TellFinder: Discovering Related Content in Big Data

Eric Hall, David Schroh and William Wright

Uncharted Software, Toronto, Ontario, Canada

Abstract

In this paper, we present lessons learned in the development of TellFinder, a tool designed to explore domain-specific web crawls using graph analysis and multi-modal visualization. The initial application of the tool was to help combat human trafficking through entity resolution and characterization based on data from sex ads crawled from a variety of publicly available sites. Understanding the nature of the content allows for extraction of domain-specific attributes such as contact information, images and keywords which can be used to aggregate, link and visualize the data. The same principles of performing a deep crawl of a domain, extracting domain-specific attributes, and producing overview and drill-down visualizations can be broadly applied.

Keywords: graph analytics, multi-modal visualization, domain-specific search

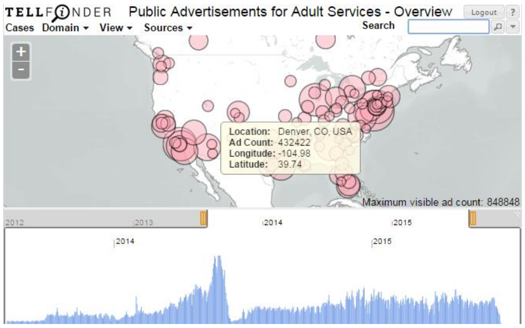
Index Terms: H.5.2 [Computer Graphics]: Graphical User Interfaces (GUI)

1 Introduction

Tools for exploring a sub-domain of the Internet can leverage knowledge of domain content to provide richer analysis using aggregation, linking and multi-modal visualization. TellFinder is a tool that was originally designed to combat human trafficking by providing tools to NGOs and law enforcement to explore the domain of sex ads on the Internet. These ads contain contact information, locations, post times, images and other features which can be used to aid in exploration of the data. TellFinder provides an overview visualization to help the user understand the scope of the crawled data, a linked entity graph, and multimodal visualization to drill down and explore specific content.

2 Domain Overview Visualization

While the primary function of TellFinder is to search, drill down, characterize and report, it is first important to provide users with an understanding of the scope of the crawled data. Without this overview, users can not have confidence in the scope of the data or understand potential gaps. Interactive map and timeline visuals on the opening screen indicate the volume of ads crawled in geographic and temporal dimensions. The overview visualization is also useful in providing a demographic understanding of the data.



TELLFINDER Public Advertisements for Adult Services - Overview

Location: Denver, CO, USA
Ad Count: 432422
Longitude: -104.98
Latitude: 39.74

Maximum visible ad count: 848848

Figure 1: Geo-temporal overview of crawled data.

3 Aggregation and Linking to Explore a Large Shared Attribute Graph

TellFinder uses aggregation and an expand-on-demand interface [3] to explore relationships within the crawled data. The initial data for TellFinder involved over 50M records with just over 1M pieces of contact information (e.g. phone numbers, web urls). An initial attempt at producing a force directed bipartite graph containing nodes for both ads and their shared attributes showed promise in that interesting structures such as clusters and bridge nodes could be observed. However, this approach was unwieldy for end users due to the data volume and so a second attempt was made which compressed highly connected clusters into single nodes. Each node in the resulting graph notionally represents a set of posts from an entity (i.e. a person or organization) with repeated contact information.

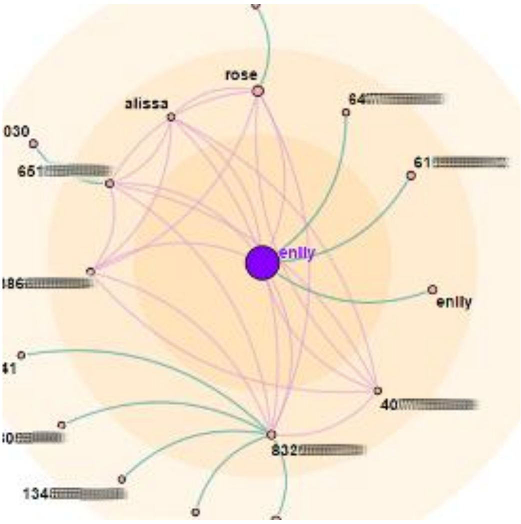
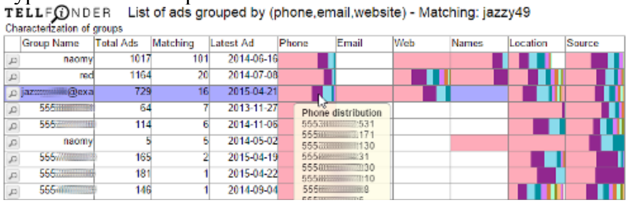


Figure 2: A cluster graph showing notional entities.

4 Entity Lists and Entity Characterization

The success of entity determination by clustering the shared attribute graph led to listing and characterizing entities in a table containing all of the ads with entities at a given location or matching a user defined query. The entity list is useful for understanding the behavior of entities at a particular location, such as typical movement patterns.



TELLFINDER List of ads grouped by (phone,email,website) - Matching: jazzy49

Characterization of groups

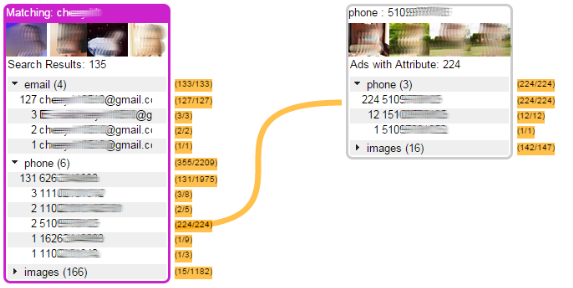
	Group Name	Total Ads	Matching	Latest Ad	Phone	Email	Web	Names	Location	Source
⌕	naomy	1017	101	2014-06-16						
⌕	red	1164	20	2014-07-08						
⌕	jaz[redacted]@exa	729	16	2015-04-21						
⌕	555[redacted]	64	7	2013-11-27						
⌕	555[redacted]	114	6	2014-11-06						
⌕	naomy	5	5	2014-05-02						
⌕	555[redacted]	165	2	2015-04-19						
⌕	555[redacted]	181	1	2015-04-22						
⌕	555[redacted]	146	1	2014-09-04						

Phone distribution
555[redacted] 531
555[redacted] 171
555[redacted] 130
555[redacted] 31
555[redacted] 30
555[redacted] 10
555[redacted] 8

Figure 3: A list of entities characterized by associated content.

5 Refinement of the Entity Graph and Entity Characterization

The entity graph described above was significantly more usable than the initial shared attribute graph but still became unwieldy due to two features of the data. First, because of efforts to circumvent deliberate obfuscation of contact information, there was a lot of noise and false positives resulting in both large aggregate nodes and excessive linking. Secondly, the addition of an image similarity service produced a large number of valid links. The solution to these problems was to produce graph nodes rich in information and affordances. Each graph node in the new "TellFinder Explorer" graph contains a breakdown of the attributes contained within the documents comprising the entity and the ability to selectively expand and view nodes related via a particular common attribute.



Matching: ch[redacted]

Search Results: 135

email (4) (133/133)
127 ch[redacted]@gmail.c (127/127)
3 [redacted]@g (3/3)
2 ch[redacted]@gmail.c (2/2)
1 ch[redacted]@gmail.c (1/1)
phone (6) (355/2209)
131 626[redacted] (131/1975)
3 111[redacted] (3/8)
2 110[redacted] (2/5)
2 510[redacted] (224/224)
1 1626[redacted] (1/9)
1 110[redacted] (1/3)
images (166) (15/1182)

phone : 510[redacted]

Ads with Attribute: 224

phone (3) (224/224)
224 510[redacted] (224/224)
12 151[redacted] (12/12)
1 510[redacted] (1/1)
images (16) (142/147)

Figure 4: Descriptive and interactive cluster nodes (Images and phone numbers have been blurred.)

Three node types are used in the explorer graph. Entity nodes represent all documents associated with a given person or organization. Attribute nodes represent all documents associated with a particular attribute value. Search result nodes represent all documents matching a given query.

6 Linked Multimodal Visualization

The entity graph was enhanced by linked visualizations characterizing the behavior of the selected entity such as movement by examining posts vs. time and location, a word cloud, a map, and a table showing the specific ads.

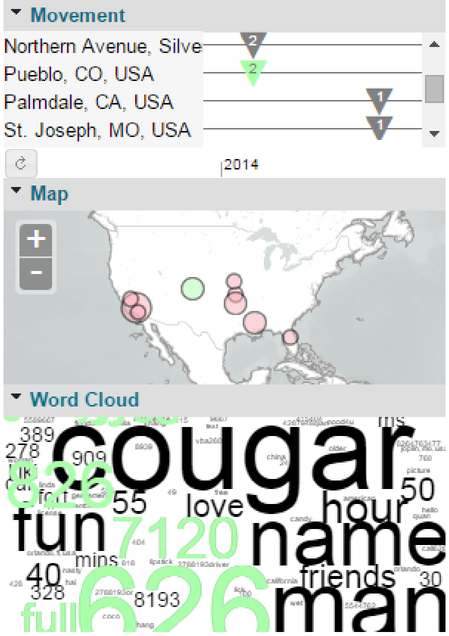


Figure 5: Multimodal visualization characterizing an entity.

7 Reporting

An essential component of the analysis workflow is the ability to record results for future reference and reporting. A side panel was added to the application to allow users to create "Case Files" into which entities and attributes could be dragged and dropped from the other widgets. These domain specific records of relevant domain objects can be used to generate meaningful reports.

8 Extension to Other Domains

The concepts described above have been successfully pivoted to work with data from different domains. Forum data relating to potential illegal firearms sales was the first domain explored. Attributes such as online identities, thread titles and weapon types have been used for aggregation and linking. Additional domains will be explored by ingesting a large corpus of raw html pages, extracting domain-specific information and then performing aggregation, linking and visualization with the same techniques described above.

9 Conclusions and Future Work

TellFinder has proven to be a practical tool for exploring a large corpus of domain specific documents with extracted content. It has been deployed as a pilot to several law enforcement organizations for evaluation and feedback. It continues to be used on a daily basis.

The TellFinder system has been extended to quickly pivot to new domains by specifying new corpuses and relevant attributes stored using ElasticSearch. Many potential additional domains will be explored going forward.

A major feature that has not been fully explored is the ability to edit TellFinder Explorer Graph nodes. This feature could be used to remove or alter invalid data or to produce user defined nodes allowing users to curate the entity resolution and then explore the behavior of the resulting aggregates.

10 Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Number FA8750-14-C-0275. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.

References

[1] Jonker D., S. Langevin, D. Gauldie and W. Wright. Influent: Scalable Transaction Flow Analysis with Entity-Relationship Graphs. Poster. IEEE EuroVis, 2014.

[2] Influent, 2014. URL: http://influent.io/.

[3] Van Ham F. and A. Perer. Search, show context, expand on demand: Supporting large graph exploration with degree-of-interest. In Visualization and Computer Graphics, volume 15, pages 953-960. *IEEE Transactions,* Nov 2009.