

⋮

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

# Intelligent Automation Incorporated

## Enhancements for a Dynamic Data Warehousing and Mining System for Large-scale HSCB Data

### Monthly Report No. 4

Reporting Period: June 21, 2016 – July 20, 2016

Contract No. N00014-16-P-3014

*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.

## Monthly Report No. 4

# Enhancements for a Dynamic Data Warehousing and Mining System Large-Scale HSCB Data

Submitted in accordance with requirements of  
Contract #N00014-16-P-3014

Performance period: June 21, 2016 to July 20, 2016  
(PI: Dr. Onur Savas, 301.294.4241, osavas@i-a-i.com)

<b>1</b>	<b>Work Performed within This Reporting Period</b>	<b>2</b>
<b>1.1</b>	<b>Development of YouTube Analytics and UI</b>	<b>2</b>
<b>1.1.1</b>	<b>YouTube Top K Statistics Computation and View</b>	<b>2</b>
<b>1.1.2</b>	<b>YouTube Timeline View</b>	<b>3</b>
<b>1.1.3</b>	<b>YouTube Media Gallery View</b>	<b>3</b>
<b>1.1.4</b>	<b>YouTube Named Entity Recognition (NER) View</b>	<b>4</b>
<b>1.1.5</b>	<b>YouTube Map View</b>	<b>4</b>
<b>2</b>	<b>Current Problems</b>	<b>5</b>
<b>3</b>	<b>Work to be Performed in the Next Reporting Period</b>	<b>5</b>

## 1 Work Performed within This Reporting Period

In this reporting period, we performed the following tasks.

- **Developed YouTube Analytics Capabilities.** We have developed computation of YouTube basic statistics including Top Videos, Top Users, Top Words, and Top Languages, and also applied NER to the text associated with YouTube posts. We have also developed UI for interactive display, and designed map and media gallery views.
- **Released Scraawl 1.16.**

### 1.1 Development of YouTube Analytics and UI

#### 1.1.1 YouTube Top K Statistics Computation and View

We have developed a fast and reliable computation of Top K statistics for YouTube videos. In particular, we compute Top Videos, Top Users, Top Words, and Top Languages. In all cases, “Top” refers to higher count. Similar to other data sources, we have also implemented a UI to show and interact with these statistics. Figure 1 shows a representative view of the UI. Top Videos, Top Users, and Top Words are shown as lists while Top Languages are shown as a pie chart. Every Top K Statistics also include a

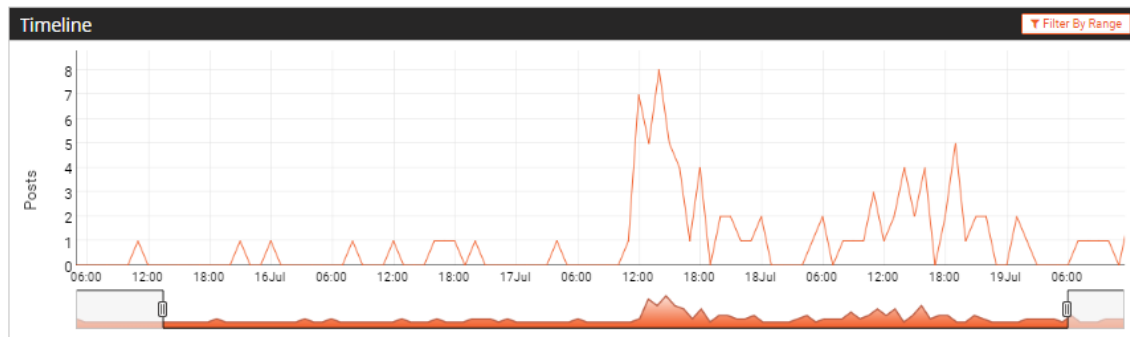
drill-down menu, where users can access using the “Details” button. In the drill-down menu, users can see Top 50 statistics and other relevant metrics. Also, users can filter to include or exclude the relevant posts.



**Figure 1: Representative Top K Statistics for YouTube Videos.**

### 1.1.2 YouTube Timeline View

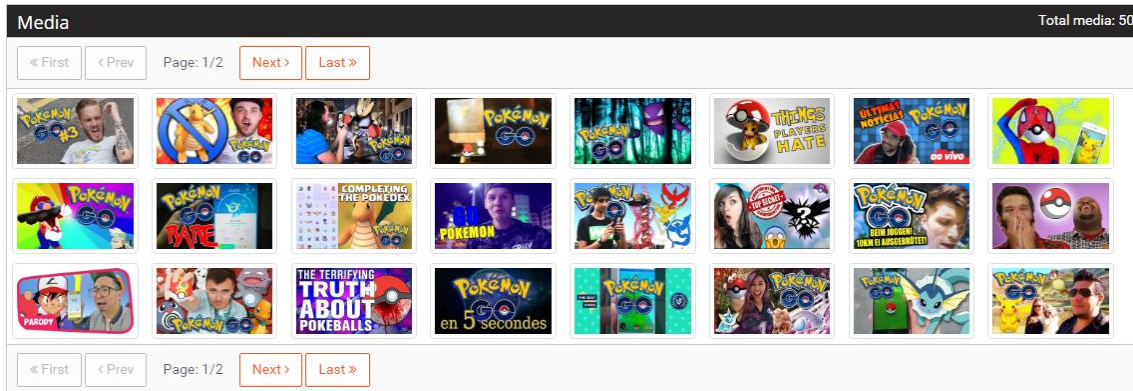
Similar to the other data feeds, we have developed a timeline view that shows the number of posts. The timeline view is interactive and the selected portion can be filtered using “Filter By Range” as shown in Figure 2.



**Figure 2: Representative Timeline View.**

### 1.1.3 YouTube Media Gallery View

We show the snapshots of the Top 50 videos in a grid under “Media Gallery” view. The video snapshots are clickable and each click directs the user to a page that has statistics about video. A representative media gallery view is shown in Figure 3.



**Figure 3: Representative Media Gallery View.**

### 1.1.4 YouTube Named Entity Recognition (NER) View

We perform Scraawl’s NER algorithm to the text associated with YouTube post, which classifies the named entities into organizations, locations, and persons. Scraawl’s NER also performs abbreviation extensions, e.g., UN is mapped to “United Nations.” A representative NER view along with other statistics is shown in Figure 4.

Status: Completed · Posts: 100 · Updated: 20 Jul, 2016 10:20 Timeline: 14 Jul, 2016 19:46 – 19 Jul, 2016 18:00 [Actions](#)

**Top Named Entities**

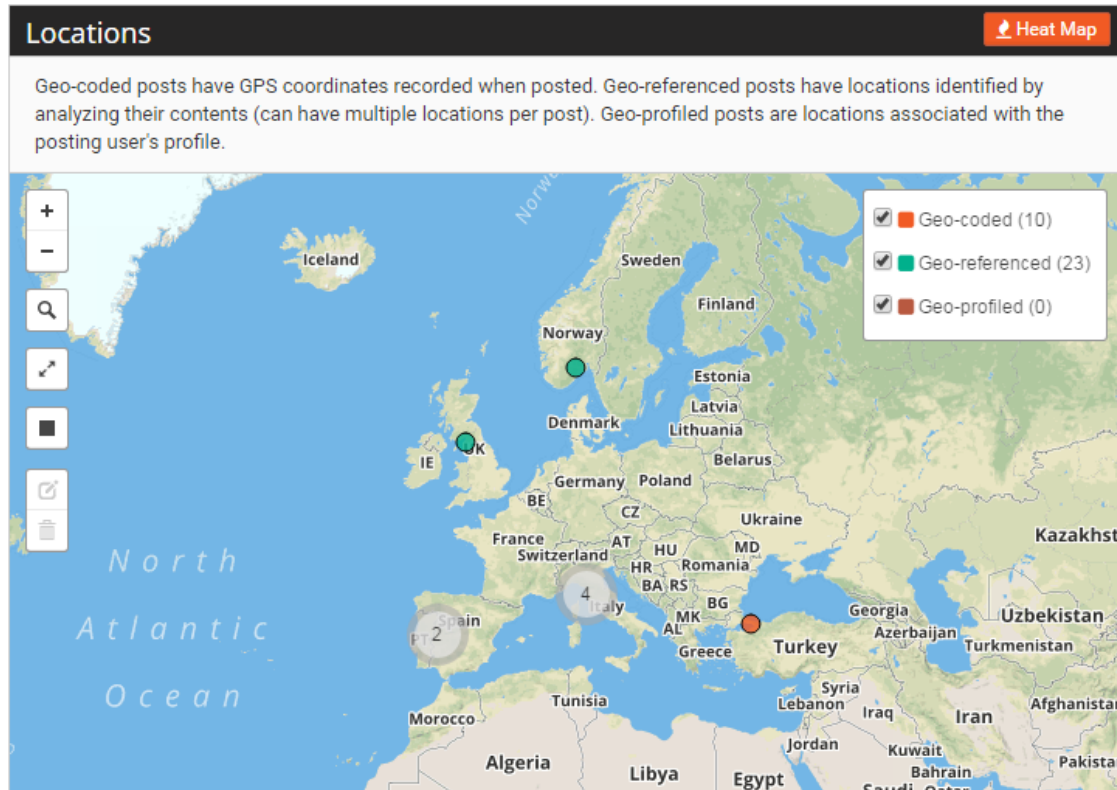
Page: 1/8 Show **All**

Entity	Type	Entity count	% of entities	Post count	% of posts	User count	% of users
<a href="#">United Nations</a>	Organization	20	50.0%	10	10.0%	13	14.4%
<a href="#">Society Of Indexers</a>	Organization	6	15.0%	5	5.0%	5	5.6%
<a href="#">Massachusetts Institute Of Technology</a>	Organization	4	10.0%	3	3.0%	3	3.3%
<a href="#">United States</a>	Location	4	17.4%	4	4.0%	4	4.4%
<a href="#">Le Jeu</a>	Person	4	4.2%	3	3.0%	3	3.3%

**Figure 4: Representative NER View.**

### 1.1.5 YouTube Map View

We use Scraawl’s Location Map and Heat Map view to display geolocation information about YouTube videos. In particular, we show geo-coded and geo-referenced YouTube posts. Geo-coded posts are those which have GPS information on them while geo-referenced posts are those which mention location information. A representative Map view is shown in Figure 5.



**Figure 5: Representative Map View.**

## 2 Current Problems

None.

## 3 Work to be Performed in the Next Reporting Period

In the next report period, we will focus on the following tasks:

- We will develop an API for news feed collection.
- We will deliver Scraawl 1.17.