

•
•
•
•
•
•
•

15400 Calhoun Drive, Suite 400
Rockville, Maryland, 20855
(301) 294-5200
<http://www.i-a-i.com>

Intelligent Automation Incorporated

Information Tailoring Enhancements for Large-Scale Social Data

Final Report

Reporting Period: September 22, 2015 – September 16, 2016

Contract No. N00014-15-P-5138

Sponsored by

ONR, Arlington VA

COTR/TPOC: Dr. Rebecca Goolsby

Prepared by

Onur Savas, Ph.D.



DISTRIBUTION A

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE					Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.						
1. REPORT DATE (DD-MM-YYYY) 26/09/2016		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) September 22, 2015 – September 16, 2016		
4. TITLE AND SUBTITLE Information Tailoring Enhancements for Large-Scale Social Data				5a. CONTRACT NUMBER N00014-15-P-5138		
				5b. GRANT NUMBER		
				5c. PROGRAM ELEMENT NUMBER		
				5d. PROJECT NUMBER		
6. AUTHOR(S) Onur Savas				5e. TASK NUMBER		
				5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Intelligent Automation, Inc. 15400 Calhoun Drive, Suite 190 Rockville, MD 20855				8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of Naval Research (BD253) 875 N. Randolph Street Arlington, VA 22203-1995				10. SPONSOR/MONITOR'S ACRONYM(S)		
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)		
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION A: Approved for public release; distribution is unlimited.						
13. SUPPLEMENTARY NOTES Report contains color						
14. ABSTRACT In this project, to achieve our goals of (i) further enhancing capability to analyze unstructured social media data at scale and rapidly, and (ii) improving IAI social media software Scraawl's features, we have (i) designed and implemented temporal community detection and influence discovery algorithms, (ii) enhanced Scraawl UI for improved usability and navigation, (iii) improved the computational framework of Scraawl, (iv) enhanced Named Entity Recognition (NER), and (v) designed and developed geo referencing capabilities. The features are released as part of Scraawl 2.0.						
15. SUBJECT TERMS social media, information tailoring, large-scale analysis, OSINT						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT SAR	18. NUMBER OF PAGES 12	19a. NAME OF RESPONSIBLE PERSON Dr. Rebecca Goolsby	
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include area code) (703) 588-0558	

Table of Contents

1	Executive Summary	3
2	Technical Work.....	4
2.1	Design and Implementation of Temporal Analytics	4
2.2	Upgrade Scraawl User Interface (UI)	6
2.2.1	Bookmarking and Case Report Capability	6
2.2.2	Enhancement of User Profiles.....	7
2.2.3	Translation of Words/Phrases in Searches.....	7
2.2.4	Usability and Navigation Improvements	7
2.3	Upgrade Scraawl Computational Framework to Increase Robustness.....	8
2.4	Enhance Named Entity Recognition (NER)	9
2.5	Incorporate Geo-reference Analytics.....	9
3	Conclusion.....	11
4	References	12

1 Executive Summary

In this project, to achieve our goals of (i) further enhancing capability to analyze unstructured social media data at scale and rapidly, and (ii) improving IAI social media software Scraawl's features, we have accomplished the following.

First, we designed and implemented temporal community detection and influence discovery algorithms and associated visualizations using Twitter data. These capabilities improved our understanding of how features associated with influence and communities with time as social conversations about a particular event begin to grow.

Second, we improved Scraawl UI by designing and implementing (i) bookmarking and case report capability, (ii) user profile enhancements such as linking social media accounts with Scraawl, (iii) translation of words/phrases in searches, and (iv) a communication methodology in the UI so that the user gets notified of data changes in the backend for improved usability and navigation.

Third, we improved the computational framework of Scraawl to make it more robust and handle higher data loads. In particular, we improved (i) the messaging architecture, (ii) data redundancy, and (iii) the service availability.

Fourth, we enhanced Named Entity Recognition (NER) capabilities of Scraawl by incorporating (i) part-of-speech tagging using GATE, (ii) enhancing the gazetteers with multilingual entities, and (iii) adding multi-lingual name matching capabilities.

Fifth, we developed analytics that evaluate the text of the tweets and user profiles to extract and identify location information, and geo reference the location on a map. We identified locations through NER and displayed geo-coordinates (longitudes and latitudes) for locations recognized by NER. We then provided the capability to locate locations mentioned by tweets on a map.

Finally, we released IAI social media software Scraawl 2.0 that incorporates these features.

2 Technical Work

2.1 Design and Implementation of Temporal Analytics

Based on our capabilities to discover influential users and community detection from Twitter, we designed temporal analysis algorithms covering all aspects of front end, back end, and services. In particular, inputs to these algorithms include the Tweet dataset (i.e., Scraawl report) and number of time fractions (e.g., 4). Based on these user provided parameters Scraawl partition Twitter data into different time fractions (according to tweet post times), run Twitter analytics over each time fraction, and display how Twitter analytics results change over time. A representative analysis is shown in Figure 1.

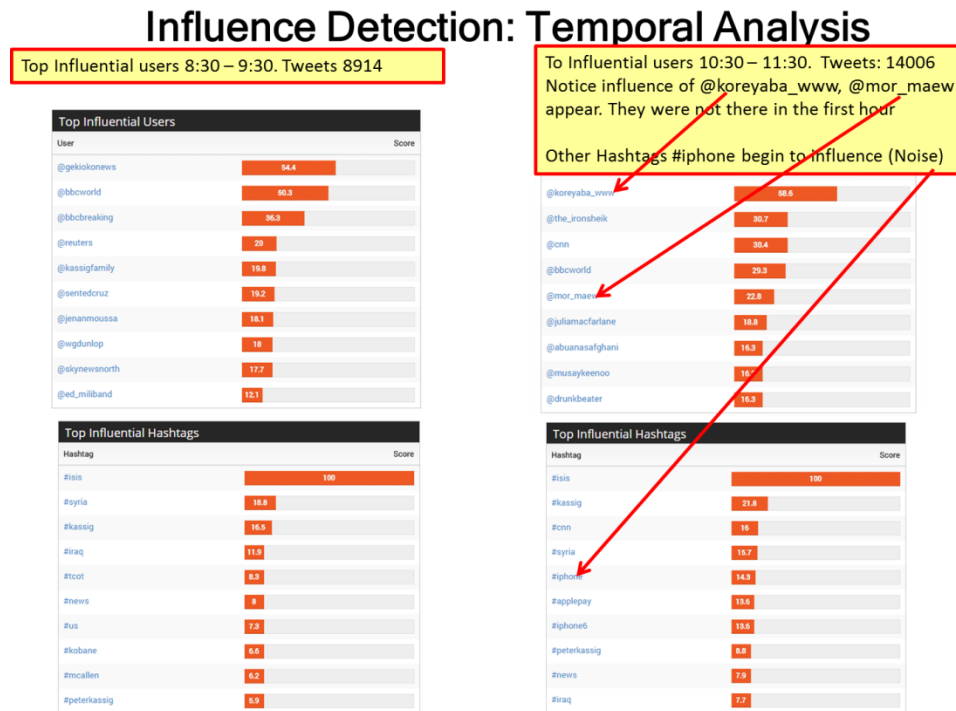


Figure 1: Temporal Analysis of Influential Users in a Report.

In particular, we implemented a smart data binning for influential nodes analytical capabilities for twitter data source to evaluate the temporal evolution of this analytical capability.

Our lightweight temporal software implementation is described as follows:

1. We extended the influential nodes web service to have the time period (*IntervalInMinutes*) for a bin in minutes as a parameter. The caller of the web service can run the analytic if he provides the value zero to the *IntervalInMinutes*, otherwise the caller will run the temporal analytics.
2. We modified the tweets database extraction routine to pull the timestamps of the time of tweet creation for every tweet as reported by twitter.
3. We developed a new smart binning routine that has as input the *StartingTime*,

EndingTime, and an *IntervalInMin*:

- a. We retrieve *hour* and *minutes* from the timestamp of the first tweet in the report.
 - b. We calculate the next *EndingTime* of the bin by finding the next minute *m* that satisfy the following equation $m = K * IntervalInMinutes$ among all the minutes in the first hour of the social media data report.
 - c. We compute all the bins by shifting the *StartingTime* and *EndingTime* of the first bin *IntervalInMinutes* until the *StartingTime* is greater than the last tweet time in the report
 - d. We replace the *EndingTime* of the last bin with the last time of a tweet in the report
4. The posts are divided into batches and grouped by *report_id* and an identical time period using a smart data binning model approach. We developed a routine to return all the tweets in each bin such as the time *t* of a tweet satisfy the following condition: *Starting_Time_Of_Bin* $\leq t < End_Time_Of_Bin$
 5. We developed a new routine to store the application results along with additional fields = { *StartingTime* of each bin, *EndingTime* of each bin, and the *BinNumber* } in a specific table of MySQL database. We extended the analytics influence results table in MySQL database and the Influence model to address the storage of the additional fields.
 6. We run our influence discovery application on those batches of social media data within the report and we store the results in MySql database.
 7. On the GUI, we developed a scrolled list of temporal batches for the user to see the evolution of the analytical results.

Example web service request for temporal influential nodes:

```
{
  "topk": "10",
  "topn": "10",
  "IntervalInMin": "15",
  "promise": "http://www.google.com",
  "error": "http://www.google.com"
}
```

Example web service response for temporal influential nodes:

```
{
  "status": "OK",
  "message": "Analytics completed"
}
```

Example web service request for influential nodes on all tweets:

```
{
  "topk": "10",
  "topn": "10",
```

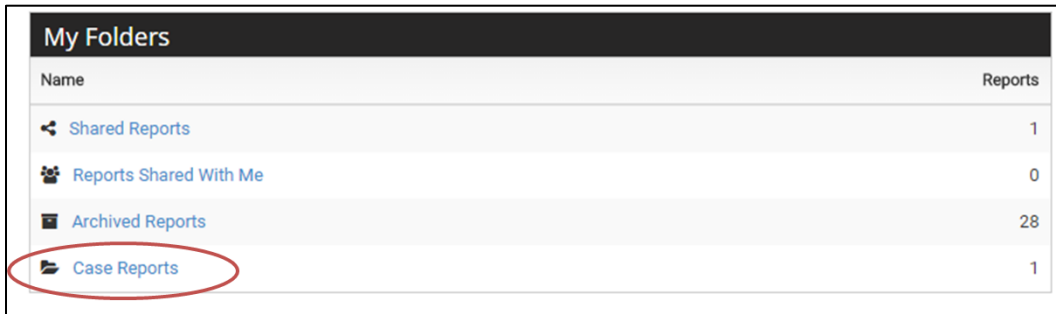
```
"IntervalInMin": "0",  
"promise": "http://www.google.com",  
"error": "http://www.google.com"  
}
```

2.2 Upgrade Scraawl User Interface (UI)

We have enhanced the responsive UI and added features to Scraawl UI, which are detailed below.

2.2.1 Bookmarking and Case Report Capability

We have added a capability to bookmark a set of tweets and export them into a case folder, which is, for all practical purposes, another Scraawl report. The case reports are added part of “My Folders” in the “My Reports” screen and are shown along with “Shared Reports,” “Reports Shared with me,” and “Archived Reports” (see Figure 2).



Name	Reports
 Shared Reports	1
 Reports Shared With Me	0
 Archived Reports	28
 Case Reports	1

Figure 2: “Case Reports” under “My Folders.”

A user can bookmark a tweet or a set of tweets and then export these bookmarked tweets into a Case Report. The most convenient way to bookmark all tweets is to make a search under “Raw Data” and use “Bookmark ALL matching” under the “Bookmark” dropdown menu. Once a tweet is bookmarked, a bookmark symbol is shown to the right of the tweet before the date column.

The bookmarked tweets can then be exported into a “Case Report” by using the “Export Bookmarked” menu item under the “Actions” dropdown menu (see Figure 3).

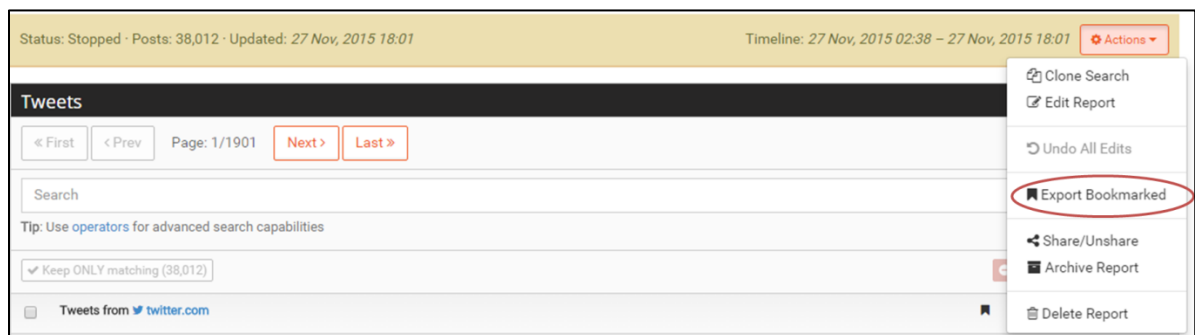


Figure 3: “Export Bookmarked” menu item.

2.2.2 Enhancement of User Profiles

The user profile screen has been enhanced with additions to limits for data access and managed access to analytics. Specific submenus are as follows.

- **Basic Information:** Basic user information such as name and e-mail is shown here. Users can also change their password and change their time zone. If a user belongs to a group then that information is shown here as well.
- **Linked Accounts:** In addition to linking Twitter accounts, users can now link their Instagram accounts. This is encouraged because users can use their token (as opposed to shared IAI token) when using public APIs to collect data.
- **Report Limits:** Number of limits for search, active search, user, shared, archived search, archived user and case reports are shown here. These limits are based on user privileges.
- **Twitter Limits:** All limits and privileges regarding Twitter data collection is shown here.
- **Instagram Limits:** All limits and privileges regarding Instagram data collection is shown here.
- **Analytics Limits:** The availability of analytics for Twitter and Instagram is shown here. These privileges are based on user or group.
- **Manage Account:** Users can delete their account, in which case their data is deleted from IAI servers as well.

2.2.3 Translation of Words/Phrases in Searches

We enabled a translation capability directly in search screen, where users can input their words and phrases in one of the 90 languages, translate into/from them, and start searching. This capability is provided in both basic and advanced searches.

2.2.4 Usability and Navigation Improvements

We have also made improvements to the usability and navigation. We improved the communication methodology in the UI so that the front end user interface gets notified of data changes in the backend. In particular we improved the efficiency of the JavaScript libraries that communicate with the backend data and analytics services and display the result to the end user. We made the front end user interface more efficient, responsive, and dynamic for the end user. This included making the UI render faster, the data fields being updated more frequently in an efficient manner. We also focused on improving the user interface for mobile devices by better adapting to the specification of the device that is requesting the data. We also improved the visualization of the analytics results so that the exploration of the results is more intuitive and accessible.

2.4 Enhance Named Entity Recognition (NER)

We enhanced Scraawl's NER capabilities of (i) resolving a large set of names, organizations, and places in English, and (ii) expanding abbreviations, e.g., UN to United Nations. In this task, we have made the following improvements to the Scraawl NER module.

Incorporated GATE Part-of-Speech (POS) tagging: We have started using General Architecture for Text Engineering (GATE) software's [1] English POS tagger as part of Scraawl NER module. GATE is an open source software to do many common task related to Natural Language Processing. Its POS tagger [2] is a modified version of the Brill tagger, which produces a part-of-speech tag as an annotation on each word or symbol. The current NER development software uses classifies a word as an entity if and only if the word is one of the gazetteers and its POS tag is Noun.

Enhanced the gazetteers and incorporated multi-lingual name matching capabilities: We have enhanced the gazetteers by including open source JRC-Names dictionary [3], and NGA Geographical Names Database [4]. With the compiled dictionaries, we have added the capability of resolving (i) 1.18+ million persons, (ii) 6700+ organizations, and (iii) virtually every town/city in both English, in their native languages, and additional common languages.

2.5 Incorporate Geo-reference Analytics

We developed analytics that evaluate the text of the tweets and user profiles to extract and identify location information, and geo reference the location on a map. We identified locations through NER and displayed geo-coordinates (longitudes and latitudes) for locations recognized by NER. We then provided the capability to locate locations mentioned by tweets on a map.

We also added MapBox, which is a mapping and searching service that is built on vector maps and rendered in real-time, and made it available as part of Scraawl's geospatial analytics. A screenshot of MapBox screen is shown in Figure 5, where geo-coded, geo-referenced, and geo-profiled tweets are shown. In addition, click/drag/filter (selecting the square icon and by drawing rectangles with mouse), zoom in/out (either with the +/- buttons or mouse wheel), home zooming, and editing/deleting selected regions are available. In addition, using the "magnifying glass" icon, a user can search locations/places, and the map automatically zooms into the selected region.

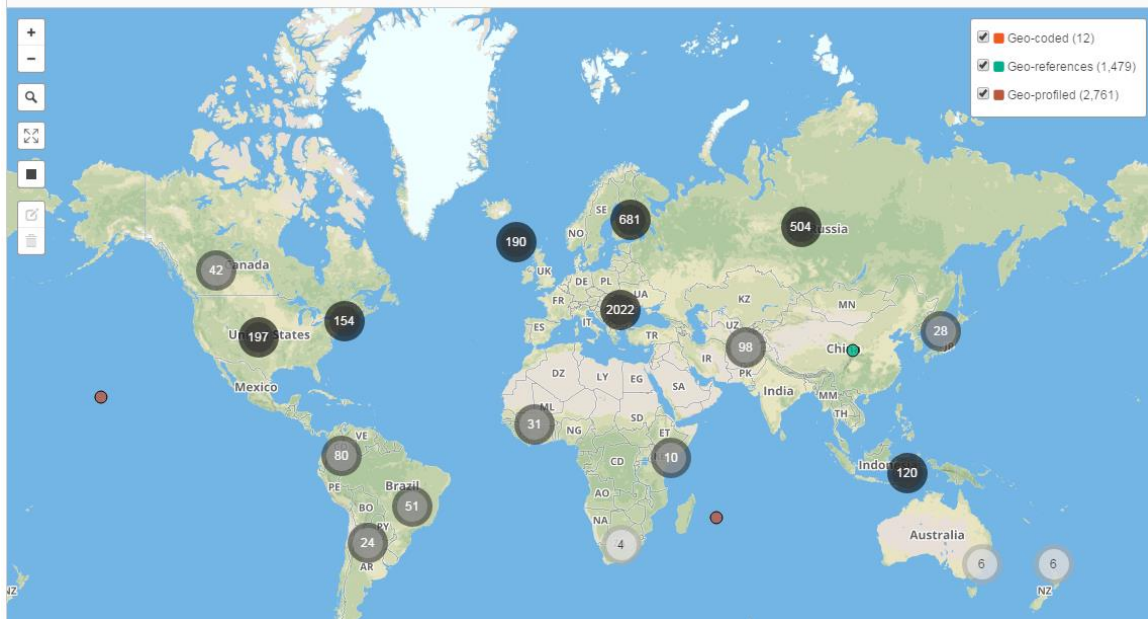


Figure 5: MapBox screen.

3 Conclusion

With the algorithms implemented and improvements made to both software and UI, users will have an enhanced information tailoring capability using Scraawl. This will enable users to better answer questions about key actors, topics, events, communities, sentiments, discourses etc. using social media. In addition, users will have more capabilities to create their own workflows by incorporating searching, filtering, and analytics in any order and as many times as possible. For example, with temporal analytics, users can first find influential users in two different time periods, and then filter into the desired time period. The user can run geo-referencing only on the filtered time period effectively increasing the signal to noise ratio.

4 References

- [1] GATE, <https://gate.ac.uk>.
- [2] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, October 2000.
- [3] JRC-Names, <https://ec.europa.eu/jrc/en/language-technologies/jrc-names>.
- [4] NGA Geographic Names Database, <https://www.nga.mil/ProductsServices/GeographicNames/Pages/default.aspx>.