OJPHI

ISDS 2013 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

ORBiT – The Oak Ridge Biosurveillance Toolkit

Laura Pullum* and Arvind Ramanathan

Computational Data Analytics, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Objective

Our objective is to provide 1) forecasting and early warning, and 2) an extensible data analytics platform for biosurveillance by enabling the use of traditional and non-traditional datasets, consisting of heterogeneous data types and modalities.

Introduction

ORBiT is implemented as a distributed analytic platform consisting of a software stack atop of Hadoop and makes use of Titan, a distributed graph database as a backend for data storage. Data from each of the traditional and non-traditional sources are hosted as a massive linked structure, with extensible interfaces for each stream. The data from the linked structure is interfaced with streaming and graph-data analytic modules. The outputs from the analytic modules are interfaced with visualization tools that enable analysts to detect spatial and temporal patterns/correlations across multiple data sources.

Methods

The ORBiT data collection interface incorporates a collection of tools to handle multiple diverse/disparate, potentially high volume data streams including: (a) social media sites such as Twitter; (b) climatological data; (c) traditional structured data records of emergency room visits and prescription sales that include data regarding physician issued prescriptions for patients, and (d) extensions to accommodate other non-traditional multimedia data such as images from Instagram. Further, the collection interface can interact with existing reporting tools for biosurveillance and, with minimal extensions, it is possible to integrate data from other data streams. A secondary aspect of the data collection interface is the ability to stream the data to the Titan distributed graph database [1] to enable efficient storage and retrieval of large-scale datasets.

The core of ORBiT's analytic components consists of a powerful NLP (natural language processing) toolkit that can effectively build a statistically relevant vocabulary or bag-of-words model to process text-related data-streams such as Twitter [2]. The NLP tools build statistically principled models of disease associated terms from existing ontologies, PubMed literature and other textual data-sources.

Once the data-streams are filtered using NLP, we used higher-order statistical tools to track/tag events of interest using multi-scale temporal windows that can be specified by the analyst/end-user [3]. Statistical feature-sets extracted from the filtered data allow one to quickly identify a baseline and tag events as outliers from these baselines. In order to track correlations across multiple data-streams and make predictions, we include several linear, non-linear and hybrid statistical inference tools that achieve good performance in terms of an applied loss function within ORBiT [4].

The analysis modules closely interface with the visual front-end, which consists of a front-end that allows the analysts (or end-users) to interact with and provide feedback to the data analytics components in the toolkit. The front-end allows the end-user to: visualize data-streams, identify and tag potentially interesting leads (from different data sources), and visualize anomalous behaviors and spatio-temporal correlations across multiple data-streams.

Conclusions

We have described ORBiT, which emphasizes our novel statistical and machine learning tools to analyze potentially large datasets and provide a visual analytics front-end for biosurveillance related tasks.

Keywords

Heterogeneous data; Extensible toolkit; Visual analytics

Acknowledgments

ORNL is operated by UT-Battelle, LLC, for the U.S. Dept of Energy under contract DE-AC05-00OR22725. The U.S. Government (USG) retains and the publisher, by accepting the article for publication, acknowledges that the USG retains a non-exclusive, paid-up, irrevocable, worldwide license to publish/reproduce the published manuscript, or allow others to do so, for USG purposes.

References

[1] M. Rodriguez and J. Shinavier. Exposing multi-relational networks to single-relational network analysis algorithms. J Infomet, 4(1):29–41, 2009.

[2] W. Chapman. Natural language processing biosurveillance. Handbook of Biosurveillance. Elsevier Inc., 2005.

[3] A. Ramanathan, A.J. Savol, P.K. Agarwal, and C.S. Chennubhotla. Event detection and sub-state discovery from bio-molecular simulations using higher-order statistics: Appli-cation to enzyme adenylate kinase. Proteins: Structure, Function, and Bioinformatics, 80(11):2536–2551, 2012.

[4] C. Chennubhotla and A. Jepson. Eigencuts: Half-lives of eigenflows for spectral clustering. In Advances in Neural Information Processing Systems (NIPS), pp 689–696, 2003.

*Laura Pullum
E-mail: pullumll@ornl.gov