Scoring of system logs to prevent IT incidents

Philippe Logette; Aissa El Ouafi

Introduction

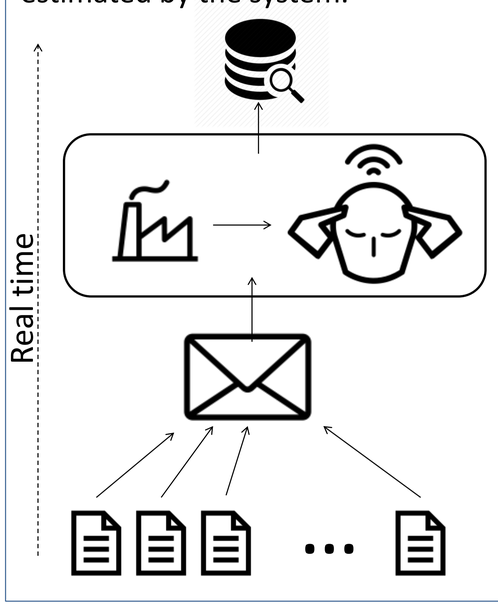
Information Systems of many companies still face incidents that could be avoided if the relevant information was analyzed at the right time. That information is often written in system logs, but traditional means, such as regular expressions based alerts, fail to highlight events accurately and timely. These patterns are laborious to set-up initially and to maintain over time when systems and applications evolve, due to the diversity of logs that do not follow a unique standard, and to the complexity of building regular expressions that would identify relevant alerts and ignore the rest – if only it is possible.

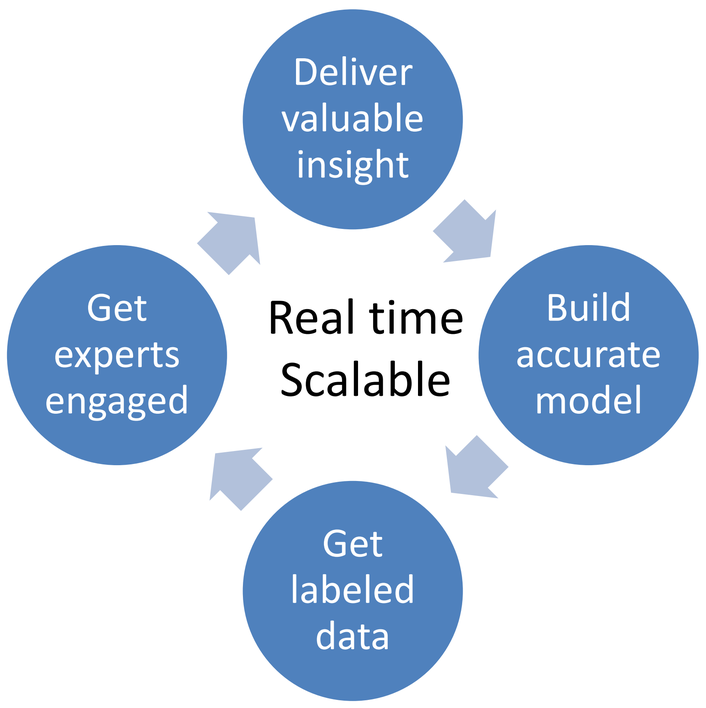
Hence, our approach is to build a machine learning system that captures experts knowledge and scores each line of logs so that logs flow can be filtered based on its estimated relevance.

Get experts engaged

Experts will only get engaged in the first place, and remain motivated to provide quality labeled data if and only if they get value out of the system and they cannot get it elsewhere.

Thus, we have built a real time pipeline to deliver all new incoming logs, scored by the relevance estimated by the system.





Challenges

Build accurate model

The model performance is the foundation to generate logs insight. As we want to be able to analyze logs as diverse as they are and will be, we have chosen to process logs as natural language and not infer any explicit structure.

Logs items and metadata are transformed through the following steps: tokenize, build bag of words with term frequency , hash features, normalize items.

The dimension of the logs items space is high, and the matrix is very sparse, and as for some natural language processing problems, we chose to work with cosine similarity. Hence, we have chosen SVM models to classify an item as relevant vs. not-relevant. A linear kernel with normalized items is used to compute cosine similarity.

Besides, we wanted to predict a probability ,not only a class. Thus, we have piped SVM classification with a logistic regression.

Get labeled data

Accurate models rely on on quality and quantity of labeled data, and require experts involvment for that. We have designed three means to capture such labels.

As users see new logs coming in live streaming, we capture direct feedback to either confirm or correct the prediction. Besides, as experts do their regular jobs and troubleshoot issues, they may encounter typical logs they do not want to miss in the future: such examples can be fed as an open feedback.

The most important labeling aims at meshing the space of logs diversity, minimizing the number of examples experts have to label. We use a semi-supervised approach to first cluster the logs based on their cosine similarity and then pick one representative to label.

Types of experts labeling

Ongoing work

This project is still at its early stage. The real time scoring platform has been set-up, as well as the machine learning pipeline to train models on the labeled data captured so far.

The logs we work on are generated by a Big Data platform with many components, at a bitrate of approximately 1 To/day.

A restricted population of experts is testing: gather feedback of all forms so that the performance and the system can be assessed and improved.

At this stage, we are tuning our parameters model as well as engineering features to get more relevant predictions. Besides, we are optimizing the semi-supervised feedback to reduce the time spent by experts on labeling data.

What's next

To improve model accuracy, we plan to explore different options. For data transformation, we will compare bag of words with n-gram dimensions, as well as entity embeddings. Algorithms other than SVM will also be considered.

The other key improvement area is the labeling of the logs diversity with minimum epxerts interaction. Our tests will combine dimensionality reduction techniques via t-SNE or autoencoders, with linear or density-based clustering algorithms. We will also take into account ambiguous feedback to propose new logs to label.
Another area of reseach we may experiment in our context is active learning.

Finally, we may introduce scoring on sequence of logs instead of single lines, as well as abnormal logs detection.

Contact

Philippe Logette
Société Générale