


MẪU BÁO CÁO CỦA MỖI HV

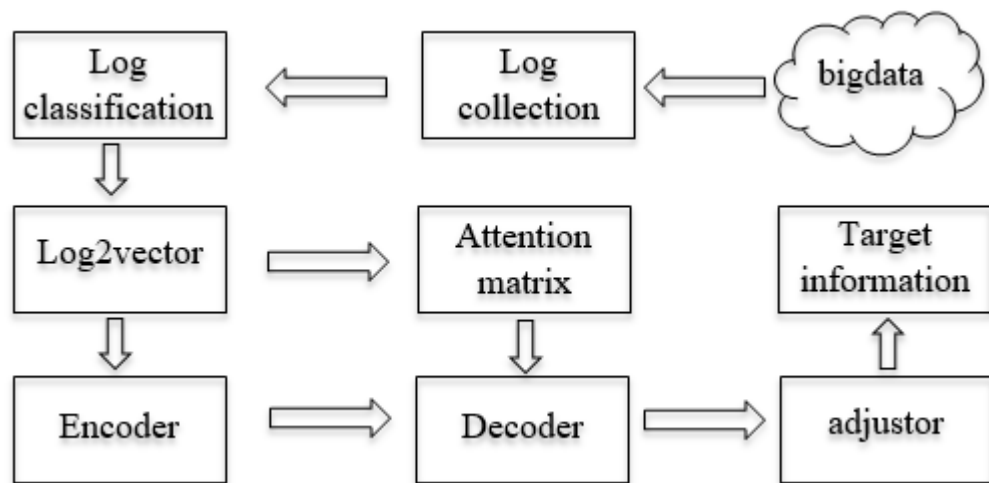
Họ và tên (IN HOA)	HOÀNG HẢI NAM
Ảnh	
Số buổi vắng	0
Bonus	19
Tên đề tài (VN)	DỰ ĐOÁN TƯƠNG LAI CHO IOT DỰA VÀO PHÂN TÍCH LOG DỮ LIỆU LỚN TRÊN MẠNG SEQ2SEQ
Tên đề tài (EN)	
Giới thiệu	<p>Sự phát triển nhanh chóng của khoa học, sự bùng nổ của công nghệ đã phát sinh một lượng lớn các thiết bị IoT. Theo báo cáo dự báo của BIIntelligence [1], sẽ có khoảng 22,5 tỷ thiết bị IoT trong năm 2021, nhiều hơn gần 16 tỷ so với năm 2017. Các thiết bị khi hoạt động sẽ có một lượng lớn dữ liệu trao đổi qua lại giữa chúng, cùng với đó là các log nhật ký phát sinh từ chính thiết bị để ghi lại các hành vi của hệ thống. Ở góc độ nhà quản trị, việc theo dõi, phân tích và dự đoán các tình huống phát sinh trong tương lai từ các log phát sinh này đã dần trở thành một công việc cần thiết để đảm bảo tốt hơn chức năng của các dịch vụ bigdata cũng như cải thiện hiệu suất và đạt được sự thông minh của hệ thống. Tuy nhiên, hệ thống Bigdata</p>

thường sử dụng nhiều thành phần để thực hiện các nhiệm vụ khác nhau. Các thành phần này có thể kể đến như Hadoop HDFS, MapReduce, Hbase, Hive, Spark, Redis, Kafka, MPP, Zookeeper, YARN và các thành phần khác. Các thành phần này thường tạo log riêng rẽ tương ứng trong quá trình chúng hoạt động. Các bản ghi log này cũng không đồng nhất, có một số độc lập và một số phụ thuộc lẫn nhau giữa các thành phần. Nhìn chung, nhật ký hệ thống bigdata có những đặc điểm đáng chú ý như đồ sộ, đa nguồn, không đồng nhất. Và đó là thách thức để phân tích và dự đoán chúng.

Do cấu trúc phức tạp của dữ liệu, rất khó để trích xuất các đặc trưng của log nhật ký hệ thống theo cách thủ công, cách tốt nhất là thực hiện nó một cách tự động. Và dựa trên log nhật ký thu thập được, các quá trình phân tích có thể tạo ra thông tin mới, dự đoán cho các tình huống cụ thể và không lường trước được [2].

Theo một nghiên cứu [3] thì việc dự đoán lỗi hệ thống dựa trên log nhật ký là khả thi. Phân tích log nhật ký có thể được sử dụng để gỡ lỗi chương trình, khắc phục sự cố, đảm bảo độ ổn định hệ thống, tránh các lỗi hỏng, dự đoán trước các cuộc tấn công có thể xảy ra và tối đa hóa doanh thu quảng cáo. Vì vậy, chúng ta cần có khả năng dự đoán chính xác các sự kiện trong tương lai.

Các bản ghi log được tạo theo thứ tự thời gian, do đó mọi người bắt đầu phân tích chúng bằng các kỹ thuật dựa trên chuỗi thời gian và luôn cố gắng để tìm ra phương pháp phù hợp. Mặc dù đã đạt được nhiều tiến bộ, nhưng vẫn còn nhiều vấn đề có thể cải thiện để giảm độ phức tạp và tăng độ chính xác. Các nhà nghiên cứu gần đây đã có những bước đột phá lớn trong phân tích dữ liệu thông minh bằng mạng nơ-ron. Nghiên cứu này cũng đề xuất một phương pháp mới sử dụng mạng nơ-ron tuần hoàn của seq2seq để tìm hiểu log nhật ký và tạo mô hình dự đoán. Mô hình huấn luyện và dự đoán được mô tả như bên dưới:



Mô tả quá trình huấn luyện và mô hình dự đoán

Đầu vào của mô hình là các log thu thập được từ các nguồn khác nhau của hệ thống các thành phần dữ liệu lớn như Hadoop, YARN, MapReduce, Hive, HBase, Spark ... Đầu ra của mô hình là kết quả dự đoán các sự kiện có thể xảy ra trong tương lai dựa trên các log đầu vào đã thu thập được.

Mục tiêu

Các bản ghi log được tạo từ các thành phần khác nhau của hệ thống bigdata được nhập vào mô hình trong thời gian thực. Bằng cách sử dụng các thuật toán như thu thập dữ liệu, khai thác dữ liệu và học máy, các log này được tiến hành phân tích, liên kết ngữ nghĩa cấp cao và dự đoán các sự kiện trong tương lai. Mục tiêu cụ thể như sau:

- Xây dựng được thuật toán xử lý, tổng hợp các log thu thập được từ các thành phần hệ thống bigdata, làm cho chúng trở thành dữ liệu có cấu trúc, đồng nhất.
- Xây dựng mô hình huấn luyện và dự đoán được các sự kiện có thể xảy ra trong tương lai dựa trên các thông tin đầu vào là các log đã được xử lý sau khi thu thập với độ chính xác cao hơn các mô hình hiện có.
- Phương pháp phân tích dự đoán dựa trên seq2seq nhằm hỗ trợ quản lý hệ thống IoT tốt hơn, đáp ứng yêu cầu phân tích log nhật ký hiệu quả và chính xác.

<p>Nội dung và phương pháp thực hiện</p>	<p>Theo mô hình đề xuất, đầu tiên chúng ta sẽ thu thập log nhật ký của các thành phần riêng lẻ từ hệ thống bigdata. Tiếp theo, các log này sẽ được chuẩn hóa thành định dạng JSON, dữ liệu phi cấu trúc được chuyển đổi thành dữ liệu có cấu trúc và được lưu trữ trong một hệ thống lập chỉ mục văn bản đầy đủ phân tán [4]. Cuối cùng, khi dữ liệu đã sẵn sàng, chúng ta sẽ bắt đầu huấn luyện mô hình và đưa ra dự đoán bằng cách sử dụng mô hình đã huấn luyện đó. Các bước thực hiện được mô tả như sau:</p> <ul style="list-style-type: none"> - Bước 1: thu thập log <p>Bản thân nhật ký bigdata là một loại dữ liệu lớn, vì vậy chúng ta sử dụng các công cụ Elasticsearch, Logstash [5], Kibana để thu thập, xử lý và truy vấn chúng. Đây là các công cụ mã nguồn mở, cho phép thu thập, phân tích, xử lý và lưu trữ log với số lượng lớn, từ nhiều nguồn thu thập khác nhau một cách tự động và không làm mất dữ liệu. Logstash để thu thập nhật ký từ các thành phần khác nhau của hệ thống Hadoop (chẳng hạn như Hadoop, YARN, MapReduce, Hive, HBase, Spark, Solr, Carbondata, v.v.). Trong quá trình thu thập, nhật ký được định dạng, chuẩn hóa và được lưu trữ trong Elasticsearch. Toàn bộ các thông tin nhiễu, không cần thiết bị loại bỏ, đồng nhất và đưa về dạng dữ liệu có cấu trúc. Chúng ta cần xây dựng thuật toán cho việc xử lý, tổng hợp log trong bước này.</p> <ul style="list-style-type: none"> - Bước 2: phân loại log <p>Khối lượng dữ liệu của log rất lớn, việc phân loại các sự kiện để đơn giản hóa việc tính toán. Và ở đây, ta sử dụng danh mục làm mức độ chi tiết cho phân tích dự đoán. Chúng ta phân loại log dựa trên các thành phần, cấp độ log (thông tin, cảnh báo, lỗi) và nội dung chính của log (bắt đầu, kết thúc). Ta sử dụng một số thuật toán như độ tương tự cosine, kmeans để tạo từ điển danh mục. Log nhật ký sẽ gán thêm thẻ tag phân loại trước khi lưu trữ trong Elasticsearch ở bước trước.</p> <ul style="list-style-type: none"> - Bước 3: Vector hóa log <p>Mỗi log nhật ký được chuẩn hóa, định dạng và phân loại sẽ có một ID duy nhất (được xác định tự động trong Elasticsearch). Tiếp theo, chúng ta xây dựng một từ điển dựa trên tập này. Sau đó, log được chuyển thành vector và được huấn luyện.</p>
---	--

Cuối cùng, vector của mỗi log được lưu trữ trong một từ điển mới. Như vậy, các log biến đổi thành các vector có kích thước. Các vector này sẽ được cung cấp cho hai bước tiếp theo là mã hóa làm dữ liệu đầu vào và cho ma trận hiệu chỉnh. Bước này được gọi là log2vector.

- Bước 4: Đa dạng hóa tập huấn luyện

Chúng ta cần làm cho tập huấn luyện đa dạng hơn để tăng độ chính xác mà không làm thay đổi thông tin ban đầu. Bằng cách sử dụng một số tham số đầu vào và xây dựng thuật toán mới để đa dạng hóa tập huấn luyện.

- Bước 5: Mã hóa chuỗi

Các log đầu vào được vector hóa ở bước log2vectors. Tiếp theo, chúng được đưa vào lớp mã hóa theo thứ tự thời gian. Lớp mã hóa sẽ chuyển đổi chuỗi thành một vector trạng thái ẩn.

- Bước 6: Giải mã vector

Vector trạng thái ẩn thu được từ bước trước được nhân với ma trận hiệu chỉnh làm đầu vào cho lớp giải mã. Và thông tin mục tiêu được tạo ra bởi lớp giải mã, nó ở dạng chuỗi. Và chuỗi này cần được xử lý bởi một bộ điều chỉnh để dữ liệu dự đoán phù hợp hơn với dữ liệu thực.

Ở bước 5 và 6 này chúng ta cần xây dựng thuật toán huấn luyện và dự đoán dựa trên seq2seq của mạng nơ-ron hồi quy. Thuật toán gồm có hai phần là huấn luyện và dự đoán. Trong quá trình huấn luyện, các log bigdata được biến đổi thành vector đầu vào mạng nơ-ron để đào tạo mô hình dự đoán. Trong phần dự đoán, chúng ta có thể nhập chuỗi log được tạo theo thời gian thực vào mô hình được huấn luyện để thu được dữ liệu dự đoán mục tiêu.

- Bước 7: Lấy thông tin mục tiêu

Thông tin mục tiêu là chuỗi dự đoán mà chúng ta cần có được. Nó có thể đến từ một số thành phần, một số hệ thống ứng dụng hoặc hệ thống bảo mật. Chúng ta phải xác định loại thông tin mục tiêu trước bước huấn luyện. Sau đó, chúng ta gán nhãn của tập huấn luyện (thông tin mục tiêu của dữ liệu thực).

	<p>Cuối cùng, sẽ tiến hành thực nghiệm các bước đã trình bày ở trên trong môi trường thực tế để đánh giá các kết quả thu được, cũng như so sánh với các phương pháp trước đó. Các chỉ số đánh giá bao gồm sai số bình phương trung bình (RMSE), điểm phương sai mở rộng (ESV), sai số tuyệt đối trung bình (MEANAE), sai số tuyệt đối trung bình (MEDIANAE), R2 SCORE, độ lệch chuẩn của dữ liệu dự đoán và độ lệch chuẩn của giá trị thực dữ liệu.</p>
Kết quả dự kiến	<ul style="list-style-type: none"> - Kết quả mô hình dự đoán so với thực tế quan sát được ở các trường hợp: khi không hiệu chỉnh, có hiệu chỉnh và sử dụng thuật toán Random Forest [6]. - Bảng so sánh về các chỉ số đánh giá giữa mô hình do bài nghiên cứu đề xuất với các mô hình có sẵn như: Random Forest Model, AdaBoost Model, Bagging Model, Gradient Boosting Model, ARIMA và HMM.
Tài liệu tham khảo	<p>[1] Andrew Meloa, The US government is pouring money into the Internet of Things, https://www.businessinsider.com.</p> <p>[2] Mauro Coccoli, Paolo Maresca, Lidia Stanganelli, The role of big data and cognitive computing in the learning process, J. Visual Lang. Comput. 38 (2017) 97–103.</p> <p>[3] Fares A. Nassar, Dorothy M. Andrews, A methodology for analysis of failure prediction data, in: IEEE Real-Time Systems Symposium, 1985, pp. 160–166.</p> <p>[4] Oleksii Kononenko, Olga Baysal, Reid Holmes, Michael W. Godfrey, Mining Modern Repositories with Elasticsearch, ACM, 2014, pp. 328–331.</p> <p>[5] Sushma Sanjappa, Muzameel Ahmed, Analysis of logs by using logstash, 2017.</p> <p>[6] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32</p>