Data Mining and Integration to Combat Child Trafficking

Hao Wang
University of Southern California
Department of Psychology
Los Angeles, CA 90089 USA
haowang@usc.edu

Andrew Philpot
University of Southern California
Information Sciences Institute
4676 Admiralty Way #1001
Marina del Rey, CA 90292 USA
philpot@isi.edu

Eduard H. Hovy
University of Southern California
Information Sciences Institute
4676 Admiralty Way #1001
Marina del Rey, CA 90292 USA
hovy@isi.edu

Mark Latonero
University of Southern California
Annenberg Center on Communication
Leadership & Policy
Los Angeles, CA 90089 USA
latonero@usc.edu

ABSTRACT

Women and children are trafficked between countries and within countries for illicit sexual purposes. This is a serious international crime. Domestic traffickers use a variety of means to advertise the illicit sexual services of the children and women they offer, including Internet classified ads, bulletin boards, and social media associated with escort and massage services (EMS). Clients ("johns") of the EMS fronts for prostitution also use the Internet and social media to compare their experiences and offer leads to one another. FBI, the principal U.S. law enforcement authority in this area, has begun implementing a number of initiatives to combat child sexual trafficking. We describe a prototype law enforcement support system to automatically compile and correlate information from open sources about trafficking and sexual abuse of women and especially children. The system, called TrafficBot, employs information retrieval, information integration, and natural language technologies to build a data warehouse allowing various visualizations of information for the benefit of law enforcement. We discuss the current capabilities of TrafficBot, how it could be used by law enforcement, and suggest some future directions.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content analysis and Indexing

I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Design, Theory, Measurement

Keywords

Sexual trafficking, Child trafficking, Law enforcement, Information Integration, Natural Language Processing

1. Sexual Trafficking

Sexual trafficking is the recruitment, transportation (within national or across international borders), transfer, harboring, or receipt of persons for the purposes of commercial sexual exploitation. Sexual trafficking is accomplished by means of fraud, deception, threat of or use of force, abuse of a position of vulnerability, and other forms of coercion.

Trafficking of persons exists in two distinct types: labor trafficking and sexual trafficking. "This new distinction avoids the problem of combining into a single category both labor violations and violations that are more akin to a forcible sexual assault [1]."

Women and men, and girls and boys, are trafficked. In different locales worldwide, the proportion of female to male prostitutes differs to local attitudes, tourism trends, etc. In the U.S., up to 90% of trafficked children are girls [31]. In this article, we will use the terms "women" and either "girls" and "children" (when underage trafficked persons are indicated); but in general this work can apply to both genders.

1.1 Transnational vs. Domestic Trafficking

Law enforcement authorities in the United States distinguish between two modes of trafficking: (1) transnational trafficking, meaning those instances of transnational women and children brought into the U.S. often under false pretenses, for illicit purposes such as prostitution, forced labor, etc., and which fall under the jurisdiction of international agreements DHS, and the State Department; and (2) domestic trafficking, where women and children who are U.S. persons are trafficked, often between states; these latter cases fall under the sole jurisdiction of the FBI. [NR][28]

1.2 FBI Counter-trafficking Efforts

FBI has developed various initiatives to address and combat the trafficking of women and children, including the Human Smuggling Trafficking Center [26].

From fiscal year 2001 through fiscal year 2005, the Civil Rights Division and United States Attorney's Offices filed 91 trafficking cases, a 405% increase over the number of trafficking cases filed from fiscal years 1996 through 2000. In these cases, Department attorneys charged 248 trafficking defendants, a 210% increase over the previous five fiscal years. In addition, 140 defendants of trafficking related crimes were convicted, a 109% increase over the previous five years. [9]

Despite an estimated prevalence of 100,000 to 150,000 sexual slaves in the U.S., fewer than 1,000 victims have been assisted through the efforts of federal, state, and local law enforcement since 2001, when services for trafficking victims were first made available. [11]

1.3 Trafficking of Children

Of particular interest to the FBI is the situation of domestically trafficked minor children who are being exploited sexually. The most common outcomes of trafficked children include homicide, suicide, drug overdose, and adult prostitution. Some trafficked and abused children become sexual abusers themselves. Trafficked children are significantly more likely to develop mental health problems, abuse substances, engage in prostitution as adults, and either commit or be victimized by violent crimes later in life. [30] The Innocence Lost initiative seeks to prevent and combat exploitation of minors, divert minors from prostitution and other associated crimes, and recover children. [27]

As with trafficked women, the trafficked children are controlled using a combination of physical, psychological, and sexual abuse as well as the use of drugs and alcohol to break down their defenses. Children are often traded between pimps and/or moved between various locations, both to evade prosecution and to further their disorientation.

Combatting child trafficking proceeds along multiple fronts. Besides the core crime of procuring, the pimps may also be prosecuted for child abuse, sexual abuse, imprisonment, slavery, and kidnapping. In a law enforcement action, the women and children are often the first to be arrested. Pimps coach minor prostitutes to claim to be adults, for various reasons:

- Adults are more likely to be arrested and sentenced (and thus returned to the streets) expeditiously.
- Law enforcement interventions may be more involved in the case of underage prostitutes (diversion rather than arrest, etc.).
- Similarly, the penalties for pimping children presumably would be higher.

Indeed, the scope of activities that constitute a crime may be considerably broader when the victim is a child. For example, simply to take a minor (in non-custodial context) across state lines is *prima facie* evidence of kidnapping, but would not be so in case of an adult.

1.4 Internet and Child Trafficking

As with licit activities, the use of the Internet in child sexual abuse is increasing with greater use of digital media. Escort and massage services (EMS) serving as fronts for prostitution and other sexual activities advertise the availability of the trafficked girls and women using online directories and social media such as Facebook and Twitter. Besides use of cell phones, some services may employ texting and social media to communicate with their clients. The facilities and their services, both licit and illicit, may be the subject of reviews, discussions, and/or third-party directories on the Internet. In particular, the clients ("johns") share notes about facilities, escort services, and particular experiences and individuals using online bulletin boards ("john boards") as well as social media, including both private and public forums.

In a widely publicized series of stories in 2010, the popular classified ads site Craigslist was found to host many ads for prostitution operations fronting as EMS [29]. Under pressure, Craigslist quickly removed the ads and began policing the categories under which they were posted. However, several other sites, including backpage.com, cityvibe.com, eros.com, humaniplex.com, myredbook.com, and sugardaddyforme.com continue to host ads for EMS, where underage and/or trafficked persons may be involved.

Open sites believed to host "john board" bulletin board and/or chat rooms where clients of sexual services communicate include eroticmp.com and theeroticreview.com. Besides the open material, there are also password-protected "members only" forums on these and other sites.

2. APPROACH

As we have seen, the Internet contains a wide variety of material dealing with EMS. Advertisements and discussions regarding consenting, licit activity is mixed with information about services with underage and/or trafficked persons or otherwise illegal activity.

Our prototype system, TrafficBot, is being developed to provide a comprehensive view of the Internet resources associated with child sexual exploitation. Harvesting and archiving of online data creates a record that can be used to establish history and movements of trafficked individuals. Canonicalization and alias detection helps identify individuals and build networks related by identities and shared attributes. Natural Language Processing and other analytical tools help to detect patterns, determine evidence, and test hypotheses. Cross-source correlation is key to detecting relationships between pimps, prostituted women and girls, and their clients. The combination of the disciplines of information retrieval, information integration, and natural language processing combine to form a system which can create intelligence which may be useful in detecting and combating the trafficking and sexual exploitation of children by law enforcement. Figure 1 depicts the architecture of the in-development system.

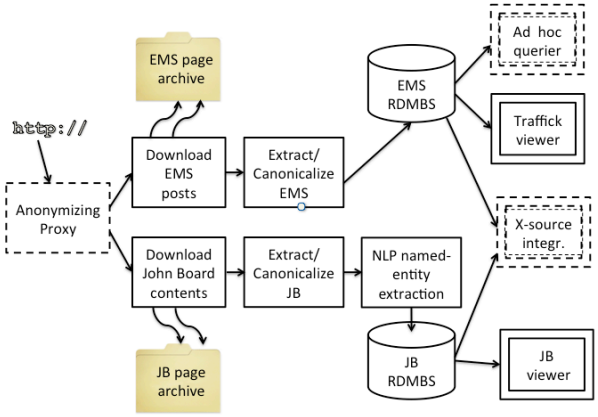


Figure 1: Architecture of TrafficBot system. Items in dashed lines are proposed or under development

2.1 Collection and Warehousing

Like those dealing with legal goods and services, the contents of EMS classified sites and bulletin boards for EMS clients changes from time to time. For trend detection as well as to provide an evidentiary "paper trail", it is important to maintain enough history to support any necessary action. In the prototype version of TrafficBot, we have chosen to crawl the city-specific classified listings every two hours, filtering out duplicates within a 24-hour window and archiving the results of the each day; the "john board" data are downloaded and archived once each day.

It is possible that the providers of a bulletin board or classified ad site prefer that their site not be crawled for law enforcement or academic research purposes, or may wish to limit bandwidth use by crawlers deemed less beneficial to their business purposes. Providers may institute blocking or even redirect unwanted requests to benign or misleading content. Accordingly, to provide the most accurate results and to avoid our work being affected, we are investigating strategies to disguise our affiliation and/or anonymize our HTTP traffic.

Once downloaded and archived, a post is destructured into city (typically embedded in the URL), location list (includes surrounding areas, destinations and/or events), title, timestamp, and age. Full posting text (stripped of HTML) is recorded as well, along with metadata including unique ID, URL, and source ID. A simple post parser is used rather than a more sophisticated landmark grammar technique [32] due to wide variability in formatting and few landmarks available in the posting text. Currently, we store these records with text untokenized (and unindexed) in a relational database.

2.2 Extraction and Canonicalization

2.2.1 From Escort/Massage Services Classifieds

Advertising for illicit EMS contained in online classified advertisement sites functions in some ways like advertising for legal goods and services. As with a conventional shopping integrator or aggregator system, a key goal in this work is detecting and disambiguating the goods or serviced offered. At present, the key attributes of interest are name, location, and phone number.

As is the case with many products and services advertised relatively informally on the Internet, the information related to EMS advertisements (both licit and illicit) and in john boards is irregularly formatted, relatively noisy and often inaccurate. Reasons for the low information quality may include:

- Posters may deliberately use unconventional spellings, formatting, etc. to make their ad seem more natural or informal, to seem more nearly unique to browsers and/or search engines, or to fit in with the prevailing style.
- Posters may have limited Internet accessibility to the Internet (e.g., using mobile phone), resulting in typing mistakes, spell correction errors, etc.
- Posters may have limited writing skills or may wish to appear that they do. One can note that while all ads are written in the first person, i.e., indicate the women offering services, in most illicit cases (as well as many other cases), the actual author is often some other person.
- Posters may wish to avoid detection or identification. In particular, phone numbers are often disguised or encoded, presumably to confuse law enforcement and/or search engines. Phone numbers or posted pictures may be shared between advertisements for different women; filenames or phone numbers may be spelled differently in an attempt to avoid the detection of these aliases. Advertised names of children and women may be changed as they are moved between cities or transferred to a new pimp, or to make an offered woman seem new or different.
- Posters may wish to evade classified advertising systems' list of prohibited words or phrases. These include descriptions of illegal activities as well as explicit sexual content.

Phone numbers presented in ads constitute the most interesting and useful case of information hiding. Phone numbers are typically spelled with unusual punctuation or spacing, character or word substitution, etc. For example, the phone number `(212) 555-1234` might be rendered as `2.I-2/5Fifty-five*one~2 3*for` (with embedded space). At present, we use regular expressions to detect and decode phone numbers. This approach works well in the majority of cases.

No canonicalization is currently performed on names, although it is conceivable to use a dictionary or edit-distance approach to map between possible synonym spellings (especially when other attributes from previous uses could provide supporting evidence).

A typical ad lists multiple locations within a general area; we can use the core city to disambiguate the subsidiary locations (neighborhoods, suburbs, or nearby towns). Encoding of names or locations using a scheme similar to that of phone numbers is possible (e.g., "R1chm0nd" or "D0R1S"), but this has been only rarely observed and so detection of such is not currently implemented.

2.2.2 From "John board" bulletin boards

Extraction from john boards is less well developed at present and the information presented in discussions is even more free-form than the advertising of EMS sites in classified ad listings. Moreover, possible since john board discussions do not produce revenue directly as advertisements do, the forums seem to have less strict enforcement of rules, and thus have a greater fraction of off-topic and spam posts. Using simple regular expression matchers, we extract from john board posts the author(s), title, body text, and location, and any included phone number(s). We store this information annotated with source ID, post ID, URLs, and image URLs. Future work anticipated will extract people names (other than the author), location names (from the free text), activities, and preferences. Besides regular expressions, we are also using the Python NLTK [33][34], particularly for named-entity recognition of persons and locations.

2.3 Alias Detection

Two forms of alias detection are in development at present.

- Association of the same (canonicalized) phone number with two different women. This might indicate that the two women are advertised as a multi-person service; that they belong to the same pimp; or that they are aliases of the same person.
- Use of the same picture for two different women. This indicates either that the two women are aliases, or that the photos are fictitious depictions of both women (but does strongly indicate they share a common link). The easiest case is where the URL is the same, but identical or derived photos also indicate this situation.

2.4 Cross-source Correlation

The aim of cross-source correlation is to use information from one source to extend or correct information from another. For example, a partial or misspelled name in a john board posting might be corrected to agree with the advertised name from the classified ads, if the date and EMS establishment identification can be made to agree.

2.5 Querying and Visualization

Two general query interfaces with simple visualizations have been implemented to display the results of harvesting from EMS classified ads and john boards. Figure 2 details the Timeline view of the EMS data. Given a location, posts are displayed in order, with contact, age, location, text, and thumbnails (dimmed) of the images associated with the ad. Clicking on a phone number leads to the detailed statistics and location information for the phone number in question, as in Figure 4.

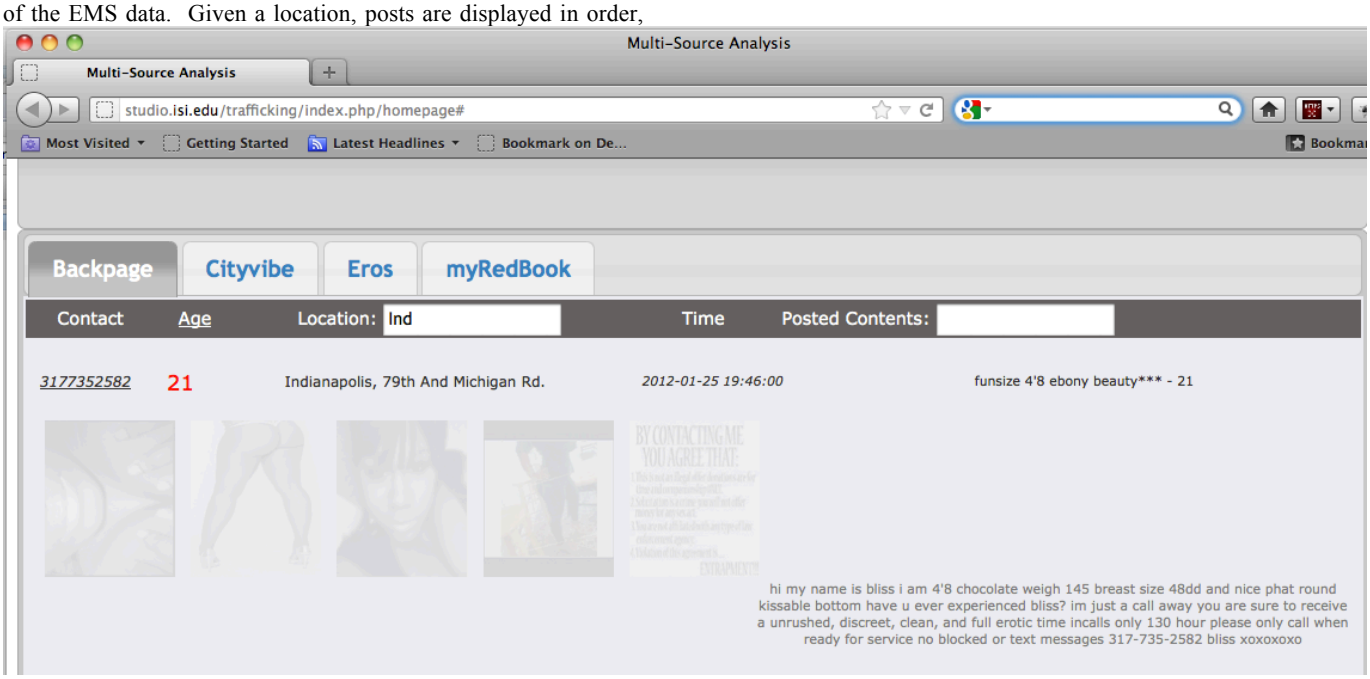


Figure 2: Traffick viewer allows general querying of trafficking database relation. Images are dimmed unless specifically clicked upon.

Similarly, the john board information can be visualized in in a Timeline view, as seen in Figure 3. When browsing deeper on a poster, as in Figure 5, the summary view includes word clouds from the totality of all posts by that participant.

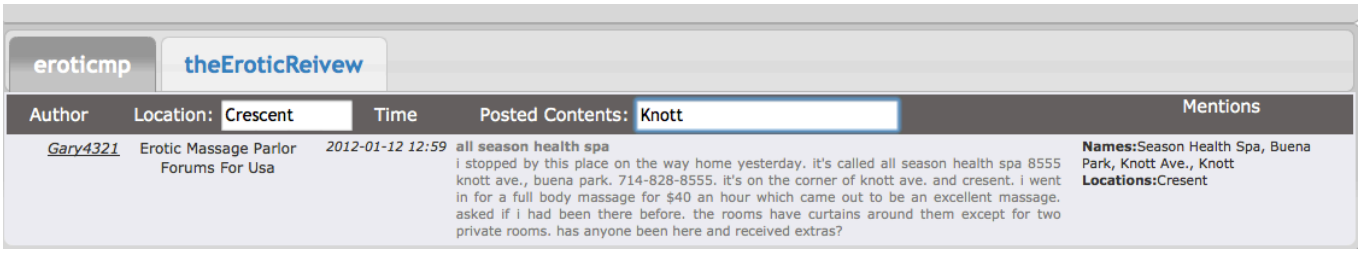


Figure 3: John Board viewer allows general querying of john board results database relation.

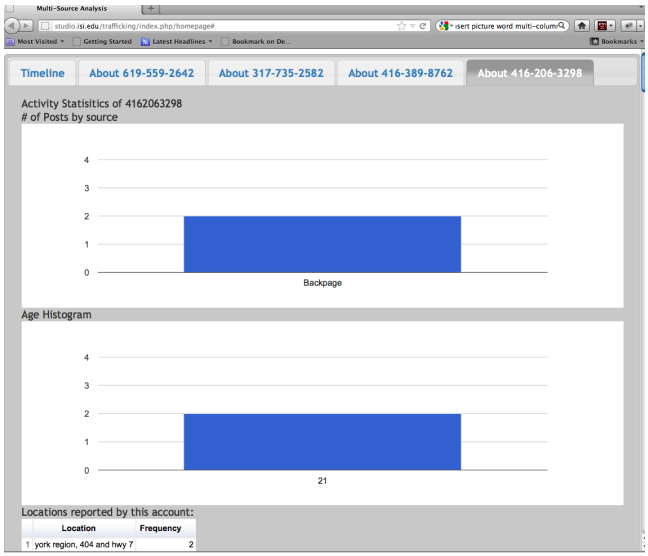


Figure 4: Traffick drill-down: frequency and location details

Figure 5: John Board drill down shows word clouds, extracted phone numbers, and extracted locations.

Finally, an ad hoc query interface facilitating the construction of data driver queries is being prototyped. This will allow narrowly focused investigations when some data believed pertinent have already been identified.

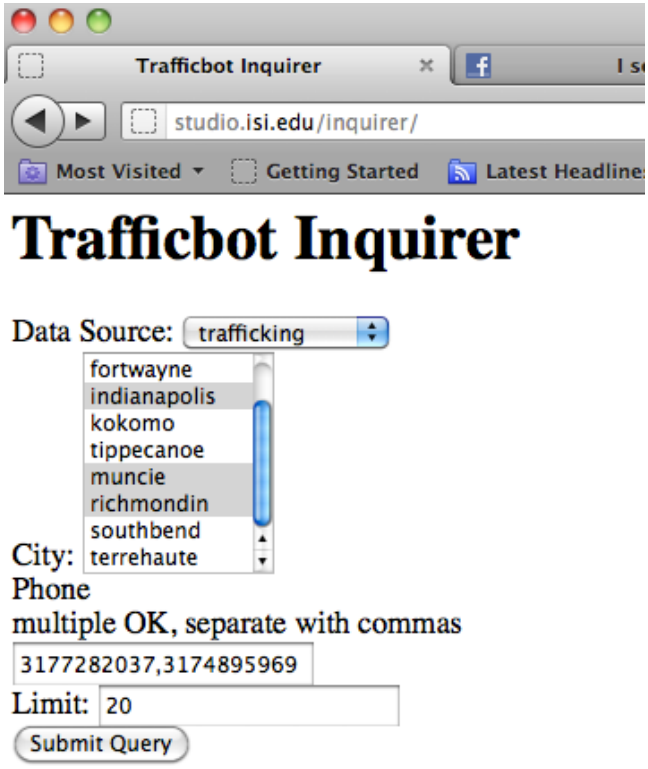
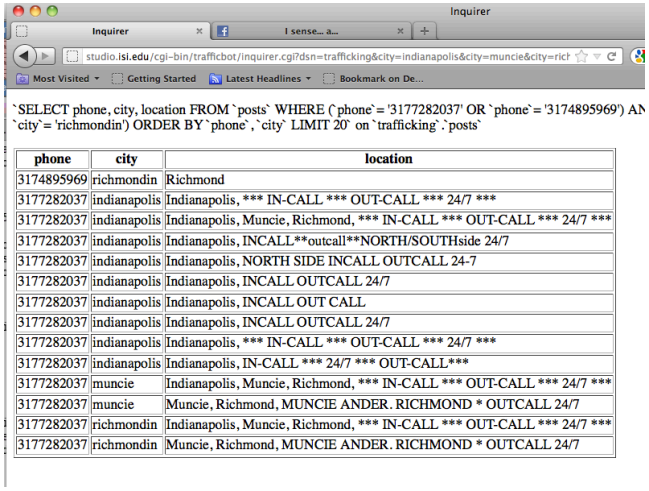


Figure 6 shows an ad hoc query interface allowing construction of arbitrary queries.



phone	city	location
3174895969	richmondin	Richmond
3177282037	indianapolis	Indianapolis, *** IN-CALL *** OUT-CALL *** 24/7 ***
3177282037	indianapolis	Indianapolis, Muncie, Richmond, *** IN-CALL *** OUT-CALL *** 24/7 ***
3177282037	indianapolis	Indianapolis, INCALL**outcall**NORTH/SOUTHside 24/7
3177282037	indianapolis	Indianapolis, NORTH SIDE INCALL OUTCALL 24-7
3177282037	indianapolis	Indianapolis, INCALL OUTCALL 24/7
3177282037	indianapolis	Indianapolis, INCALL OUT CALL
3177282037	indianapolis	Indianapolis, INCALL OUTCALL 24/7
3177282037	indianapolis	Indianapolis, *** IN-CALL *** OUT-CALL *** 24/7 ***
3177282037	indianapolis	Indianapolis, IN-CALL *** 24/7 *** OUT-CALL***
3177282037	muncie	Indianapolis, Muncie, Richmond, *** IN-CALL *** OUT-CALL *** 24/7 ***
3177282037	muncie	Muncie, Richmond, MUNCIE ANDER. RICHMOND * OUTCALL 24/7
3177282037	richmondin	Indianapolis, Muncie, Richmond, *** IN-CALL *** OUT-CALL *** 24/7 ***
3177282037	richmondin	Muncie, Richmond, MUNCIE ANDER. RICHMOND * OUTCALL 24/7

Figure 7 shows the result of the ad hoc query. Note the same phone number coming from three different cities' listings.

3. APPLICATION

3.1 Normal or Body Text

Even in its current incomplete prototype state, the TrafficBot application provides a significant benefit over what a traditional ad hoc law enforcement scan of EMS online resources might provide, and with significantly fewer human resources required.

Periodic and systematic crawling of the same resource base allows for historical views, providing insight into mobility of trafficked women and children, phone numbers, and possibly pimps.

Canonicalization of important data fields such as locations and phone numbers means that searches and joins over a given datum are more likely to yield more and higher quality results.

Tag clouds and named entity recognition identification and merging of locations means that pertinent data can drive the investigation.

Ad hoc query capability provides a middle ground between the basic tabular format and a specialist writing SQL queries against the harvested schema.

4. FUTURE WORK

The primary focus of current work is scaling TrafficBot to the actual size of the task. As women and girls are trafficked throughout the United States, it only makes sense to extend the analysis to all major cities with significant trafficking problems and where appropriate online resources are available. While we have focused on federated national sites up to this stage, it may become useful to extend our analysis with sites that cater to a few or only one metropolitan area. As we scale up, the necessity to cover our tracks using some sort of anonymizing screen will pose a technical challenge as well.

TrafficBot's two main components, the EMS crawler and the john board analyzer, are themselves very immature prototypes.

The EMS module phone number detection and extraction can be extended and made more robust by applying active learning techniques. Document classification and clustering techniques will help us better identify duplicate or near-duplicate posts, which lead to better alias and associate identification. Phone numbers' area codes, especially when they differ from the ad's stated location, may provide clues as to home locations of the girls or the pimps. Named entity resolution and use of name dictionaries should help to make identifications more precise.

The john board analyzer currently makes no attempt to identify posts that are off-topic (advertisements, spam, etc.) Removal of hiding of probable spam will improve precision.

Both applications treat the main text as a bag of words but do little to allow fast indexing for keyword search or analysis. Use of an information retrieval index such as Lucene may greatly improve the power and speed of ad hoc keyword searches.

In longer-term work, we plan to investigate the use of face recognition and age-estimation software.

Finally, we will devote considerable attention to gathering statistics that enable cross-correlation of information aross postings.

5. ACKNOWLEDGMENTS

Our thanks to the FBI Crimes Against Children Unit.

6. REFERENCES

[1] The Protection Project, "What is Trafficking?", Paul H. Nitze School of Advanced International Studies, Johns Hopkins University, 2000 <http://209.190.246.234:80/vt/tra.htm>. in iast.net/fastfacts.html.

[2] USAID Office of Women in Development, Trafficking in Persons: USAID's Response, September 2001.

[3] U. S. Department of State, Victims of Trafficking and Violence Protection Act 2000: Trafficking in Persons Report, July 2004.

[4] Richard, Amy O'Neill, International Trafficking in Women to the United States: A Contemporary Manifestation of Slavery and Organized Crime, DCI Exceptional Intelligence Analyst Program, Center for the Study of Intelligence, November 1999.

[5] "Commercial sexual exploitation position statement." UNICEF UK. (2004, January 28).

[6] U.S. Department of State, Office of the Under Secretary for Global Affairs. (2005, June). *Trafficking in Persons Report - June 2005.*

[7] United Nations Office on Drugs and Crime. (2004, April). *Trafficking in Persons Global Patterns.*

[8] U.S. Department of State. (2006, March 8). *Country Reports on Human Rights Practices - 2005.*

[9] U.S. Department of Justice, Civil Rights Division. (2006, February). *Report on Activities to Combat Human Trafficking Fiscal Years 2001-2005.*

[10] Bales, K. (n.d.). *International Labor Standards: Quality of Information and Measures of Progress in Combating Forced Labor.*

[11] U.S. Department of Justice, Civil Rights Division. (2006, February). *Report on Activities to Combat Human Trafficking Fiscal Years 2001-2005.*

[12] United Nations Office on Drugs and Crime. (2004, April). *Trafficking in Persons Global Patterns.*

[13] Trinidad, A. (2005). *Child pornography in the Philippines.* Psychosocial Trauma and Human Rights Program, UP Center for Integrative and Development Studies and UNICEF Manila, p 14.

[14] Marks, K. (2004, June 28). "In the clubs of the Filipino sex trade, a former RUC officer is back in business." *The Independent.*

[15] United Nations Economic Commission of Europe. (2004, December 12). *Economic roots of trafficking in the UNECE region fact sheet 1.*

[16] Hughes, D. (2002, September 23). *The corruption of civil society: maintaining the flow of women to the sex industries.* Encuentro Internacional Sobre Trafico De Mujeres Y Explotacion, Andalusian Women's Institute, Malaga, Spain.

[17] Thompson, L. (2005, June 22). *The Sexual Gulag: Profiteering from the Global Commercial Sexual Exploitation of Women and Children.* Testimony before the Financial Service Committee, Subcommittee on Domestic and International Monetary Policy, Trade, and Technology, U.S. House of Representatives.

[18] 18. Canadian Medical Association Journal. (2004, July 24). "Prostitution laws: health risks and hypocrisy."

[19] 19. Farley, M. (Ed.). (2003). *Prostitution, trafficking, and traumatic stress.* Binghamton, NY: The Hayworth Maltreatment and Trauma Press.

[20] Ibid.

[21] Lederer, Laura J., Human Rights Report on Trafficking of Women and Children: A Country-by-Country Report on a Contemporary Form of Slavery, The Protection Project, The Paul H. Nitze School of Advanced International Studies, Johns Hopkins University, February 2001.

[22] Dolan, Christine, A Report on the Exploitation of Children Emanating from the Balkan Crises: A Shattered Innocence, The Millennium Holocaust, International Centre for Missing and Exploited Children, April 2001.

[23] Richard, Amy O'Neill, International Trafficking in Women to the United States: A Contemporary Manifestation of Slavery and Organized Crime, DCI Exceptional Intelligence Analyst Program, Center for the Study of Intelligence, November 1999.

[24] Ibid.

[25] Conference on Sexual Trafficking, "Gloria Steinem's Submitted Remarks," Washington, DC, September 13, 1999.

[26] http://www.fbi.gov/about-us/investigate/civilrights/human_trafficking/initiatives

[27] http://www.fbi.gov/about-us/investigate/vc_majorthefts/cac/innocencelost

[28] http://www2.ed.gov/about/offices/list/osdfs/factsheet.html

[29] http://abcnews.go.com/WN/popular-website-craigslist-outlet-sex-trafficking-child-exploitation/story?id=11367581#.TxeEsyOJsbg

[30] http://www.dosomething.org/tipsandtools/11-facts-about-human-trafficking

[31] https://en.wikipedia.org/wiki/Commercial_sexual_exploitation_of_children

[32] http://www.nltk.org

[33] http://www.amazon.com/Natural-Language-Processing-Python-Steven/dp/0596516495