The EMBERS Architecture for Streaming Predictive Analytics

Andy Doyle*, Graham Katz*, Kristen Summers*, Chris Ackermann*, Ilya Zavorin*, Zunsik Lim*,
Sathappan Muthiah†, Liang Zhao†, Chang-Tien Lu†, Patrick Butler†, Rupinder Paul Khandpur†,
Youssef Fayed‡, Naren Ramakrishnan†
**CACI Inc., Lanham, MD 20706*
†*Virginia Tech, Blacksburg, VA 24061*
‡*BASIS Technology, Herndon, VA 20171*

Abstract—**Developed under the IARPA Open Source Initiative program, EMBERS (Early Model Based Event Recognition using Surrogates) is a large-scale big-data analytics system for forecasting significant societal events, such as civil unrest incidents and disease outbreaks on the basis of continuous, automated analysis of large volumes of publicly available data. It has been operational since November of 2012, delivering approximately 50 predictions each day. EMBERS is built on a streaming, scalable, share-nothing architecture and is deployed on Amazon Web Services (AWS).**

I. Introduction

Anticipatory intelligence is an important frontier of 'big data' research, wherein myriad data streams are fused together to generate predictions of critical societal events such as civil unrest incidents, disease outbreaks and election outcomes in order to identify threats and aid decision making for national security, law enforcement, and intelligence missions. EMBERS [1] is an anticipatory intelligence system supported by the Intelligence Advanced Research Project Activity (IARPA) OSI (Open Source Indicators) program which produces detailed forecasts of critical societal events in Latin America and the Middle East and North Africa on the basis of publicly available (open-source) data.

EMBERS has been operational since November 2012 and has delivered over 16,000 fine-grained event-predictions on the basis of a wide range of data, from high-volume, high-velocity, noisy and unstructured social media such as Twitter to lower-volume, higher-quality structured data sources, such as OpenTable reservation cancellations or humidity measurements. EMBERS predictions, which specify the date, location and type of event, have been evaluated retrospectively against ground truth data created by human analysts. The quality of the predictions and their accuracy have steadily improved as the system has been developed. Detailed description of the predictive models and the evaluation, both methods and metrics, is to be found elsewhere [1]. In this paper we focus on providing a comprehensive description of the architecture of the EMBERS system.

II. Architecture

A. Background

EMBERS was developed by a dispersed team of eight research universities and industry partners, with diverse expertise in computer science, machine learning, disease modeling, social science, linguistic processing, and systems integration. This distributed team required a loosely-coupled functional architecture which would allow team members to develop system components independently, but allow for rapid and simple integration. Additionally, the vast majority of inputs to the system were anticipated to be continuously ingested streams of data needing near-real-time processing, so a streaming architecture was deemed appropriate.

B. Software Architecture

In order to meet these needs, the EMBERS system was designed to be made up of a large number of simple independent components wired together in a pipes and filters architecture [2]. A simple simple message-passing design was adopted for data exchange, using JSON messages transmitted over ZeroMQ sockets. The component programs of the EMBERS system (described below) implement multiple independent data transduction steps, from data ingest, normalization and indexing to entity extraction, geo-coding, and keyword counting. Architecturally EMBERS is closely related to the UIMA [3] model for pipeline processing of content and similar systems such as Storm [4] and IBM InfoSphere Streams [5]. EMBERS differs from these approaches in that it minimizes required infrastructure and uses open interfaces for data.

C. Operational Infrastructure

EMBERS is deployed on the commercial AWS (Amazon Web Services) cloud infrastructure as a cluster. The cluster is a collection of nodes (EC2 virtual machines) which host a collection of services (individual programs) that read from or write to streams (ZeroMQ queues or S3 files). Services are distributed in the cluster according to their resource needs. The layout of the cluster, including the name, number and type of VM; the configuration of each service; and the input and output queues for a service is specified in a configuration file making cluster setup and deployment trivial. Communications among programs is network-transparent, allowing services to be moved between nodes with no impact on the queue topology.

The current cluster configuration consists of 12 machines with a total of 21 virtual CPUs and 75G of RAM. The EMBERS system comprises approximately 100 individual

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

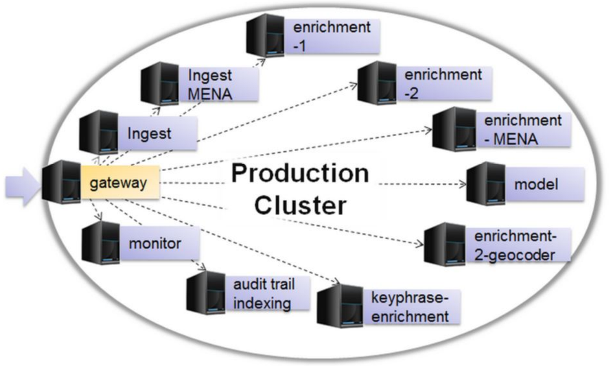


Figure 1: Layout of the EMBERS Cluster.

programs. Processing, dataflow and machine-state parameters are monitored with Ganglia and Nagios.

III. Processing Pipeline

The EMBERS system consists of three major processing components: Data Ingest, Message Enrichment and Analytic Modeling (Fig. 2). In addition to the the primary mode of streaming analysis, the architecture also supports batch processing and database-based cached storage for data aggregation and persistence. Data is archived from key points in the processing pipeline, making it easy to replay the system from historical data. Additionally, messages are indexed so that derivation chains and audit-trails can be quickly computed.

A. Data ingest

Approximately a dozen different data-source types are ingested into the EMBERS system, ranging from weekly government reports to feeds of Twitter posts. Other data sources include curated data such as HealthMap [6], [7] alerts and Google Flu Trends data as well as RSS newsfeeds and blogs. Most EMBERS data sources are text-based but a significant number, such as Google Flu Trends are numeric and some are more complex sources such as GDAS (Global Data Assimilation System [8]), which provides climate information derived from satellite data. EMBERS also ingest some "derived" data, such as the content of URLs referenced in a Twitter posting. Currently the system ingests about 19.2G of data per day, which corresponds to about 4.6M messages a day in total.

B. Enrichment

Textual data passes through a series of text-analytic enrichment processes that feed downstream processing. Basic processing such as tokenization, part of speech tagging and lemmatization, as well as named entity extraction (NEE) is performed by Basis Technology's RLP and REX products. This preprocessing serves as input to subsequent deeper semantic analysis as well as further downstream processing.

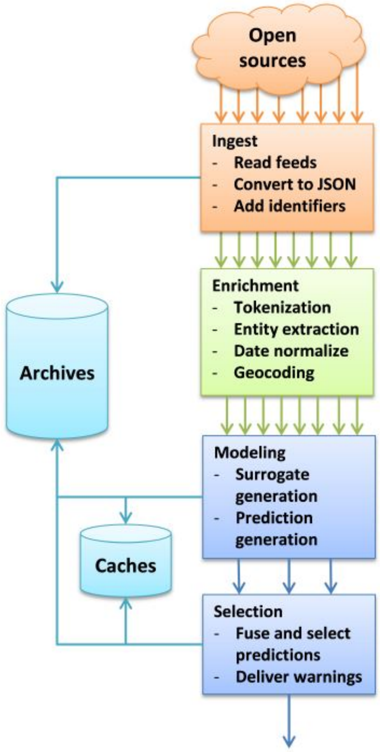
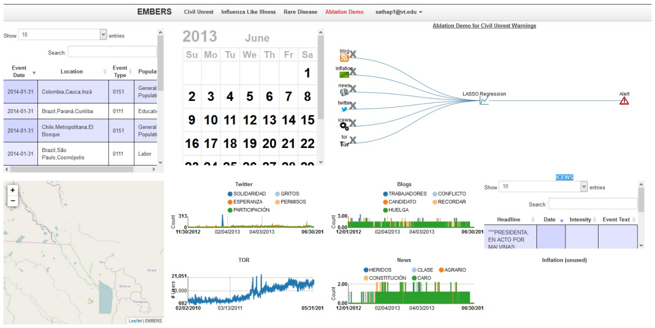


Figure 2: EMBERS System Components.

EMBERS performs three targeted semantic analysis tasks: date normalization, geo-coding and sentiment analysis. These provide important inputs to the predictive models. The EMBERS date normalization module (based on TIMEN package [9], English, Spanish and Portuguese and on the HeidelTime package [10] for Arabic) determines for each date expression (e.g. *Friday*, *mañana*) in the text, the most likely date it refers to, based on both textual cues and metadata. The geocoding module identifies at the granularity of the city the geographical focus of the input text. For microblog postings, a simple set of rules for exploiting geo-spatial metadata (such as mobile device tags and user profile information) is leveraged to geo-code the posting. For longer texts, such news articles or blog posts, a more complex system based on probabilistic soft logic (PSL [11]), which identifies the likely geo-focus of the text on the basis of the location entities extracted from the text. The sentiment analysis system uses the ANEW [12] lexicon (and translations) to derive a three dimensional sentiment score (VALENCE, DOMINANCE, AROUSAL).

C. Analytic Modeling

Predictions about future events come from a set of models that ingest data feeds to produce specific predictions about the data, location, involved population and type of event. These models, which operate independently of each other, use a variety of underlying algorithms such as logis-



Event Date	Location	Event Type	Popula
2014-01-31	Colombia,Cauca,Inzá	0151	General Populati
2014-01-31	Brazil,Paraná,Curitiba	0111	Educati
2014-01-31	Chile,Metropolitana,El Bosque	0151	General Populati
2014-01-31	Brazil,São Paulo,Cosmópolis	0111	Labor

Headline	Date	Intensity	Event Text
"""PRESIDENTA, EN ACTO POR MALVINAS:			

Figure 3: EMBERS visualizer. Top right shows contributing data sources. Each source can be selectively removed to show the effect on the prediction generated.

tic regression, geo-temporal clustering and keyword-based counting to make their predictions (individual models are described in [1]). In keeping with the parallel architecture, different models may apply different techniques to the same data feeds to produce different predictions.

In a final processing stage, the the set of predictions generated by the ensemble of models is evaluated and an optimal "fused" set of predictions selected for delivery. This fusion step incorporates information about past model performance as well as baseline data about observed past events, prediction density and expected event-count.

IV. Visualization Tool

EMBERS tracks dataflow through the system, recording the processing pipeline associated with message. The visualization tool allows users to visualize this pipeline for each prediction made. This "audit-trail" shows the contributions of data sources to that prediction and to run rudimentary ablation tests (Fig. 3).

V. Conclusion

EMBERS presents a working example of a big data streaming architecture designed to process large volumes of social media data and produce predictions using a variety of modeling approaches. While EMBERS is primarily a research platform, the operational experience with the system indicates that the streaming message-based architecture is a viable approach to big data system implementation and that it performs well in some real world scenarios that tested its ability to detect large atypical events.

Acknowledgment

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz *et al.*, "'Beating the news' with EMBERS: Forecasting Civil Unrest using Open Source Indicators," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[2] F. Bushmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal, *Pattern Oriented Software Architecture: A System of Patterns*. John Wiley & Sons, 1996.

[3] D. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering*, 2004.

[4] "Storm: Distributed and fault-tolerant realtime computation." [Online]. Available: http://storm.incubator.apache.org/documentation/Rationale.html

[5] B. Gedik and H. Andrade, "A model-based framework for building extensible, high performance stream processing middleware and programming language for IBM InfoSphere Streams," *Software: Practice and Experience*, 2012.

[6] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project," *PLoS medicine*, 2008.

[7] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports," *Journal of the American Medical Informatics Association*, 2008.

[8] M. Rodell, P. Houser, U. e. a. Jambor, J. Gottschalck, K. Mitchell, C. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich *et al.*, "The Global Land Data Assimilation System," *Bulletin of the American Meteorological Society*, 2004.

[9] H. Llorens, L. Derczynski, R. J. Gaizauskas, and E. Saquete, "TIMEN: An Open Temporal Expression Normalisation Resource." in *Proceedings of Language Resources and evaluation*, ser. LREC, 2012.

[10] J. Strötgen and M. Gertz, "Heideltime: High quality rule-based extraction and normalization of temporal expressions," in *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2010, pp. 321–324.

[11] M. Brocheler, L. Mihalkova, and L. Getoor, "Probabilistic Similarity Logic," *arXiv preprint arXiv:1203.3469*, 2012.

[12] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Citeseer, Tech. Rep., 1999.