เว็บแอปพลิเคชันสำหรับการวิเคราะห์บันทึกข้อมูลจราจรคอมพิวเตอร์ขนาดใหญ่

A WEB-BASED APPLICATION FOR BIG DATA LOGS ANALYSIS

สลิลทิพย์ ธรรมศิริ[1]

เอกสิทธิ์ พัชรวงศ์ศักดา[2]

บทคัดย่อ

ในปัจจุบันเว็บแพลตฟอร์มอีคอมเมิร์ซได้รับความสนใจมากขึ้น เพราะความสะดวกสบายในการซื้อขายสินค้า ดังนั้นจึงมีข้อมูลธุรกรรมทางออนไลน์จำนวนมากที่ถูกจัดเก็บไว้ในฐานข้อมูล นอกจากนี้ยังมีไฟล์บันทึกข้อมูลการเข้าใช้งานเว็บไซต์ที่บันทึกข้อมูลการเข้าถึง เช่น IP address ของผู้ใช้งาน วันที่เข้าใช้งานและสถานะ ที่ถูกจัดเก็บในรูปแบบของ log file ในการใช้งานข้อมูลบันทึกนี้ ทางผู้วิจัยได้ติดตั้งแอปพลิเคชันบนเว็บที่สามารถวิเคราะห์ข้อมูลบันทึกการเข้าใช้งานเว็บไซต์จำนวนมาก และนำเสนอในรูปแบบของ dashboard ด้วยผลลัพธ์จากการวิเคราะห์นี้ ผู้ดูแลระบบสามารถตรวจดูความถี่ของ IP address แต่ละรายการ และดูข้อมูลช่วงเวลารวมไปถึงชั่วโมงเร่งด่วนที่ผู้ใช้งานเข้าใช้เว็บไซต์ นอกจากนี้ ระบบสามารถคาดการณ์จำนวนผู้ใช้งานที่จะเข้ามาใช้งานเว็บไซต์ล่วงหน้าใน 1 ถึง 6 ชั่วโมงข้างหน้าอีกด้วย

คำสำคัญ: ข้อมูลจราจรทางคอมพิวเตอร์, ปริมาณการใช้งานเว็บ, การเรียนรู้ของเครื่อง, อนุกรมเวลา

---

[1] นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

[2] ที่ปรึกษาสารนิพนธ์หลัก

ABSTRACT

Nowadays, an e-commerce platform gets more attention because it's convenient to buy products. Therefore, it has a large number of online transactions stored in a database. Moreover, there is a web access log file that records access data, e.g., client IP address, date, and status. To utilize this log data, we implemented a web-based application that can analyze a large volume of log data, and present it as a dashboard. With this result, an administrator can view the frequency of each IP address and see the peak hours that the many users accessed. Additionally, the system can forecast the number of access in the next 1 to 6 hours.

Keywords: Log File, Web Traffic, Machine learning, Time Series

บทนำ

ในปัจจุบันเว็บไซต์เข้ามามีบทบาทและความสำคัญต่อชีวิตประจำวันเราเป็นอย่างมาก ทุกองค์กรทั้งภาครัฐและเอกชน ได้ให้ความสำคัญของการมีเว็บไซต์เพิ่มมากขึ้น จุดประสงค์เพื่อประชาสัมพันธ์ หรือเพื่อการค้าขาย หลายองค์กรจึงมีการทำเว็บไซต์ เพื่อเป็นการยืนยันตัวตนให้เป็นที่รู้จักและเป็นการโฆษณาไปในตัว ในแง่ของการทำธุรกิจการสามารถใช้เว็บไซต์เพื่อช่วยกระตุ้นยอดขาย รวมถึงสร้างฐานลูกค้าใหม่ ๆ ได้อีกด้วย

โดยทั่วไปไม่ว่าเราจะใช้งานเว็บไซต์ไหนก็จะมีการจัดเก็บข้อมูลจราจรทางคอมพิวเตอร์ (Log file) เป็นข้อมูลที่แสดงเกี่ยวกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ แสดงให้เห็นถึง ต้นทาง, ปลายทาง, วันที่, เวลา หรืออื่น ๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ การบันทึกการใช้งานข้อมูลด้วย Log file คือวิธีที่จะช่วยแก้ปัญหาที่อาจจะเกิดขึ้น ไม่ว่าจะเป็นการส่งข้อมูลที่ผิดพลาดจากผู้ใช้งาน การขโมยข้อมูล หรือการกระทำที่เกิดจากการถูกโจมตีจากโปรแกรมที่ไม่พึงประสงค์ ที่ทำให้เกิดความเสียหายในระบบ การมีข้อมูลจาก Log file จะสามารถตรวจสอบที่มาที่ไปของข้อมูลได้ ทำให้หาสาเหตุของปัญหาและแก้ไขได้รวดเร็วยิ่งขึ้น

ในด้านของการทำธุรกิจออนไลน์การนำ Log file ของเว็บไซต์มาวิเคราะห์ข้อมูลผู้ใช้งาน จะสามารถทำให้เข้าใจพฤติกรรมของลูกค้าที่ใช้งานเว็บไซต์และเพื่อนำไปพัฒนาเว็บไซต์ให้มีประสิทธิภาพมากยิ่งขึ้น

งานวิจัยนี้จึงนำเสนอข้อมูลการวิเคราะห์ Web access log ผ่านเว็บแอปพลิเคชันในรูปแบบของ Dashboard เพื่อให้มองเห็นภาพรวมและพฤติกรรมของผู้ใช้งานเว็บไซต์ รวมไปถึงการวิเคราะห์ปริมาณการเข้าใช้งานเว็บไซต์ เพื่อใช้สำหรับวางแผนการจัดสรรทรัพยากรของระบบให้รองรับปริมาณการใช้งาน

วัตถุประสงค์ของงานวิจัย

1. เพื่อสร้างโมเดลสำหรับพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์
2. เพื่อสร้างเครื่องมือที่แสดงผลการวิเคราะห์ข้อมูลในหลายมุมมอง เข้าใจง่ายในรูปแบบของ Dashboard

ขอบเขตงานวิจัย

1. เป็นข้อมูล Web access log จาก Kaggle ที่นำมาใช้งาน
2. เป็นข้อมูลรูปแบบ Semi-structured
3. ข้อมูลเข้าใช้งานเว็บไซต์ตั้งแต่วันที่ 22 มกราคม 2562 จนถึงวันที่ 26 มกราคม 2562 มีจำนวนข้อมูลทั้งหมด 10,365,152 แถว

ทฤษฎี และผลงานที่เกี่ยวข้อง

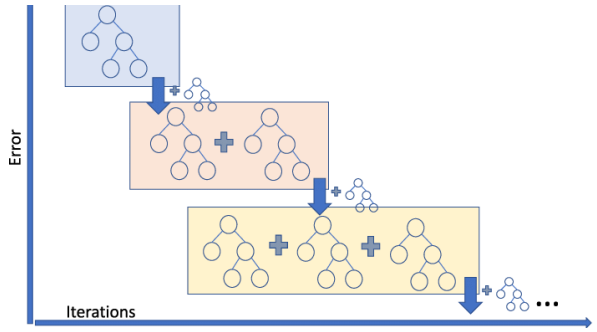
งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างเครื่องมือที่แสดงผลกการวิเคราะห์ข้อมูลในหลายมุมมอง เข้าใจง่ายในรูปแบบของ Dashboard และเพื่อพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยจำเป็นต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

1. การจัดเก็บข้อมูลการจราจรทางคอมพิวเตอร์ (Log file)

หรือที่เรียกอีกแบบว่าการเก็บข้อมูลการใช้งานคอมพิวเตอร์และอินเทอร์เน็ต คือ ข้อมูลเส้นทางการใช้งานอินเทอร์เน็ต หรือการเชื่อมต่อสื่อสารกันระหว่างเครื่องคอมพิวเตอร์ รวมถึงมือถือและแท็บเล็ต ที่มีไว้บันทึกเหตุการณ์ต่าง ๆ ข้อมูลที่เก็บไว้จะแสดงถึงต้นทาง, ปลายทาง, วันที่, เส้นทาง หรืออื่น ๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์

2. Gradient Boosted Trees

เป็นชุดของแบบจำลองการถดถอย (Regression) และการจำแนกประเภท (Classification) ทั้งสองวิธีเป็นวิธีการเรียนรู้แบบ Forward-learning มีพื้นฐานมาจาก Decision tree ซึ่งจะเป็นการสร้าง Decision tree ต่อกันหลายแบบ โดยที่แต่ละ Decision tree จะเรียนรู้จาก Error ของแบบจำลองก่อนหน้า ซึ่งเป็นการปรับปรุงประสิทธิภาพของแบบจำลองให้สูงขึ้น และประเมินผลแต่ละแบบจนกว่าจะได้ Decision tree ที่สมบูรณ์



ภาพที่ 1 ลักษณะการทำงานของ Gradient Boosted Trees

3. Autoregressive Integrated Moving Average (ARIMA)

เทคนิค ARIMA เป็นหนึ่งในการวิเคราะห์ข้อมูลแบบอนุกรมเวลา คือการอาศัยข้อมูลในอดีตมากำหนดรูปแบบของข้อมูลและการพยากรณ์ข้อมูลในอนาคต องค์ประกอบของแบบจำลอง ARIMA ประกอบด้วย 3 ส่วน ได้แก่ Autoregression process หรือ AR(p), Integrated (d) และ Moving average process หรือ MA(q)

Autoregression process หรือ AR(p) เป็นกระบวนการอธิบายตัวแปร Y ด้วยค่าของตัวแปร Y(Yt-1, Yt-2, Yt-3,…, Yt-p) ค่า Y ที่ใช้สามารถมีได้หลาย order, Integrated (d) d คือจำนวนความ

แตกต่างที่จำเป็นในการทำให้อนุกรมเวลาคงที่ และ Moving average process หรือ MA(q) คือการดูโมเดลว่าพึ่งพาระหว่างการสังเกตและข้อผิดพลาดที่เหลือจากแบบจำลองค่าเฉลี่ยเคลื่อนที่ที่ใช้กับการสังเกตที่ยังไง

รูปแบบ	สมการของ (Y)
ARIMA(1,1,0)/AR(1)	$\Delta Y_t = \alpha + \phi_1 \Delta Y_{t-1} + \varepsilon_t$
ARIMA(2,1,0)/AR(2)	$\Delta Y_t = \alpha + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} + \varepsilon_t$
ARIMA(0,1,1)/MA(1)	$\Delta Y_t = \alpha + \theta_1 \varepsilon_{t-1} + \varepsilon_t$
ARIMA(0,1,2)/MA(2)	$\Delta Y_t = \alpha + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$
ARMA(1,1,1)	$\Delta Y_t = \alpha + \phi_1 \Delta Y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$

ภาพที่ 2 ตัวอย่างสมการพยากรณ์ ARIMA model

4. งานวิจัยที่เกี่ยวข้อง

Tejas Shelatkar, Stephen Tondale, Swaraj Yadav and Sheetal Ahir, (2020). Web Traffic Time Series Forecasting using ARIMA and LSTM RNN. ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในการศึกษาปริมาณการเข้าใช้งานเว็บไซต์ โดยมีการเปรียบเทียบประสิทธิภาพของโมเดลทั้งหมด 2 แบบได้แก่ ARIMA model และ LSTM RNN จากผลการทดลองจะเห็นได้ว่าการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ ที่เป็นอนุกรมเวลาโดยใช้ LSTM-RNN มีประสิทธิภาพและแม่นยำมากกว่า ARIMA model

Jungkee Kim, (2018). Web Server Log Visualization. งานวิจัยนี้ได้นำเสนอข้อมูลของ Web log และการวิเคราะห์บันทึกการใช้เว็บไซต์ของ NPAC โดยการใช้ Shell script ในการบันทึกข้อมูลและใช้ Perl ในการสร้างข้อมูลรายวัน, รายสัปดาห์ และ โดเมน และบันทึกข้อมูลลงในฐานข้อมูล Oracle สำหรับขั้นตอน Pre-process ข้อมูลจะใช้ข้อมูลเพียงสามฟิลด์จากที่บันทึกไว้ในฐานข้อมูลได้แก่ Remote Host, Access Page และ Date-time โดยนำข้อมูลเหล่านี้เข้าโปรแกรม Java เพื่อแสดงผลข้อมูลการวิเคราะห์ออกมาเป็นกราฟ

ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นการนำเสนอเครื่องมือที่ช่วยพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ และเครื่องมือการแสดงผลการวิเคราะห์ข้อมูลแบบพื้นฐาน โดยมีขั้นตอนและแนวทางการวิจัยดังนี้

1. การเก็บข้อมูล

งานวิจัยนี้ได้ใช้ข้อมูลการบันทึกของเว็บเซิร์ฟเวอร์ที่บันทึกการเข้าใช้งานของเว็บไซต์ www.zanbil.ir จาก Kaggle ซึ่งเป็นข้อมูลการเข้าใช้งานเว็บไซต์อีคอมเมิร์ซของประเทศอิหร่าน ตั้งแต่วันที่ 22 มกราคม 2562 ถึง 26 มกราคม 2562

2. การเตรียมข้อมูล

2.1 นำข้อมูลเข้าฐานข้อมูล

นำข้อมูล Access log เข้าฐานข้อมูลโดยใช้ภาษา PHP ทำคำสั่ง loop ข้อมูลขึ้นฐานข้อมูลทั้งหมด

2.2 Feature Extraction

สำหรับข้อมูลจากฐานข้อมูลจะอยู่ในลักษณะ Semi-structure จะต้องทำการแบ่งข้อมูลออกมาเป็นส่วนต่าง ๆ และทำการเปลี่ยนชื่อแอตทริบิวต์เพื่อให้ง่ายต่อการใช้งาน และทำการเปลี่ยนแอตทริบิวต์วันที่จาก Norminal ให้เป็น Date ก่อนที่จะนำไปใช้งาน

3. การวิเคราะห์ข้อมูลพื้นฐานและบันทึกค่าลงฐานข้อมูล

หลังจากเสร็จสิ้นกระบวนการเตรียมข้อมูลแล้ว ข้อมูลจะถูกนำมาวิเคราะห์หาผลลัพธ์ในหลายๆด้าน เพื่อนำไปบันทึกลงฐานข้อมูลและแสดงผลบน dashboard

3.1 Count data group by request status ทำการหาค่าผลรวมจำนวนของแต่ละ request status

3.2 Count data group by request url ทำการหาค่าผลรวมจำนวนของแต่ละ request url

3.3 Count IP address group by IP address and datetime ทำการหาค่าผลรวมจำนวนของ IP address ที่มีการ request เข้าเว็บไซต์โดยอิงข้อมูลตาม IP address และ วันเวลา (รายนาที)

3.4 Web access traffic ทำการหาค่าผลรวมจำนวนของ request โดยอิงข้อมูลตามวันเวลา

3.5 Most active IP address ทำการหาค่าผลรวมจำนวนของ IP address ที่มีการ request เข้าเว็บไซต์โดยอิงค่าตาม IP address

3.6 Count request status group by request status and datetime Map ค่า request status ให้อยู่ในรูปแบบเดียวกัน เช่น (request status 200, 201, 202 จะถูก map ค่าให้เป็น 2XX) และทำการหาค่าผลรวมจำนวนของ request status เข้าเว็บไซต์โดยอิงค่าตาม request status และ วันเวลา

3.7 Count active url by path Split ค่า url ออก โดยใช้ / (Slash) เป็นตัวแบ่งข้อมูลออกจากกัน ก่อนนำข้อมูลของ path ไปหาค่าผลรวม

3.8 Count request by user agent ทำการหาค่าผลรวมจำนวนของ user agent ที่มีการ request เข้าเว็บไซต์โดยอิงค่าตาม user agent

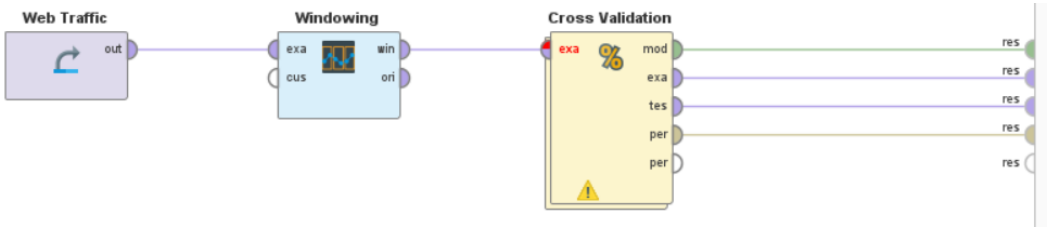
3.9 Map device user agent แปลงค่า user agent ให้กลายเป็น device (desktop, tablet, smartphone) โดยใช้ curl API ส่งค่า user agent ไปยังเว็บผู้ให้บริการข้อมูล user agent และนำค่าที่ return กลับมาบันทึกลงฐานข้อมูล

3.10 Map IP address geolocation แปลงค่า IP address ให้กลายเป็นประเทศโดยใช้ curl API ส่งค่า IP address ไปยังเว็บผู้ให้บริการข้อมูล IP address geolocation และนำค่าที่ return กลับมาบันทึกลงฐานข้อมูล

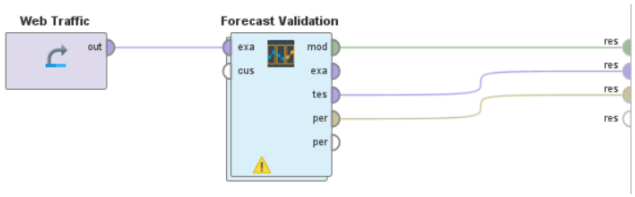
4. การสร้างแบบจำลอง

หลังจากได้ข้อมูลปริมาณการเข้าใช้งานเว็บไซต์แล้ว ก็นำไปสร้างแบบจำลองโดยงานวิจัยนี้ได้ทดลองกับแบบจำลอง Gradient Boosted Trees และ ARIMA model

ในส่วนของการสร้างแบบจำลอง Gradient Boosted Trees จะใช้โอเปอร์เรเตอร์ Windowing ในการสร้างตัวแปรปริมาณการเข้าใช้งานเว็บไซต์ย้อนหลังแบบ 3 ชม., 6 ชม., 9 ชม., 12 ชม., 15 ชม., 18 ชม., 21 ชม. และ 24 ชม. และส่วนของแบบจำลอง ARIMA จะใช้โอเปอร์เรเตอร์ Forecast Validation ในการสร้างตัวแปรปริมาณการเข้าใช้งานเว็บไซต์ย้อนหลังแบบเดียวกับโอเปอร์เรเตอร์ Windowing ก่อนที่จะนำข้อมูลเข้าไปยังแบบจำลอง Gradient Boosted Trees และ ARIMA



ภาพที่ 3 ตัวอย่างขั้นตอนการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ของแบบจำลอง Gradient Boosted Trees



ภาพที่ 4 ตัวอย่างขั้นตอนการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ของแบบจำลอง ARIMA

5. เครื่องมือที่ใช้ในงานวิจัย

งานวิจัยนี้ใช้ภาษา PHP ในการเขียนข้อมูล Web access log เข้าสู่ฐานข้อมูล และใช้ RapidMiner ตั้งแต่ขั้นตอนการเตรียมข้อมูล (Data processing) ไปจนถึง ขั้นตอนการสร้างแบบจำลอง และใช้ Laravel Framework ในการแสดงผลข้อมูลที่ได้จากการวิเคราะห์และพยากรณ์

ผลการศึกษา

งานวิจัยนี้ได้ทำการทดลองสร้างแบบจำลอง Gradient Boosted Trees และ ARIMA โดยใช้โปรแกรม RapidMiner ได้ผลดังตารางที่ 1 และ ตารางที่ 2

ตารางที่ 1 ตารางแสดงผลการทดสอบประสิทธิภาพของโมเดล Gradient Boosted Trees

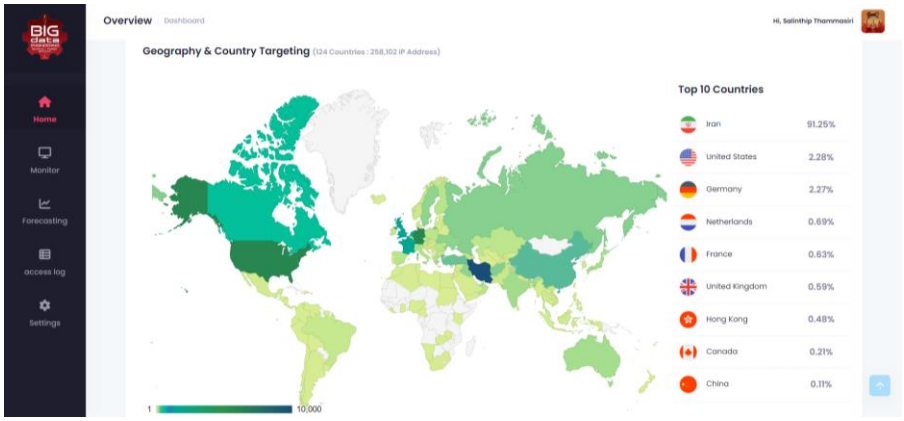
Windowing (size time)			Performance		
Window	**Step**	**Horizon**	**RMSE**	**Absolute Error**	**Relative Error**
3 Hours	1 Hours	1 Minutes	366.210	277.051	21.24%
6 Hours	1 Hours	1 Minutes	395.264	291.497	21.36%
9 Hours	1 Hours	1 Minutes	373.654	275.233	20.24%
12 Hours	1 Hours	1 Minutes	335.589	257.569	20.93%
15 Hours	1 Hours	1 Minutes	314.679	244.693	21.54%
18 Hours	1 Hours	1 Minutes	340.993	259.125	21.47%
21 Hours	1 Hours	1 Minutes	343.225	256.483	22.85%
24 Hours	1 Hours	1 Minutes	350.484	263.864	20.60%

ตารางที่ 2 ตารางแสดงผลการทดสอบประสิทธิภาพของโมเดล ARIMA

Windowing (size time)			Performance		
Window	**Step**	**Horizon**	**RMSE**	**Absolute Error**	**Relative Error**
3 Hours	1 Hours	1 Minutes	239.639	239.639	17.90%
6 Hours	1 Hours	1 Minutes	249.171	249.171	18.64%
9 Hours	1 Hours	1 Minutes	221.621	221.621	17.46%
12 Hours	1 Hours	1 Minutes	221.478	221.478	17.91%
15 Hours	1 Hours	1 Minutes	220.450	220.450	18.29%
18 Hours	1 Hours	1 Minutes	215.373	215.373	18.04%
21 Hours	1 Hours	1 Minutes	217.468	217.468	18.33%
24 Hours	1 Hours	1 Minutes	222.689	222.689	17.72%

จากตารางผลการทดสอบของโมเดล Gradient Boosted Trees และ ARIMA พบว่าโมเดล ARIMA ที่ใช้ตัวแปร Window size time เป็น 18 Hours มีประสิทธิภาพในการพยากรณ์ปริมาณการเข้า

ใช้งานเว็บไซต์ที่ดีที่สุดโดยให้ค่า RMSE อยู่ที่ 215.373 ค่า Absolute Error อยู่ที่ 215. 373 และค่า Relative Error อยู่ที่ 18.04% จากการทดลองหากนำข้อมูลที่ได้จากการวิเคราะห์พื้นฐานและผลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ไปใช้จริงบน dashboard ดังภาพที่ 5 และภาพที่ 6



ภาพที่ 5 ตัวอย่างการแสดงผลข้อมูล 10 อันดับสูงสุดของประเทศที่เข้าใช้งานเว็บไซต์

ภาพที่ 6 ตัวอย่างการแสดงผลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์บน Dashboard

สรุปและอภิปรายผลการวิจัย

1. จากการดึงข้อมูลจำนวน 10,365,152 แถว จากฐานข้อมูลเข้าโปรแกรม Rapidminer เพื่อทำการวิเคราะห์ข้อมูล พบว่าใช้เวลาในการดึงข้อมูลนาน ทางผู้วิจัยจึงแก้ปัญหาด้วยการดึงข้อมูลมาทำการ Feature extraction และ Cleansing data ให้ข้อมูลอยู่ในรูปแบบที่พร้อมใช้งานและทำการบันทึกข้อมูลที่ผ่านการ Clean แล้วเข้า Repository สำหรับนำมาใช้วิเคราะห์ข้อมูลต่อ

2. ข้อมูลแบบรายวันมีปริมาณเพียง 5 วัน ซึ่งเป็นปริมาณที่น้อยเกินไป ทำให้ไม่สามารถนำข้อมูลใช้กับแบบจำลองที่เป็น deep learning ได้ จึงได้ใช้แบบจำลอง ARIMA และ Gradient Boosted Trees ในการสร้างแบบจำลองและเปรียบเทียบประสิทธิภาพ ซึ่งทั้ง 2 แบบจำลองสามารถใช้กับชุดข้อมูลของงานวิจัยนี้ที่เป็นอนุกรมเวลา

3. ได้พัฒนาแบบจำลอง สำหรับพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ โอเปอร์เรเตอร์ Forecast Validation ที่ปรับค่าพารามิเตอร์ Window size time ที่ 18 Hours, Step ที่ 1 Hours และ Horizon ที่ 1 Minutes และโมเดล ARIMA ที่ปรับค่าพารามิเตอร์ p = 2 , ค่า d = 0 และค่า q = 1 ซึ่งให้ค่า RMAE อยู่ที่ 215.373 ค่า Absolute Error อยู่ที่ 215.373 และค่า Relative Error อยู่ที่ 18.04%

ข้อสังเกต

จากผลการทดสอบพบว่าข้อมูลที่นำมาใช้งานกับแบบจำลองมีค่าต่ำสุด (min value) และค่าสูงสุด (max value) ของข้อมูลห่างกันเกินไป ส่งผลให้ค่า RMSE, Absolute error และ Relative Error ที่ใช้สำหรับวัดประสิทธิภาพของแบบจำลองสูง ซึ่งทั้งสามค่าที่ใช้วัดประสิทธิภาพของแบบจำลองนี้ยิ่งค่าน้อยจะยิ่งดี จากการทดลองกับแบบจำลองเดียวกันด้วยชุดข้อมูลอื่นที่มีค่าต่ำสุด (min value) และค่าสูงสุด (max value) ของข้อมูลที่ไม่ห่างกันมาก ส่งผลให้ตัววัดประสิทธิภาพของแบบจำลอง RMSE, Absolute error และ Relative error มีค่าที่ต่ำ

ข้อเสนอแนะ

1. ข้อมูลที่นำมาใช้งานเมื่อทำการ Cleaning data แล้วข้อมูลในรูปแบบวันที่มีปริมาณที่น้อยจึงทำให้เกิดความยากในการใช้งาน

2. ข้อมูล Access log ที่มีข้อมูลจำนวนมาก การดึงข้อมูลจากฐานข้อมูลปกติและการนำข้อมูลมาวิเคราะห์อาจใช้เวลานานและโปรแกรมอาจรันไม่ไหว อาจต้องเปลี่ยนจากการเก็บข้อมูลในฐานข้อมูลไปเก็บที่ Hadoop แทน

3. ถ้าข้อมูล Access log ที่นำมาใช้งานมีข้อมูลของ Destination IP ด้วยจะสามารถนำมาต่อยอดการวิเคราะห์อื่น ๆ ได้อีก เช่น Anomaly Detection, Social Network Analysis

4. พัฒนาประสิทธิภาพของโมเดล หรือทดลองทำโมเดลอื่นเพิ่มเติม เพื่อให้ได้ผลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ที่แม่นยำมากขึ้น

บรรณานุกรม

ไขข้อข้องใจ Log File คืออะไร? ทำไมร้านกาแฟต้องจัดเก็บข้อมูล ตาม พ.ร.บ. คอมฯ.สืบค้น 19 พฤษภาคม 2565, จาก https://ditc.co.th/knowledge/log-file/

เอกสิทธิ์ พัชรวงศ์ศักดา. (2563). Practical Data Mining with Rapid Miner Studio 9. บริษัท เอเชีย ดิจิตอลการพิมพ์ จำกัด.

Junyan Shao. (2019). Web Traffic Time Series Prediction Using ARIMA & LSTM. สืบค้น 23 พฤษภาคม 2565, จาก https://medium.com/@jyshao53/web-traffic-time-series-prediction-using-arima-lstm-7ef3911845ae

Jungkee Kim. (2018). Web Server Log Visualization. *International journal of advanced smart convergence, 7(4)*, pp.101-107.

Nuthdanai Wangpratham. (2020).การพยากรข้อมูลอนุกรมเวลาด้วยเทคนิค ARIMA ด้วย Python สืบค้น 10 มิถุนายน 2565, จาก https://nutdnuy.medium.com/การพยากรณ์ข้อมูลอนุกรมเวลาด้วยเทคนิค-arima-ด้วย-python-44809eb8e990

Tejas Shelatkar, Stephen Tondale, Swaraj Yadav and Sheetal Ahir. (2020). Web Traffic Time Series Forecasting using ARIMA and LSTM RNN.