Experiences in Building an Enterprise Data Analytics

Ashot N. Harutyunyan, Arnak V. Poghosyan, and
Naira M. Grigoryan
VMware
Email: {aharutyunyan;apoghosyan,ngrigoryan}@vmware.com

Abstract

The Information Age made data easily accessible and omnipresent, currently with features of big volume, high velocity, and large variety, never seen before. For sciences, that is an unbelievable opportunity to explain the world better. Moreover, in the post-Information Age, businesses make any attempt to collect data and deeply benefit from it to achieve highly innovative technologies in terms of automation, performance, and efficiency. We share our experiences in building an enterprise data analytics for managing modern cloud computing infrastructures, as well as make parallels with information theory problems.

1 Introduction

In the era of Big Data, technologies are made of data. They increasingly tend to design smart applications with data-driven intelligence to profoundly benefit from the advantage of measured/monitored storm of data overwhelming human capabilities to process it and retrieve actionable knowledge. Therefore, industries are vastly investing in building relevant data analytics platforms and solutions to adequately address the challenges of the new era. Those challenges imply research for novel data scientific and machine learning approaches for real-time and proactive view into various systems.

As a provider of software-defined data centers and cloud computing infrastructures through virtualization, VMware dominates in the market of management of those systems by measuring and leveraging data. To have the full and proactive view of the systems real-time in terms of performance (health), capacities (IT resources), configuration and compliance, company's cloud management solutions [1,2] monitor both sources of IT data: structured and unstructured, respectively. The goal is to effectively and efficiently manage hundreds of thousands of IT objects with millions of different parameters (metrics) over time, terabytes of logs per day, and millions of events for anomaly identification and/or prevention. Building a generic data analytics platform to target such a goal in a context-independent way is a hard problem. Our experiences in providing a real-time performance analytics for data centers are summarized in a system (see [3], [6,7], and [8]) of several modules encompassing (Fig. 1)

- *behavioral analysis* for time series data and *extreme value analysis*: typical vs. atypical behavior to judge about anomalies based on data categorization, change point and periodicity detections;
- *abnormality degree* estimation for an outlying process to measure its severity or form an anomaly event;
- *ranking of events* in terms of their impact factor and *problem root causing*;
- *principal feature analysis and event reduction*;
- *data compression*;
- *prediction of alterations* in the system (allows sparing computational resources needed to run expensive behavioral pattern extraction procedures)

and other building blocks.

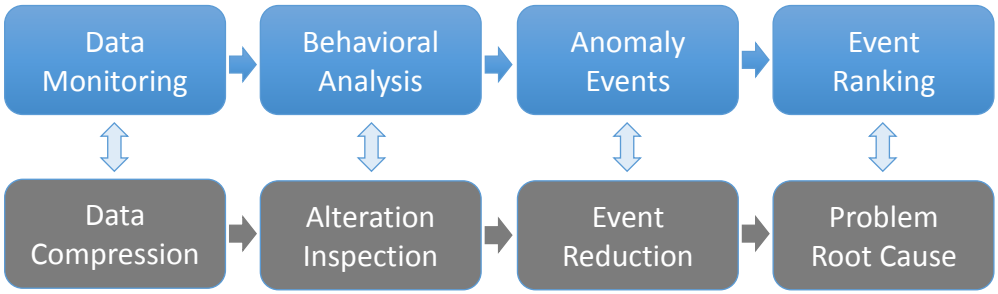

Fig. 1. Building blocks of a multi-layer analytics.

2 The Role of Information Theory

Concepts of information theory and its measures help in tackling problems we are working on, such as pattern and anomaly detection in logs [4,5], extreme value analysis applying a maximum entropy principle [8], identification of problem root causes [3], and feedback-enhanced analytics [7].

We observe inspirational parallels with information theory especially when dealing with data compression problems. One of our solutions in this domain applies correlations within data sets to compress metrics in a "meaningful" and efficient way. It parallels with Wyner-Ziv coding (Fig. 2). Specifically, the method is based on finding the principle (independent) features of a metric group (in other words, its basis) and compress the rest of dependent metrics using the corresponding linear combination coefficients. Moreover, a univariate (for a single metric) lossy compression method subject to fidelity criteria we are developing currently parallels with rate-distortion theory. The main idea behind the approach is to design a data compression model subject to the application needs such as anomaly detection, preserving relevant "interesting" patterns with high resolution and losing accuracy in other patterns.

Below we would like to enclose an example of using entropy measure to estimate confidence of users in beliefs they utilize as our data-driven recommendations and adjust those beliefs accordingly (see [7]), as well as show how it applies as a generic tool for ranking items based on user ratings as additional entertaining examples.

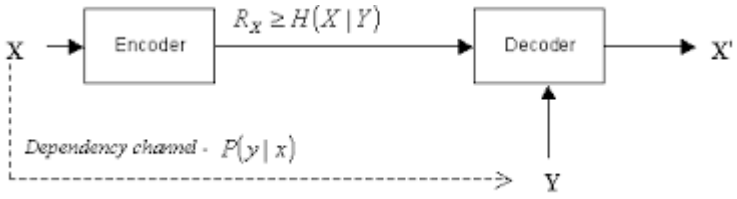

Fig. 2. Wyner-Ziv source coding.

Anomaly Detection in Logs with User Feedback. Here the basic data-agnostic analysis concerns the anomaly detection problem in log files via Dynamic Normalcy Graphs (DNG) [5]. We show how the general approach on feedback-based belief evaluation in [7] specializes into a specific solution for this correlation-based abnormality analysis and recommendation generation. We assume that the main notations and concepts of the work [7] can be used without detailed explanations. In summary, the goal is to enhance the efficiency of the DNG as a causation tool via processing of user feedback statistics on correlation breakage alarms. That can be performed if we evaluate the confidence for each correlation (belief) in DNG from that statistics and apply it in computation of abnormality degree of data stream. In some sense, it is an update of the conditional probabilities in DNG.

Formally, DNG is a collection of beliefs each representing the conditional probability from node j to i, so we denote the corresponding beliefs by $B_{i,j} = P(i \mid j)$. Let the user be asked to answer to the question if he/she is satisfied (and to what degree) by abnormality recommendation regarding a missing event type which is related to belief $B_{i,j}$. And let the users provide feedback taking value from $[0,1]$ (assume $l = 3$ and quantized feedback are from the intervals $[0,0.25)$, $[0.25,0.75)$, and $[0.75,1]$) at each time t_k when facing the breakage in $B_{i,j}$. So the feedback for $B_{i,j}$ is a series of ratings:

$$F(B_{i,j}) \equiv \{f(t_k, B_{i,j})\}_{k=1}^{K} \equiv \{f_k(B_{i,j})\}_{k=1}^{K}$$

where 1 is full satisfaction and 0 is complete dissatisfaction.

Based on $F(B_{i,j})$ we want to make a convergence evaluation in user opinion and output a confidence $C(B_{i,j})$ which can be incorporated into the existing abnormality analysis to tune its performance (an optimization of false positive alarms). This confidence will support the degree of validity of the initial belief $B_{i,j}$ and lead to a new DNG-based mismatch calculation. In other words, an updated $P'(i \mid j)$ can be obtained which is a combination of original DNG and one obtained from the feedback processing. All the postulates formulated in Section II of [7] are valid here. If there is a convergence to some degree of user satisfaction in the recommended beliefs, then the basic conditional probabilities are updated for further usage in anomaly detection with their confidences. The results in the original probabilities can increase or decrease as new beliefs about the system are incorporated. Correspondingly, their role in abnormality (mismatch) computation may change.

The notations in Section II of [7] easily apply to the beliefs $B_{i,j}$ of the DNG. According to the same reference, in case of comparably large entropy

$$H(\bar{S}(B_{i,j})) \in \left(1 - \frac{2}{3}\log_3 2, 1\right]$$

there is no convergence in feedback and hence the system has no update. The value $1-(2/3)\log_3 2$ corresponds, for instance, to the following scenario: $h_1 = 0$, $h_2 = 1/3$, $h_3 = 2/3$. If entropy is smaller than $1-(2/3)\log_3 2$ we determine the needed confidence $C(B_{i,j})$ by checking at which interval is the bias in the uncertainty. The bias is determined by the mode of the histogram $h_{\max} = \max\{h_1, h_2, h_3\}$.

Let $m(h_{\max})$ be the weighted average of the values $S(f_k(B_{i,j}))$ calculated by the technique of [7] (Section II) and corresponding to the mode $h_{\max}$. Then the confidence of $B_{i,j}$ is determined by the entropy:

$$C(B_{i,j}) = 1 - H(\bar{S}(B_{i,j}))$$

This means, for instance, that if the entropy is high, then the confidence in user feedback is zero:

$$C(B_{i,j}) = 0, \text{ if } H(\bar{S}(B_{i,j})) \geq 1 - \frac{2}{3}\log_3 2 .$$

If $C(B_{i,j})$ is a positive, then we define a feedback-based belief $B_{i,j}^f$ or $P_f(i \mid j)$ as

$$P_f(i \mid j) \equiv m(h_{\max}).$$

Now we can combine the basic data agnostic and the feedback-based beliefs on our DNG to have a new belief as

$$P'(i \mid j) = \alpha P(i \mid j) + (1-\alpha) P_f(i \mid j)$$

where

$$\alpha = 1, \text{ if } C(B_{i,j}) = 0,$$

entropy is big and

$$\alpha = C(B_{i,j}), \text{ if } C(B_{i,j}) > 0$$

entropy is low, there is a convergence.

In this way the basic DNG transforms to a new correlation structure.

To further experiment with the entropy-based confidence we applied the prototype algorithm to two public databases on consumer ratings for books and movies. In these cases, we deal with equally ranked prior recommendations for items. In other words, all the books and films included in those data sets have initial beliefs ranked to 1 until the user feedback modifies that basis assumption. So, what the program outputs on both databases is a feedback-based ranked list of items.

The next subsections demonstrate some results obtained while experimenting with the mentioned data sources, respectively. Note also that we produce two categories of item lists, one for items with converged user opinion (positive confidence) and the other with uncertain results. For the latter we do not show any rank, although it can be performed if we relax the requirement on the confidence.

In both experiments the ratings interval quantization level is $l = 3$.

Results for Book Ratings. The first experiment was performed on the book ratings data from http://www.informatik.uni-freiburg.de/~cziegler/BX/. We fed our algorithm with 600,000 ratings by 278,859 users on 271,379 books. For this data portion, we included into our analysis only the set of books that have been rated at least 30 times.

Note that the ratings timestamps are not available in the dataset. Therefore, the ratings temporal weighting is not applied in this case.

Table 1 illustrates the first three highest rank books according to our algorithm and another three famous works ("The Little Prince", "Animal Farm", and "Lolita") from 20th century that are of high rank but comparably low in overall feedback confidence. Moreover, those classics exhibit a larger uncertainty in the user feedback (perhaps also due to the fact that the first three works led to popular films).

Table 2 shows works by popular authors that exhibit severe disparity in reader opinions. This means that the uncertainty in reader's rating is so high that they are within the most disagreeable items in our analysis, although being historically significant and impactful works.

Table 1. Several books with their ranks.

Title	Author	Year	Conf.	Rank
The Return of the King	Tolkien	1955	0.81	0.99
Harry Potter and the Goblet of Fire	Rowling	2000	0.83	0.99
Charlotte's Web	White	1952	0.68	0.99
The Little Prince	de Saint-Exupéry	1943	0.61	0.98
Lolita	Nabokov	1955	0.55	0.96
Animal Farm	Orwell	1945	0.54	0.96

Table 2. Several books with zero confidences.

Title	Author	Year	Conf.
Call of the Wild	London	1903	0
Jonathan Livingston Seagull	Bach	1970	0
The Catcher in the Rye	Salinger	1951	0
The reader	Schlink	1995	0

Results for Movie Ratings. The second data set processed by our approach was from http://www.grouplens.org/node/73, namely the 100k-MovieLens rating database (20,000 ratings by 459 users on 1410 movies). We included into our analysis only the set of those movies that have been rated at least 15 times.

Table 3 displays some of the highest ranked movies (with rather distinct confidences) that are of different eras. For comparison, the well-known IMDB rating (varying from 1 to 10) for listed films is also included. Note that the entropy-based rank coincides with IMDB rating for Godfather.

Table 3. Several films with their ranks.

Title	Release	Conf.	Rank	IMDB Rating
Taxi Driver	1996	0.85	1	8.5
Three Colors: Red	1994	1	1	8
12 Angry Men	1957	0.65	1	8.9
Casablanca	1942	0.79	1	8.7
Pinocchio	1940	0.67	1	7.6
The Wizard of Oz	1939	0.46	0.99	8.2
Amadeus	1984	0.61	0.97	8.4
Godfather	1972	0.61	**0.92**	**9.2**
Schindler's List	1993	0.59	0.92	8.9

The movies in Table 4 are examples of high IMDB rated films, including Academy award-winning ones, with wildly varying audience opinions, since the users' ratings show high uncertainty.

Table 4. Films with zero confidences.

Title	Release	Conf.	IMDB Rating
Mighty Aphrodite	1995	0	7
The Lion King	1994	0	8.4
The Fifth Element	1997	0	7.6
Men in Black	1997	0	7.2
Toy Story	1995	0	8.3
Twelve Monkeys	1995	0	8.1
Seven	1995	0	8.7

References

[1] VMware vRealize Operations Manager, http://www.vmware.com/products/vrealize-operations.html.

[2] VMware vRealize Log Insight, http://www.vmware.com/products/vrealize-log-insight.html.

[3] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "An anomaly event correlation engine: Identifying root causes, bottlenecks, and black swans in IT environments," *VMware Technical Journal*, vol. 2, no. 1, pp. 35-45, 2013.

[4] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "Pattern detection in unstructured data: An experience for a virtualized IT environment," IFIP/IEEE International Symposium on Integrated Network Management, Ghent, Belgium, May 27-31, pp. 1048-1053, 2013.

[5] A.N. Harutyunyan, A.V. Poghosyan, N.M. Grigoryan, and M.A. Marvasti, "Abnormality analysis of streamed log data," Proc. *IFIP/IEEE Network Operations and Management Symposium*, May 5-9, Krakow, Poland, pp. 1-7, 2014.

[6] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "An enterprise dynamic thresholding system", Proc. USENIX 11th International Conference on Autonomic Computing, June 18-20, Philadelphia, PA, pp. 129-135, 2014.

[7] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "Ranking and updating beliefs based on user feedback: Industrial use cases," Proc. 12th *IEEE International Conference on Autonomic Computing*, July 07-10, Grenoble, France, pp. 227-203, 2015.

[8] A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "Managing cloud infrastructures by a multi-layer data analytics," Proc. *IEEE International Conference on Autonomic Computing*, July 18-22, Wuerzburg, Germany, pp. 351-356, 2016.