กระบวนการทำ ETL สำหรับการทำเหมืองข้อมูลของผู้ใช้ห้องสมุด เพื่อตรวจสอบความคุ้มค่าในการใช้วารสารอิเล็กทรอนิกส์

Extract-Transform-Load Process for Data Mining of Library Users to Examine the E-Journal Usage Worthiness

ปัณดารีย์ สุนทรวราภาส, อัษฎายุธ อุบลกาญจน์

สำนักทรัพยากรการเรียนรู้คุณหญิงหลง อรรถกระวีสุนทร
มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่
pandaree.s@psu.ac.th, aussadayut.u@psu.ac.th

บทคัดย่อ

Extract-Transform-Load (ETL) เป็นกระบวนการเพื่อเตรียมข้อมูลสำหรับการวิเคราะห์ในการทำเหมืองข้อมูล ETL เกี่ยวข้องกับการดึงข้อมูลจากแหล่งต่าง ๆ เช่น ฐานข้อมูล เพื่อเปลี่ยนเป็นรูปแบบที่เหมาะสม และสุดท้ายนำข้อมูลเหล่านั้นถ่ายโอนกลับสู่คลังข้อมูลก่อนการวิเคราะห์ด้วยเทคนิค data mining ในบทความนี้ได้นำเสนอกระบวนการ ETL สำหรับการขุดข้อมูลของห้องสมุดเพื่อตรวจสอบความคุ้มค่าในการใช้งานวารสารอิเล็กทรอนิกส์ นอกจากนี้จะมีหลายวิธีการของการตรวจสอบข้อมูลในระหว่างกระบวนการทำ ETL เช่น การทำความสะอาด, การรวม, การสร้าง, การจัดรูปแบบ ซึ่งพบว่ามีประโยชน์และมีประสิทธิภาพมากในทางเทคนิคในการตรวจสอบข้อมูลที่เตรียมไว้ในกระบวนการของ ETL

คำสำคัญ: การวิเคราะห์ข้อมูล, วารสารอิเล็กทรอนิกส์, ผู้ใช้บริการ

ABSTRACT

The Extract-Transform-Load (ETL) is the preprocessing to prepare data for analysis in data mining process. The ETL involves retrieving data from different sources like database to change to an appropriate format and finally load into data warehouse before analysis with data mining technique. In this paper, we present the ETL process for data mining of library users to examine the E-Journal usage worthiness. Also, the paper is included with several techniques of data validation during the ETL process. There are, cleaning, integration, constructing, formatting and mapping that are found very useful and effective technique to validate the prepared data in the ETL process.

Keyword: Extract, Transform, Load, ETL, Library, Data Mining, Data Analytics

บทนำ

ห้องสมุดถือเป็นศูนย์กลางความรู้ที่ให้บริการข้อมูลทั้งทรัพยากรทางด้านวิชาการและที่ไม่ใช่ทางด้านวิชาการ โดยเน้นผู้ใช้บริการเป็นสำคัญ ซึ่งอันที่จริงแล้วเมื่อมีผู้ใช้บริการเข้ามาใช้บริการทรัพยากรสารสนเทศของห้องสมุดก็จะส่งผลต่อประสิทธิภาพการให้บริการในหลาย ๆ ด้าน ดังนั้นจึงเป็นเรื่องสำคัญที่ห้องสมุดจะต้องคำนึงถึงความพึงพอใจของผู้ใช้บริการด้วยการตอบสนองความต้องการผ่านการศึกษาพฤติกรรมของผู้ใช้บริการด้วย

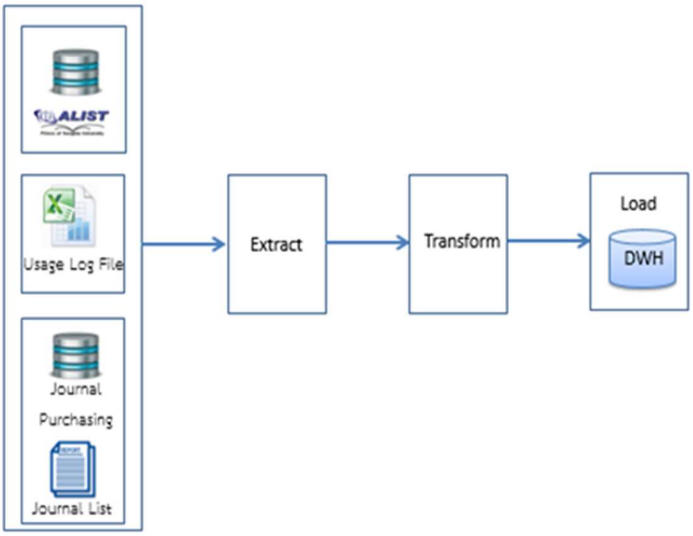
ห้องสมุดมหาวิทยาลัยส่วนใหญ่ในประเทศไทยได้นำระบบคอมพิวเตอร์มาใช้ในการบริการประเภทต่าง ๆ ดังนั้นจึงสามารถรวบรวมข้อมูลในแง่มุมต่าง ๆ ได้ เช่น ข้อมูลเกี่ยวกับการยืม การสืบค้น รวมถึงข้อมูลที่เกี่ยวข้องกับพฤติกรรมของผู้ใช้บริการ ในความเป็นจริงหัวใจสำคัญหลักของบริการห้องสมุดมหาวิทยาลัยคือการจัดหาทรัพยากรสารสนเทศทางวิชาการที่เกี่ยวข้องให้ตรงกับความต้องการของผู้ใช้บริการ ต้องทำให้มั่นใจว่าบริการต่าง ๆ จะได้รับความพึงพอใจจากผู้ใช้บริการ ผ่านการปรับปรุงบริการต่าง ๆ ปัจจุบันนี้ได้มีการนำเทคนิคการทำเหมืองข้อมูลไปใช้อย่างกว้างขวาง เพื่อค้นหาข้อมูลและความรู้ที่มีประโยชน์สำหรับการปรับปรุงการบริการรวมถึงการตัดสินใจ อย่างเช่นการใช้ data mining เพื่อทำนายความต้องการของสื่อประเภทต่าง ๆ และเพื่อวางแผนการจัดสรรงบประมาณให้ได้อย่างเหมาะสม ในการดำเนินการด้วยเทคนิค data mining ข้อมูลจำเป็นต้องถูกรวบรวมและทำความสะอาดก่อนที่จะนำไปทำการวิเคราะห์ซึ่งเรียกกระบวนการนี้ว่า Extract-Transform-Load หรือ ETL ซึ่งการทำ ETL เป็นกระบวนการดึงข้อมูลจากหลาย ๆ แหล่งข้อมูลนำไปจัดรูปแบบและตรวจสอบความถูกต้องของข้อมูลที่ไม่ถูกต้อง จากนั้นนำข้อมูลที่ได้รับการตรวจสอบแก้ไขแล้วป้อนกลับลงสู่ในคลังข้อมูลต่อไป

วัตถุประสงค์

1. เพื่อศึกษากระบวนการทำ ETL เป็นขั้นตอนหรือกลไกในการเพิ่มประสิทธิภาพงานของห้องสมุด
2. เพื่อนำข้อมูลที่ได้ไปสร้างแบบจำลองข้อมูล (Data Model) สำหรับตรวจสอบความคุ้มค่าในการใช้วารสารอิเล็กทรอนิกส์

ขั้นตอนและวิธีการดำเนินการ

งานวิจัยนี้แบ่งขั้นตอนกระบวนการทำ ETL เป็น 3 ขั้นตอน ได้แก่ การดึงข้อมูล การแปลงข้อมูล และการนำข้อมูลกลับสู่คลังข้อมูล การศึกษาครั้งนี้ใช้โปรแกรมซอฟต์แวร์ KNIME เพื่อทำกระบวนการ ETL โปรแกรม KNIME เป็นแพลตฟอร์มการวิเคราะห์ข้อมูลแบบโอเพ่นซอร์สและเป็นเครื่องมือที่เหมาะสมสำหรับใช้ในการวิเคราะห์ข้อมูล ผู้วิจัยแสดงให้เห็นถึงการทำงานของ 3 กระบวนการ ETL หลักได้แก่ การแยก แปลง และโหลดตามที่แสดงในภาพที่ 1



ภาพที่ 1. กระบวนการทำ ETL ของข้อมูลห้องสมุด

1. การสกัดข้อมูล

กระบวนการนี้เริ่มต้นด้วยการนำข้อมูลจากแหล่งต่าง ๆ ใน 3 แหล่งได้แก่ ระบบห้องสมุดอัตโนมัติสำหรับสถาบันอุดมศึกษาไทย (ALIST), ข้อมูลการใช้งานวารสารอิเล็กทรอนิกส์, ข้อมูลรายชื่อวารสารอิเล็กทรอนิกส์และข้อมูลจากระบบการจัดซื้อวารสารอิเล็กทรอนิกส์

ALIST เป็นฐานข้อมูลของระบบห้องสมุดอัตโนมัติในระบบนี้สำหรับการดึงข้อมูลของผู้ใช้บริการ ซึ่งข้อมูลของผู้ใช้บริการประกอบด้วย บาร์โค้ด (ID), ชื่อสมาชิก, เพศ,เลขสมาชิก, ที่อยู่ปัจจุบัน, หมายเลขบัตรประชาชน, หมายเลขโทรศัพท์, วันที่สมัครสมาชิก, วันที่หมดอายุ, อีเมล, คณะ, สาขาวิชา ฯลฯ รวมทั้งสิ้นมี 32 คุณลักษณะ แต่มีเพียง 5 คุณลักษณะที่นำไปใช้ในการเตรียมข้อมูลสำหรับการประเมินความคุ้มค่าในการใช้งานวารสารอิเล็กทรอนิกส์ ได้แก่ บาร์โค้ด, ชื่อ, ประเภทสมาชิก, คณะ, และสาขาวิชา จำนวนข้อมูลของสมาชิกในตารางนี้คือ 102,610 รายการ ตารางอื่น ๆ คือ ประเภทสมาชิก ที่มีการรวบรวมบาร์โค้ด และชื่อประเภทสมาชิกที่ประกอบด้วยสมาชิก 32 ประเภท ทั้ง 2 ตารางถูกส่งออกในรูปแบบไฟล์ Excel จากระบบ ALIST

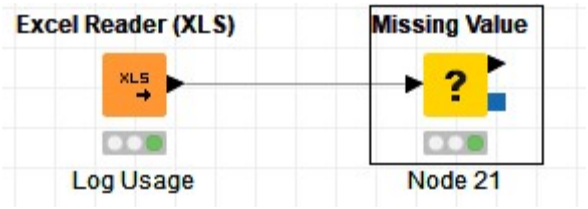
ข้อมูลการใช้งานวารสารอิเล็กทรอนิกส์ เป็นไฟล์บันทึกที่รวบรวมเกี่ยวกับการใช้งานวารสารอิเล็กทรอนิกส์ของสมาชิกห้องสมุดทั้งเครือข่ายภายในและภายนอกมหาวิทยาลัย ไฟล์รูปแบบนี้ยังเป็น. xls และประกอบไปด้วย 38 คุณลักษณะ อย่างไรก็ตามในแง่ของการประเมินความคุ้มค่าในการใช้งานวารสารอิเล็กทรอนิกส์การศึกษานี้เลือกเพียง 7 คุณลักษณะ ได้แก่ วันที่, เวลา, เซสชั่น, ผู้ใช้, ชนิดของบริการ, ชื่อโฮสต์และ URL ไฟล์นี้มีขนาดใหญ่มาก ซึ่งมีการรวบรวมข้อมูลที่มากกว่า 30,000 ระเบียนต่อวัน

แหล่งข้อมูลสุดท้ายคือ ข้อมูลจากระบบการบอกรับวารสาร ในระบบนี้ต้องใช้ไฟล์ข้อมูล 2 ไฟล์ เช่น ไฟล์การจัดซื้อวารสารอิเล็กทรอนิกส์และรายการวารสารอิเล็กทรอนิกส์ การจัดซื้อวารสารอิเล็กทรอนิกส์เป็นไฟล์ Excel ที่รวบรวมการจัดซื้อวารสารอิเล็กทรอนิกส์ในแต่ละปี และอีกไฟล์คือรายการของวารสารอิเล็กทรอนิกส์ในห้องสมุดในช่วงเวลา 3 ปีที่ผ่านมา และจำนวนวารสารอิเล็กทรอนิกส์ทั้งหมดคือ 6,432 รายการ และถูกจัดเก็บไว้ในรูปแบบไฟล์ Excel

2. การเปลี่ยนแปลง/เปลี่ยนรูปข้อมูล

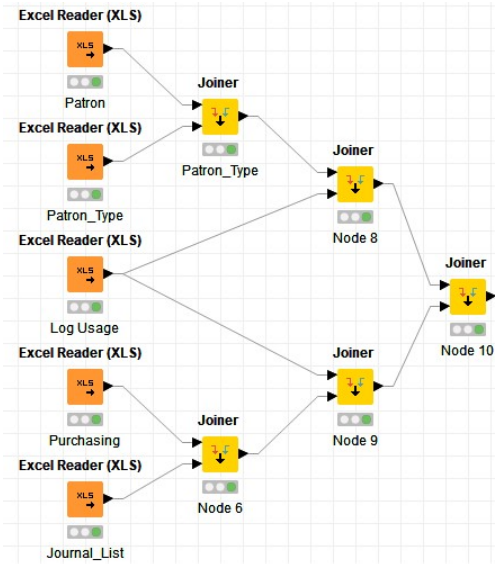
กระบวนการแปลงข้อมูลเป็นรูปแบบที่สามารถวิเคราะห์ได้ (Transform Data) ซึ่งสามารถทำได้โดยการกำจัดข้อมูลที่ผิดปกติ (Noisy Data) และข้อมูลที่อาจส่งผลให้เกิดการวิเคราะห์ข้อผิดพลาด ซึ่งรวมถึงการลดขนาดของข้อมูลที่ไม่จำเป็นสำหรับการวิเคราะห์ (Data Reduction) ด้วยเทคนิคต่อไปนี้

2.1 การทำความสะอาดข้อมูล หลังจากชุดข้อมูลได้รับการคัดเลือกเรียบร้อยแล้ว จากนั้นนำมากำจัดค่าผิดปกติไม่ถูกต้องและซ้ำซ้อนดังแสดงในภาพที่ 2 กระบวนการนี้เพื่อให้แน่ใจในคุณภาพและความถูกต้องของข้อมูล ข้อมูลที่สมบูรณ์เท่านั้นที่จะถูกเลือกสำหรับกระบวนการวิเคราะห์



ภาพที่ 2. ชุดข้อมูลที่ไม่สมบูรณ์

2.2 การรวมข้อมูล เป็นวิธีการนำข้อมูลจากหลาย ๆ แหล่งมารวมอยู่ในตารางข้อมูลเดียวเป็นชุดข้อมูลใหม่เชื่อมโยงความสัมพันธ์ของข้อมูลในแต่ละแหล่งข้อมูลเพื่อให้สามารถวิเคราะห์ข้อมูลได้พร้อมกันในคราวเดียวได้อย่างง่าย ดังแสดงในภาพที่ 3



ภาพที่ 3. การรวมข้อมูลจากแหล่งต่าง ๆ

2.3 การสร้างข้อมูล เป็นการสร้างลักษณะข้อมูลใหม่หรือสร้างชุดข้อมูลใหม่ให้แล้วเสร็จ เป็นการตรวจสอบเพิ่มเติมสำหรับข้อมูลเพื่อให้ข้อมูลมีความสมบูรณ์ที่สุด

2.4 การจัดรูปแบบข้อมูล เป็นการปรับข้อมูลให้เสร็จสมบูรณ์ถูกต้องและเหมาะสมกับพารามิเตอร์ที่ใช้ในการวิเคราะห์ เช่น การบันทึกข้อมูลเพศชายและหญิง เมื่อทำการฟอร์แมตใหม่ให้ใช้หมายเลข 1 = male และ 2 = female

2.5 การจับคู่ของข้อมูล เป็นกระบวนการของการทำการจับคู่ของข้อมูลระหว่างตารางต้นทางและปลายทาง แต่ละตารางระเบียนหรือแอตทริบิวต์ในเป้าหมายได้มาจากตารางระเบียนหรือแอตทริบิวต์เดียวกันในแหล่งที่มา ตารางที่ 1 แสดงข้อมูลผู้ใช้จากตารางต้นทางเดียวในขณะที่ตารางที่ 2 แสดงข้อมูลผู้ใช้จากหลาย ๆ แหล่ง ซึ่งทั้งคู่บ่งบอกถึงการจับคู่ตารางแบบหนึ่งต่อหนึ่ง ตารางที่ 3 ระบุการทำจับคู่แบบหลายต่อหลายและตารางที่ 4 การทำจับคู่หลายต่อหลาย

ตารางที่ 1 การแปลงตารางแบบ one to one

Source table :: column	Target table :: column	Selection Condition
Barcode	Barcode	Transform all new barcode
Patron_Name	Patron_Name	Transform all new patron_name
Gender	Gender	Transform all new gender

ตารางที่ 2 การแปลงตารางแบบ many to one

Source table :: column	Target table :: column	Selection Condition
Service	Hostname	Transform to hostname
Url	Hostname	Transform to hostname
referal	Hostname	Transform to hostname

ตารางที่ 3 การแปลงตาราง one to one โดยการรวมเร็คคอร์ดหลายรายการต่อหนึ่งรายการ

Source table :: column	Target table :: column	Selection Condition
Patron_Type_Bachelor	Patron_Type	Transform all new patron_type
Patron_Type_Master	Patron_Type	Transform all new patron_type
Patron_Type_Lecturer	Patron_Type	Transform all new patron_type

ตารางที่ 4 การแปลงตารางแบบ one to one โดยการรวมเร็คคอร์ดหลายรายการต่อหลายรายการ

Source table :: column	Target table :: column	Selection Condition
Journal_name	purchase	Transform journal list of purchase

จากข้อมูลในตารางที่ 5 เป็นตัวอย่างของชุดข้อมูลก่อนกระบวนการ ETL พบว่าข้อมูลเพศของผู้ใช้บริการ “6110121774” หายไป ข้อมูลอายุของรหัสผู้ใช้บริการ “6110121033” ที่มีอายุ 0 ปีมีแนวโน้มที่จะไม่สอดคล้องกับความเป็นจริง เพราะผู้ใช้บริการประเภทต่าง ๆ ควรมีอายุมากกว่า 0 ปี สำหรับหมายเลขโทรศัพท์มือถือ

ของรหัสผู้ใช้ “6110121048” ที่มีทั้งหมด 9 หลัก ในความเป็นจริงหมายเลขโทรศัพท์มือถือปัจจุบันคือ 10 หลัก ดังนั้นข้อมูลควรปรับหมายเลขโทรศัพท์มือถือของผู้ใช้เป็น 10 หลักเช่นกัน

ตารางที่ 5 ตัวอย่างของชุดข้อมูลที่ไม่สมบูรณ์ก่อนกระบวนการ ETL

Barcode	Gender	Age	Mobile Phone
6110121033	F	0	0896962265
6110121048	M	26	093441520
6110121774		32	0815391711
6110121887	F	37	0819512121

นอกเหนือจากข้อมูลที่ผิดปกติในตารางที่ 5 เราจำเป็นต้องทำความสะอาดชุดข้อมูลเพื่อให้ได้ชุดข้อมูลที่ถูกต้องและครบถ้วนก่อนที่จะทำการวิเคราะห์ สำหรับการแก้ไขข้อมูลบางครั้งสามารถแก้ไขได้โดยใช้การพิจารณาของตนเองเช่น หมายเลขโทรศัพท์มือถือที่ไม่สมบูรณ์สามารถเติมได้สูงสุด 10 หลัก อีกวิธีคือ การขอข้อมูลจากเจ้าของข้อมูลเองหรืออาจต้องใช้หลักการทางสถิติเพื่อค้นหาค่าที่เหมาะสมที่จะใช้แทน สิ่งสำคัญคือการเก็บข้อมูลส่วนบุคคลของผู้ใช้ไว้เป็นความลับ ข้อมูลผู้ใช้ไม่ควรแสดงในตารางที่ 6-7 ดังนั้นข้อมูลที่มีอยู่ในคลังข้อมูลต้องไม่ระบุว่า ข้อมูลเป็นของผู้ใช้บริการใด ๆ ดังแสดงในตารางที่ 8

ตารางที่ 6 ข้อมูลสมาชิก

Barcode	Patron_Name	Patron_Type
6110121048	Thanadech Sukhum	Bachelor Degree
6110121774	Somkit Meeyam	Bachelor Degree
6110121033	Pandaree Soonthonwarapas	Master/Ph.d.
6110125443	Mali Songtong	Master/Ph.d.
6110121851	Jaidee Meefah	Bachelor Degree
⋮	⋮	⋮

ตารางที่ 7 ข้อมูลการใช้วารสารอิเล็กทรอนิกส์

Date	Time	Sessionid	Hostname
01/12/2020	10.05	6110121048	https://dl.acm.org
05/12/2020	13.00	6110121774	www.webofknowledge.com
12/12/2020	14.55	6110121033	https://ieeeplore.ieee.org
19/12/2020	16.05	6110125443	https://onlinelibrary.wiley.com
25/12/2020	18.14	6110121851	https://dl.acm.org
⋮	⋮	⋮	⋮

ตารางที่ 8 คลังข้อมูล

Date	Time	Sessionid	Patron_Type	Hostname
01/12/2020	10.05	6110121048	Bachelor Degree	https://dl.acm.org
05/12/2020	13.00	6110121774	Bachelor Degree	www.webofknowledge.com
12/12/2020	14.55	6110121033	Master/Ph.d.	https://ieeeplore.ieee.org
19/12/2020	16.05	6110125443	Master/Ph.d.	https://onlinelibrary.wiley.com
25/12/2018	18.14	6110121851	Bachelor Degree	https://dl.acm.org
⋮	⋮	⋮	⋮	⋮

3. การถ่ายโอนข้อมูล

เมื่อกระบวนการเปลี่ยนแปลง/เปลี่ยนรูปข้อมูลเสร็จสมบูรณ์ก็สามารถถ่ายโอนข้อมูลเหล่านั้นกลับไปสู่ยังคลังข้อมูลเพื่อดำเนินการวิเคราะห์โดยใช้เทคนิคการทำเหมืองข้อมูลต่อไป

สรุปผล อภิปรายผล ข้อเสนอแนะและการนำไปใช้ประโยชน์

สรุปผล อภิปรายผล

การทำ ETL จะทำให้ข้อมูลที่ได้มีความถูกต้อง น่าเชื่อถือและสามารถนำข้อมูลเหล่านั้นไปทำการวิเคราะห์การใช้วารสารอิเล็กทรอนิกส์ได้อย่างง่าย เพราะข้อมูลเหล่านั้นมีผลต่อการตัดสินใจในการจัดหาวารสารอิเล็กทรอนิกส์ให้แก่ผู้ใช้บริการของสำนักทรัพยากรการเรียนรู้คุณหญิงหลงฯ นอกจากนี้ผลการวิจัยนี้ยังสอดคล้องกับผลวิจัยของ สุพจน์ ชุมสิทธิ์ (2560) ที่แสดงให้เห็นว่ากระบวนการ ETL ทำให้สามารถเข้าถึงข้อมูลได้โดยง่าย แสดงออกมาในรูปแบบรายงาน และสามารถปรับมุมมองในการวิเคราะห์ให้ตรงตามความต้องการในการนำข้อมูลไปใช้ให้เกิดประสิทธิภาพมากยิ่งขึ้น และผลงานของ Homayouni et al. (2018) ที่แสดงว่า ETL มีความสำคัญกับการเตรียมข้อมูลสำหรับข้อมูลขนาดใหญ่ก่อนนำไปใช้ในการทำการวิเคราะห์

ข้อเสนอแนะและการนำไปใช้ประโยชน์

กระบวนการทำ ETL สำหรับการทำเหมืองข้อมูลของผู้ใช้ห้องสมุดเพื่อตรวจสอบความคุ้มค่าในการใช้วารสารอิเล็กทรอนิกส์ มีวิธีการและเทคนิคที่พบว่า มีประโยชน์และมีประสิทธิภาพมากซึ่งสามารถช่วยตรวจสอบข้อมูลในกระบวนการ ETL ก่อนการนำไปใช้ในการวิเคราะห์ด้วยเทคนิคทางเหมืองข้อมูล (Data Mining) ประกอบด้วย การทำความสะอาด การรวม การสร้าง การจัดรูปแบบ และการจับคู่ และในอนาคตควรจะเพิ่มเติมเกี่ยวกับวิธีการทดสอบประเมินผลประสิทธิภาพของการทดสอบความถูกต้องของข้อมูล จะทำให้สามารถนำข้อมูลที่ได้ไปทำการวิเคราะห์ผลการใช้วารสารอิเล็กทรอนิกส์ที่จะสามารถนำไปประเมินได้ว่า คุ้มค่ากับการจัดซื้อวารสารต่อไปในอนาคตด้วย

รายการอ้างอิง

สุพจน์ ชุมสิทธิ์. (2560) *การพัฒนาระบบธุรกิจอัจฉริยะด้านระบบบัญชีสำหรับการรถไฟแห่งประเทศไทย* (วิทยานิพนธ์ปริญญามหาบัณฑิต). มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

Homayouni, H., Ghosh, S., & Ray, I. (2018). *An Approach for Testing the Extract-Transform-Load Process in Data Warehouse Systems.* Proceedings of the 22nd International Database Engineering & Applications Symposium on - IDEAS 2018. doi:10.1145/3216122.3216149