15400 Calhoun Drive, Suite 190
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

Progress Report No. 2

Reporting Period: April 22, 2016 – May 21, 2016

Contract No. N00014-16-P-3014

Sponsored by
ONR, Arlington VA
C⊙TR/TP⊙C: Dr. Rebecca Goolsby

Prepared by
Onur Savas, Ph.D.

DISTRIBUTION A

Approved for public release; distribution is unlimited.

Progress Report No. 2

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: April 22, 2016 to May 21, 2016
(PI: Dr. Onur Savas, 301.294.4241, osavas@i-a-i.com)

1 Work Performed within This Reporting Period

In this reporting period, we performed the following tasks.

- **Enhanced the capabilities for top K user and hashtag subgraph querying and visualization.** We have enhanced subgraph matching queries to retrieve top K users/hashtags and their neighborhoods, and further improved Scraawl graph visualization to visualize the results.

- **Enhanced the capabilities for friendship graph querying and visualization.** We have implemented a service to retrieve friends and followers of an arbitrary set of users, and further improved Scraawl graph visualization to visualize the results.

1.1 Top K User and Hashtag Subgraph Querying and Visualization

1.1.1 Top K Query Implementation

As reported in Progress Report No. 1, we consider the Twitter interaction graph modeled as an undirected graph $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of edges. For a collection of tweets $\{\tau(\theta) | \ \theta \in \mathbb{Z}^+\}$, where each tweet $\tau(.)$ can be uniquely identified by its unique *tweet ID* $\theta \in \mathbb{Z}^+$, let $\tau(\theta)$ be tweeted by user $u_{\tau(\theta)}$ and let $u_{\tau(\theta)}$ have retweeted, mentioned, or replied to K_θ users $\mathcal{I}(\tau(\theta)) = \{v^1_{\tau(\theta)}, v^2_{\tau(\theta)}, \dots, v^{K_\theta}_{\tau(\theta)}\}$. Of course, if no retweets, mentions, or replies are present, then $\mathcal{I}(\tau(\theta)) = \emptyset$. We can then unambiguously specify the Twitter Interaction Graph G with $V = \{u_{\tau(\theta)} \cup \mathcal{I}(\tau(\theta)) | \ \theta \in$

$\mathbb{Z}^+\}$ and $E = \{u_{\tau(\theta)} \times \mathcal{I}(\tau(\theta)) | \theta \in \mathbb{Z}^+\} = \{(u_{\tau(\theta)}, v^1_{\tau(\theta)}), (u_{\tau(\theta)}, v^2_{\tau(\theta)}), \dots, (u_{\tau(\theta)}, v^K_{\tau(\theta)}) | \theta \in \mathbb{Z}^+\}$.

Our interest lies in the querying of this graph. We first define a generic query operator over the graph that will return a subset of the graph, i.e., $Q(.): G \mapsto G$. In this reporting period, we have designed and implemented top[1] K neighborhood queries. Formally, we define the top K neigborhood query as

$$Q\left(\bigcup_i v_i\right) = G(V', E') \; where \; V' = \left(\bigcup_i v_i \cup N(v_i)\right) and \; s.t. \; E' = \bigcup_{i,j} (v_i, v_j(i)), \; where \; v_j(i) \in N(v_i)$$

for a set of "top" $\bigcup_i v_i$ vertices and $N(v_i)$ denotes the immediate neighborhood of v_i. Recall from Report 1 that we have designed and implemented a graph querying system using the graph database OrientDB. We have added the above querying capability implemented using OrientDB and made it available through Scraawl.

1.1.2 Top K query

In particular, we have implemented top K neighborhood queries for the following top K.

Top K Users: The top tweeting users.

Top K Hashtags: The top tweeted hashtags.

Top Connected Users: Top users ranked by degree in the Social Graph.

Top Connected Hashtags: Top hashtags ranked by degree in the Social Graph.

We have also enhanced the visualization for top K queries. We have improved the visualization by (i) automatically adjusting zooming and translating so that the graph fits into the visible screen, (ii) incorporating web workers, i.e., background computation threads, so that graph layout is computed without interfering with user's interaction with

[1] As defined according to context as shown in Section 1.1.2.

the page, and (iii) adjusted the parameters of the graph layout so that there is less cluttering. Figure 1 is a representative visualization for a top 10 query as integrated into Scraawl.

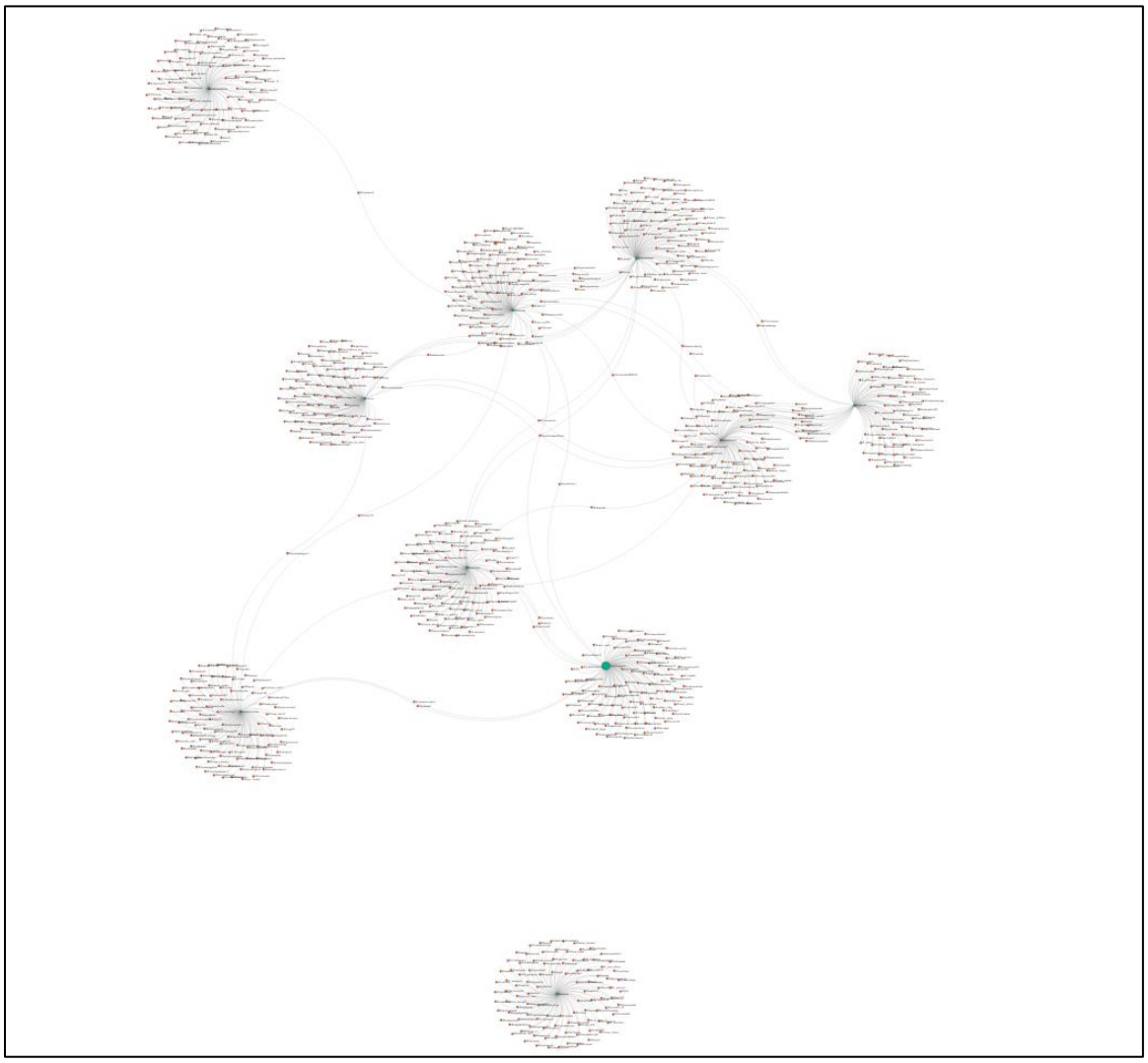

Figure 1: Representative visualization of Top 10 users.

1.1.3 Graph Visualization for Friendship Graph

We have also designed and implemented a service that retrieves the friends and followers of a given user and visualizes it in Scraawl. Figure 2 is a representative graph visualization of the friendship graph of two users. The "queried users" are depicted as orange, the common friends/followers are shown as green, and the rest of the users are drawn in gray. Unlike the Social graph, the friendship graph is directed hence the direction of links are represented in the visualization. The relationship "being a friend" and "being a follower" are distinguished by using orange and gray colors.

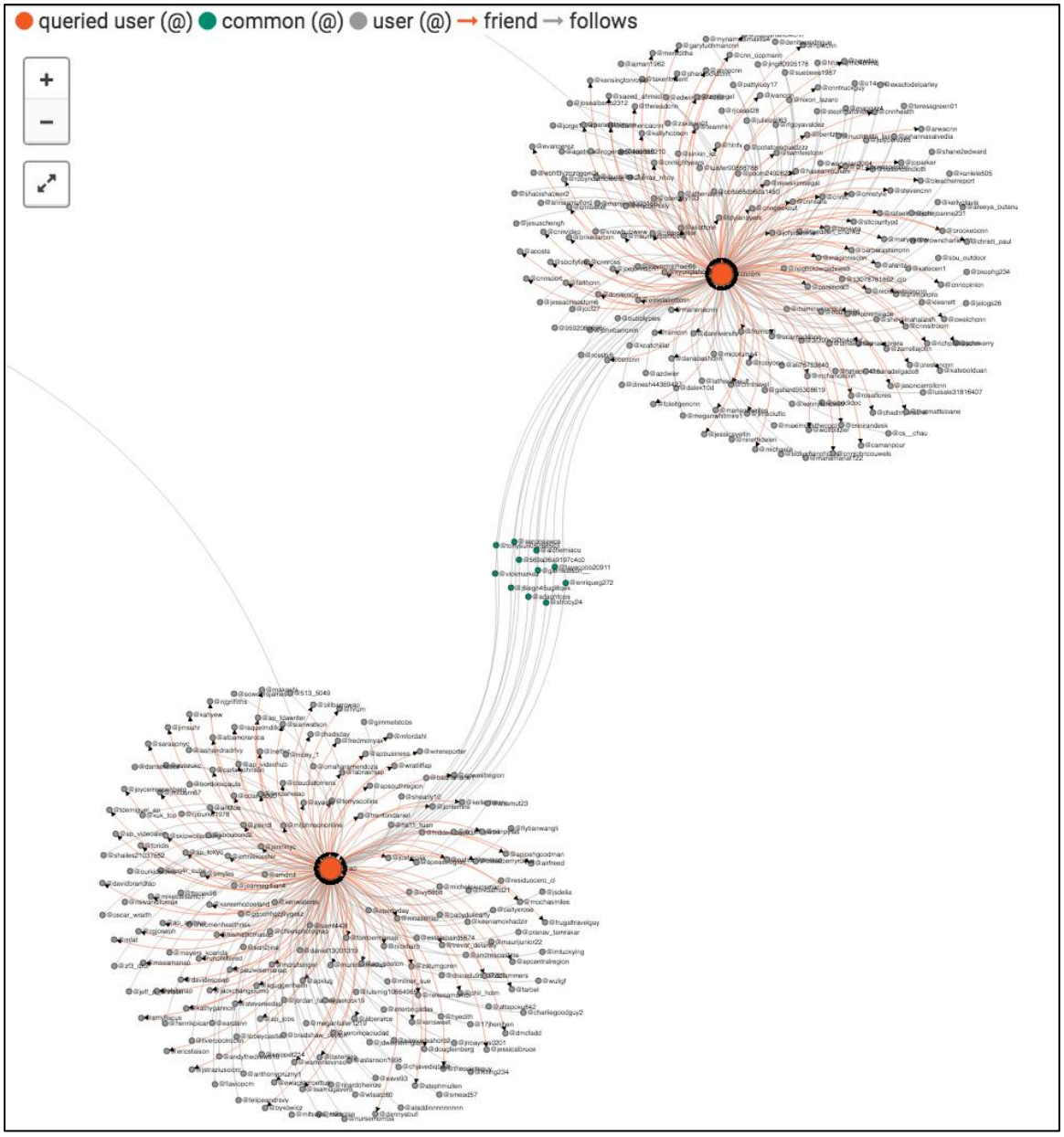

Figure 2: Representative Friendship Graph.

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 start executing Task 2.
- We will deliver Scraawl 1.15.0.