engines, such as Google.com, in addition to searching specific websites in order to collect as much information as they can about a particular screen name. They are specifically instructed to search Formspring.me, Myspace, Twitter, and YouTube. For each positive instance, the research assistant records the screen name, the Internet channel, and a link to the user's page on that channel. For example, Likabby's myspace page appears at http://www.myspace.com/yika100.

When gathering data, the research assistants often come across websites that may appear to be the same user, but are not. They are instructed to collect these as "negative ids". This information will be important for testing our authorship detection algorithms in the future. For example, there is a user on Gaia online that initially appeared to be Likabby, but was not. There is also a Likabby on Twitter that is an entirely different person.

When collecting data, some websites are identified that may or may not be the person we are interested in identifying. These ids and links are tracked and labeled as "possible ids." For example, there is a Likabby on youtube which may or may not be owned by our Formspring Likabby.

It takes our assistants an average of one hour to research each userid. They document their results in an XML file, which will be used in the next phase of our project.

Next Steps and Obstacles

Our next step is to crawl these links and build an extensive XML database that can be used for analysis and testing. We anticipate that this crawl will be completed by the end of August 2012.

A crawl of this nature is going to be more complex than just a standard web crawl. We want to capture all of the text (and possibly images) from each link collected by our research assistants, but, because these are social media sites, there will be text from multiple screen names on each site. We need to make sure that we correctly identify the text belonging to each screen name.

Once this crawl is completed, we will have a rich corpus containing samples of written text from many users (if each of our 85 users interacts with another 10 or 20 people on their social networking sites, we will have 850-1700 identities to work with. We will crawl the links for the possible and negative ids also and these links will also have multiple interactions with other users, providing an even richer source of data. At this time we do not have an estimate for the number of interactions we will capture or the size of the final database we will produce.

Usage of the Corpus

The data that we capture can be used in a variety of ways for a number of interesting research tasks. We plan to begin by developing algorithms to segregate the data into the individual userids, based solely on the site text. This is a filtering or classification problem, similar to classifying news stories as sports vs. business, for example. This task will give us insight into what features of the text are most useful for determining authorship.

The unique feature that differentiates this new corpus from other data sets that have been collected from the Internet is that we are starting with a truth set. The original XML files created by our research assistants will serve as ground truths for both training and testing our authorship detection models. We know up front which of the sites (and the associated posts) belong to particular users of interest.

We can also use this corpus to analyze different language patterns by online community. For example, we can use all of the Twitter links and all of the Youtube comment links to contrast the language patterns used on Twitter vs. Youtube. This will allow us to determine if detection engines for cyber aggression need to be site specific, or if a detection algorithm can be effectively used across multiple platforms.

Ultimately we hope to be able to develop sophisticated algorithms that can be used to identify cyber aggressors from small samples of text. For example, if we identify someone who is using grooming language to prey on teens and tweens, we can use a sample of this language to identify all posts by the same author on myriad sites across the web. Having multiple occurrences will not only provide more identifying information about the perpetrator, but will also potentially produce more evidence of harassment activity, which can be used by prosecutors or website administrators (to block additional activity by an aggressor).

Conclusion

We are confident that this corpus will provide useful data for researchers interested in studying the language patterns used by cyber aggressors. Cross-disciplinary collaboration is needed to fully protect youths from cyber crime. We believe that research-based methods that rely on solid approaches with strong theoretical underpinnings provide the best opportunities for identifying and prosecuting these criminals.

Acknowledgements

We thank the anonymous reviewers for their helpful remarks. This material is based in part upon work supported by the National Science Foundation under Grant No. 0916152. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation**.**

Citations

[1] Bagavandas, M. and G. Manimannan. (2009). "Identification of consistent and distinct writing styles: A clustering-based stylistic analysis." *Language Forum*, 35 (2), pp. 57-72.

[2] *Consumer Report Magazine*. (2011). State of the Net: Facebook concerns. ConsumerReports.org. June 2011.

[3] Guillen-Nieto, V., C. Vargas-Sierra, M.Pardino-Juan, P. Martinez-Barco, and A. Suarez-Cueto. (2008). Exploring state-of-the-art software for forensic authorship identification. *International Journal of English Studies*, 8 (1), pp. 1-28.

[4] Kontostathis, A., L. Edwards, and A. Leatherman. (2009). Text Mining and Cybercrime In *Text Mining: Application and Theory*. Michael W. Berry and Jacob Kogan, Eds., John Wiley & Sons, Ltd. 2009. pp. 149-164.

[5] McGhee, I., J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. (2011). Learning to Identify Internet Sexual Predation. *International Journal on Electronic Commerce*. 15 (3). pp. 103-122.

[6] Mitchell, K., D. Finkelhor, L. Jones, and J. Wolak. (2010). Use of Social Networking Sites in Online Sex Crimes Against Minors: An Examination of National Incidence and Means of Utilization. *Journal of Adolescent Health*, 47(2). pp. 183-190.

[7] Olson, L.N., J.L. Daggs, B.L. Ellevold, and T.K.K. Rogers. (2007). Entrapping the Innocent: Toward a Theory of Child Sexual Predators' Luring Communication. *Communication Theory*, 17 (3), pp. 231-251.

[8] Pasfield-Neofitou, S. (2011). Online domains of language use: Second language learners' experiences of virtual community and foreignness. *Language Learning & Technology*, 15 (2), pp. 92-108.

[9] Weichselbaum, S. and E. Durkin. (2011). Facebook lures youngsters with parents' OK. NYDailyNews.com. Posted December 11, 2011.

[10] Willard, N. (2007) *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*, Research Press.

[11] Wolak, J., D.Fikelhor, and K. Mitchell. (2004). Internet-initiated sex crimes against minors: Implications for prevention based on findings from a national study. *Journal of Adolescent Health*, 35 (5), pp.424.e11-424.e20.