Open Access
Baghdad Science Journal **Vol.16 (Special Issue) 2019**

DOI: http://dx.doi.org/10.21123/bsj.2019.16.2(SI).0436

An Optimised Method for Fetching and Transforming Survey Data based on SQL and R Programming Language

Forat Falih Hasan* ***Zulikha Jamaluddin*****

Received 3/9/2018, Accepted 12/11/2018, Published 20/6/2019

Abstract:

The development of information systems in recent years has contributed to various methods of gathering information to evaluate IS performance. The most common approach used to collect information is called the survey system. This method, however, suffers one major drawback. The decision makers consume considerable time to transform data from survey sheets to analytical programs. As such, this paper proposes a method called 'survey algorithm based on R programming language' or SABR, for data transformation from the survey sheets inside R environments by treating the arrangement of data as a relational format. R and Relational data format provide excellent opportunity to manage and analyse the accumulated data. Moreover, a survey system based on structured query language and R programming language is designed to optimize methods to manage survey systems by applying large features offered via combining multi data science languages. The experiments verified enhancements of flexibility, technical tools, and data visualization features employed to process the collected data from different aspects; therefore, the proposed approach demonstrates a simple case study to enhance the evaluation requirements of the proposed technique. Finally, the estimated results of this research can be used to improve the methods of information management on different aspects such as survey systems and other data models that hold the relational and non-relational models using SABR. This method demonstrated improved accuracy of data collected, reduced data processing time and arranged data to the willing model.

Key words: Data transformation, NoSQL, R programming, Structured query language.

Introduction:

Most of the decision must be made in real time to ensure its effectiveness (1, 2). For that purpose, the process of collecting data from different sources and transforming them into the format and formula desired by the decision makers become a challenging tasks. In this context, the problem can be defined as an assumption of a group of N people who answered the questionnaire related to the specific study. Each survey sheet consists of two parts: the first table [Info matrix = (i,j)] that holds a general information of the participant with unlimited attributes and rows, and the second table [SQ matrix = (i,j)] which holds questionnaire questions and answers fields with unlimited attributes and rows.

Universiti Utara Malaysia, Malaysia.
Corresponding Author: *Forat.db@gmail.com, **Zulie@uum.edu.my

The main problem in this situation is how to collect those data using a high accuracy method in a short time while providing a wide platform for thorough data analysis in the r programming environment.

In this paper, the problems of transferring data from the survey sheets to the desired data repositories are dissolved by the proposed approach. The proposed approach for collecting and analysing the survey sheets creates a new method for decision makers through its features. The main advantage of this approach is that the collected data can be formatted according to the requirements of the decision makers as well as other analytical techniques. In other words, it supports the automatic formats of the data sets according to the end users' criteria.

R is an open source community that gives more flexibility and extensibility with thousands of packages. It is considered the most common and powerful language to handle statistical data and

provides a great environment to deal with big data analytics. Furthermore, R gives the opportunity to run structured query language (SQL) which is considered the most powerful language to treat relational data (3, 4). SQL plays a vital role in managing relational data because of its higher performance and support of unlimited functions to deal with relational aspects. SQL supports many applications in many fields like web applications and information discovery applications (5, 6).The combination of the features of R programming and SQL opens up new horizons in the field of combining and analysing data. This environment provides unlimited flexibility and tools to treat the data according to the desired criteria.

In this paper, the proposed combination of R and SQL is applied in the field of information systems by using the Likert-Scale that is a standard approach to collect and analyze data. Thus, the main contribution of this research is on reducing the complexities and the difficulties of collecting and organizing a group of survey sheets in the desired location and in a desired format and styles. Such complexity on relational and non-relational data that come from different sources are taxing for the researchers, therefore, our *survey algorithm based on R programming language* (SABR) model for the purpose of dissolving the problems of data transformation from various survey sheets (sources) to the desired repository, is a novel contribution.

Related Work:

Recently, the decision makers started evaluating the performances of the IS quickly, then use the survey systems to daily increase the clear views of the current works. There exist multiple methods to evaluate the dependency of the IS situations, such as adopting one or more than one survey per days for a specific case. This huge data that flow from various sources needs an integrated environment to be analysed and obtained the desired results. For this reason, the issues are addressed for transforming data coming out from various survey sheets for collecting them in one repository, as well as changing the format of relational and non-relational models due to the cumulative data on both sides that need decision makers and technical tools. (7) proposes QSL for administering the questionnaire and managing data in the different types of E-questionnaire. Besides, (8) presents a privacy-preserving survey system to collect and analyse the data based on secure developed protocols. (9) tried to solve the problem of survey systems in a small company. They proposed methods of analysis to study the relations between the items. Further, (10) presents a survey system based on a website application (I-System) and analysing data collection by using well-known data mining algorithms. Finally, (11) demonstrates an approach to integrate and capture the information that flows between two tables in a survey system. As a result, it can be concluded that all the presented referenced works did not demonstrate a distinctive way with the dynamic survey sheets that contain a table in the form of restricted direction models. In addition, there exists no proposed method that enhances the ata modes to relational or non-relational and vice versa.

Preliminary:

To describe the proposed work, this section presents some terms related to the research approach.

1. **Survey Sheet:** refers to the group of data that is gathered by individual questions using the electronic paper format. In this study the survey sheets consist of two main tables, Information table and SQ table.

A. Information Table: The information table holds the general information about the people that answer the survey questions and consists the following attributes as shown in Figure 1.:

1. ID: Primary Key
2. Gender
3. Age
4. Marital Status
5. Education
6. Education Field
7. Years at Company
8. Years with Current Job

	A	B	C	D	E	F	G	H
1	ID	Gender	Age	MaritalStatus	Education	EducationField	YearsAtcompany	YearsWithCurrentjob
2	101	Male	28	Single	Master	Technical Degree	4	2

Figure 1. Information table in Excel format.

B. SQ Table: The survey questions table hold all the question details and consists of the following attributes as shown in Figure 2:

1. ID: The foreign key that is linked directly to the ID attribute in the Information table after processing by the R programming language to construct a one-to-many data relationship.
2. Section: If the survey system used consists of more than one parts, the part number can be mentioned in this field.
3. QN: The question number attribute refers to the sequence of the question in the survey sheet.
4. Question: It holds the survey question text.
5. Strongly Agree, Agree, Average, Disagree and Strongly Disagree columns refer to answers indicating the levels of evaluating the asked questions from one to five degrees.

	A	B	C	D	E	F	G	H	I
1	ID	Section	qn	Question	Strong_agree	Agree	Average	Disagree	Strong_disagree
2	101	1	q1	(IS) can be used easily to access information that I need					
3	101	1	q2	(IS) has all the useful features (help, sort, etc.) that I ever require					
4	101	1	q3	(IS) is always available to be used during operational hours as stated in service level agreement					
5	101	1	q4	(IS) does data/information processing in a short, timely fashion					
6	101	1	q5	(IS) has a good security measures against possible threats or vulnerabilities					

Figure 2. SQ table in excel format.

2. SABR: Is the short form of *Survey Algorithm Based on R Programming Language*. SABR is a proposed algorithm written under the environment of R programming language supported by the features of both data science languages, R and SQL.

Methodology:

The method is described in accordance to steps carried out in the study namely data collection, data processing, and algorithm development.

Data Collection: In order to test and run the proposed work the real data is firstly gathered. We administered questionnaires consisting of questions as shown in Fig. 2. The questions targeted at measuring the information system quality (12) at AL-Kitab University. There are in total 28 survey sheets collected from the respondent.

General View of the Data Processing Steps: At this stage, there are four main steps that lead towards generating the desired results as shown in Figure 3. After collecting the distributed survey sheets and arranging them, the first step is to determine the location of source files (survey sheets) that are already stored in the computer storage. The second step is uploading the survey sheets into the R programming environment. The third step refers to the procedure that transforms the data from the survey sheets to the desired tables and format (relational and non-relational) as shown in Figure 4. The fourth and last step is analyzing and printing the cumulative data stored in the Information and SQ tables according to the predetermined goals.

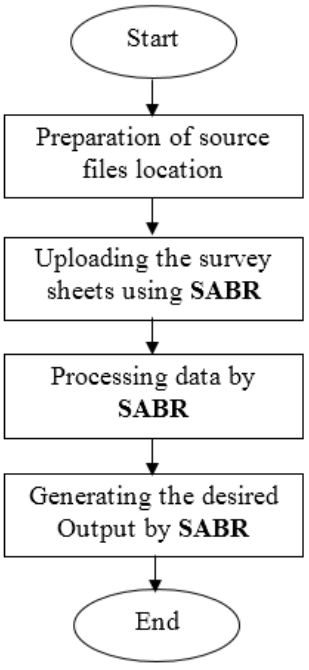

Figure 3. General view of the data processing.

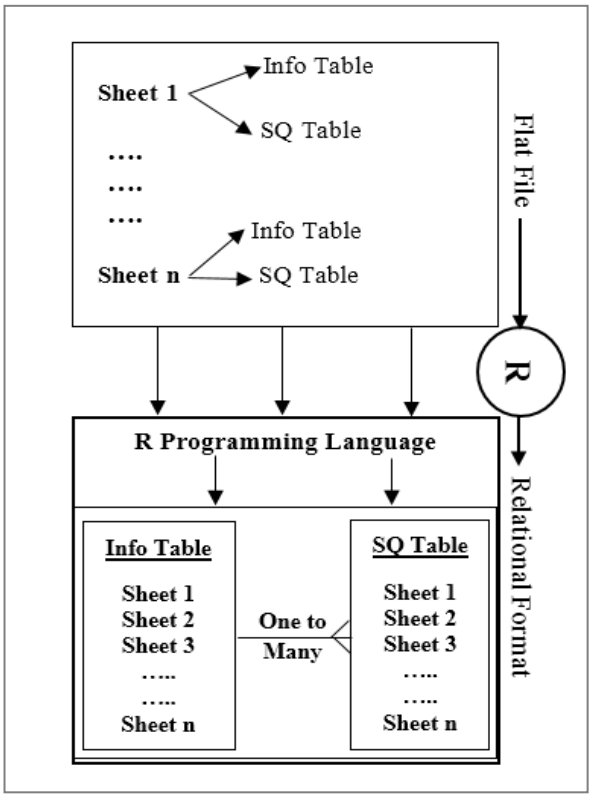

Figure 4. Data from non-relational to relational format.

Algorithm (SABR) development: The overall work of the proposed approach is as shown in Figure 5. It can be enumerated in the following steps:

1. All the survey sheets collected are arranged from (1 to n) in one folder.
2. A new information table is created as a repository to store the data of all the information tables from sheet 1 to sheet n.
3. A new SQ table is created as a repository to store the data of all the SQ tables from sheet 1 to sheet n.
4. The numbers and the path of the survey sheets' location are assigned to the algorithm.
5. The data is fetched from the sources tables to both repositories information and SQ tables as shown below:

$$Info = Info + Info\ sheet\ (i)$$

$$SQ = SQ + SQ\ sheet\ (i)$$

6. Finally, when all the data are successfully uploaded from the survey sheets to the information and SQ tables, the one-to-many data relationship between them can be done using ID attributes in both tables.

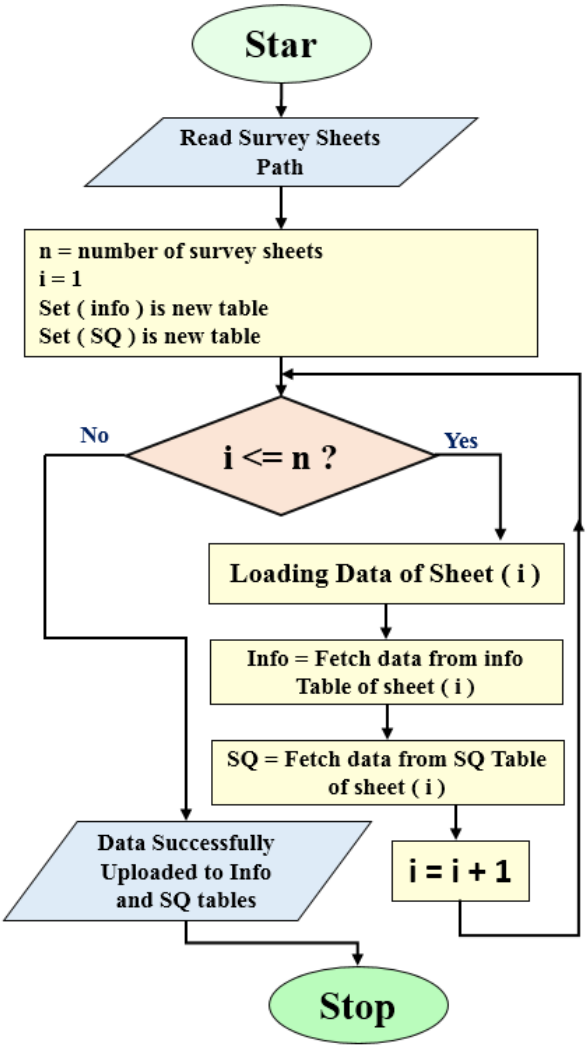

Figure 5. Flowchart of the proposed algorithm.

Based on the SABR procedures, Table 1 and Fig. 6 below show the final view of the generated tables that are based on the 28 persons in a specific group.

Table 1. Information table in R format.

ID	Gender	Age	Marital Status	Education	Education Field	Years At company	Years With Current job
101	Male	28	Single	Master	Technical Degree	4	2
102	Male	30	Married	Master	Medical	3	1
103	Female	29	Married	Bachelor	Technical Degree	5	3
104	Female	42	Married	Doctor	Medical	2	1
105	Male	35	Married	Bachelor	Management	4	2
106	Male	51	Married	Doctor	Engineering	5	3
107	Male	28	Single	Bachelor	Technical Degree	6	5
108	Female	48	Married	Master	Life sciences	3	1
109	Female	35	Single	Master	Engineering	3	2
110	Male	56	Married	Doctor	Management	2	1
111	Male	20	Single	Master	Technical Degree	2	1
112	Female	23	Married	Diploma	Technical Degree	2	1
113	Male	24	Single	Bachelor	Engineering	1	1
114	Female	29	Married	Diploma	Engineering	2	1
115	Female	30	Divorced	Master	Management	3	1
116	Male	32	Married	Bachelor	Technical Degree	4	2
117	Female	30	Married	Diploma	Management	3	1
118	Male	48	Married	Doctor	Medical	5	3
119	Female	53	Married	Doctor	Technical Degree	5	2
120	Male	55	Married	Master	Engineering	4	2
121	Male	57	Married	Master	Engineering	5	3
122	Male	55	Married	Bachelor	Medical	4	2
123	Male	57	Single	Master	Medical	5	2
124	Male	50	Married	Master	Technical Degree	3	1
125	Male	50	Single	Doctor	Technical Degree	4	2
126	Male	28	Single	Doctor	Technical Degree	4	2
127	Male	29	Single	Diploma	Technical Degree	4	2
128	Male	42	Married	Master	Engineering	3	2

After transforming and storing the data in the final station, the designed system automatically runs a package of queries to draw the data shape. The main purpose of this step is to provide a clear picture of the community study because the category of the study has a direct effect on the decision-making process and plays a significant role on the shape of the strategic planning.

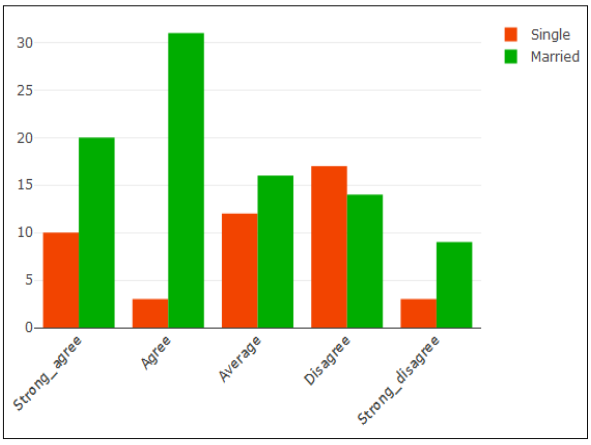

Figure 6. First view of data.

The Figure 6 shows the relationship between the evaluation range and the marital status and it clearly presents the relation. Furthermore, many graphical results are generated in this section by both R and SQL, but only a few is shown.

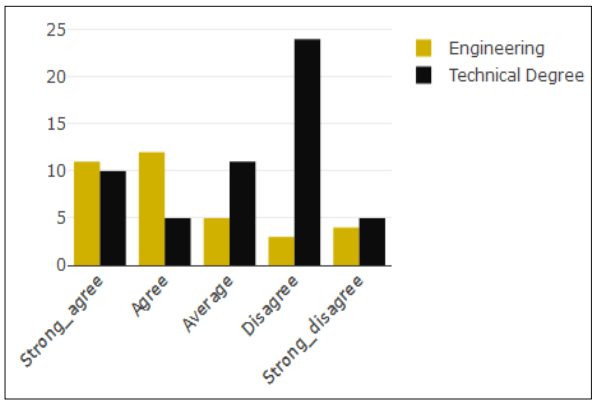

Figure 7. A third view of the data.

The data visualization algorithm is directly linked with the output of the queries. It works automatically according to the input and output and auto adjusts the data sources according to the end objectives. Furthermore, the algorithm includes the Plotly tool that is considered a great package for high data visualization. The data are treated from the relational, non-relational and statistical side and finally visualized and reported according to the final decision as shown in Figure 7.

Simulation Result:

In this section, the realized results are obtained by applying packages of sequence queries written by R and SQL on the generated tables for advance analysing, each query presents a specific situation. Thus, this study proposes a group of queries to provide a basic analysis and a general view of the accumulated data. The proposed work provides more flexibility to the dynamic arrangement of the dataset according to the decision maker's requirements as shown in queries below.

Query1. Assume the decision maker wants to make a relation between the following criteria

Attributes = [Age, Male, Female, Total, Percentage]

Rows = [(18-24), (25-30), (31-40), (41-50), (51-60), Total, Percentage]

Based on the above requirements, the proposed system engine processes the dataset and produces the results as shown in Table 2 to fulfil the requirements of the decision maker.

Table 2. Results based on query one (Query1).

Age	Male	Female	Total	Percentage
18 - 24	2	1	3	11
25 - 30	5	4	9	32
31 - 40	2	1	3	11
41 - 50	4	2	6	21
51 - 60	6	1	7	25
Total	19	9	28	100
Percentage	68	32	100	**%**

Based on the result above, the query 1 output achieves a (formula-1) to measure the accuracy of the proposed algorithm in the transformation of the data from the survey sheets to the desired tables (Information and SQ tables) as shown below

The (formula_1) denoted as (N, T) where:

N: number of survey sheets.

T: the sum of the total numbers of both male and female attributes.

Then

N= 28 sheets.

T = ((Male = 19) + (Female = 9)) / the sum equal (T = 28).

So (N = T) the formula_1 is True.

Query2. Assume the decision maker wants to draw a path to track the unarranged answer sheet by using the ID attribute, the results are shown in Table 3.

Table 3. Results based on query two (Query2).

ID	Gender	Strongly Agree	Agree	Average	Disagree	Strongly Disagree	Total
101	Male	0	2	3	0	0	5
102	Male	1	2	2	0	0	5
103	Female	0	0	3	2	0	5
104	Female	0	0	1	3	1	5
105	Male	2	2	1	0	0	5
106	Male	1	2	0	0	2	5
107	Male	5	0	0	0	0	5
108	Female	1	1	1	1	1	5
109	Female	5	0	0	0	0	5
110	Male	0	3	1	0	1	5
111	Male	0	0	0	2	3	5
112	Female	0	3	2	0	0	5
113	Male	0	1	4	0	0	5
114	Female	5	0	0	0	0	5
115	Female	0	5	0	0	0	5
116	Male	0	0	3	2	0	5
117	Female	0	5	0	0	0	5
118	Male	0	4	1	0	0	5
119	Female	5	0	0	0	0	5
120	Male	0	0	0	3	2	5
121	Male	0	4	1	0	0	5
122	Male	5	0	0	0	0	5
123	Male	0	0	5	0	0	5
124	Male	0	0	0	3	2	5
125	Male	0	0	0	5	0	5
126	Male	0	0	0	5	0	5
127	Male	0	0	0	5	0	5
128	Male	0	5	0	0	0	5

Based on the above results, the Query2 output achieved a (formula-2) as shown below.
The (formula_2) denoted as (N, M, and R) where:

M= (select Id, total from code21 where ID <> 5).
N: number of survey sheets.
R: (select sum (Total) from code21 where ID = 5)
Then
M= 0.
N= sheets.
R = 140.
*So (N * 5 =R) the formula_2 is True.*

If *m* value is not equal to zero, the proposed system automatically returns the ID and tracks to the problem path; based on that a clear picture can be provided to the decision maker about the desired study.

Query3. This query summarizes the basic statistical methods that can be applied in survey fields like the measure of central tendency and the Measure of Dispersion. For example, if a decision maker wants to calculate the frequency, mean and standard deviation for each survey question based on the evaluation degrees from five to one, therefore the results will be as shown in Table 4 and Figure 8.

Table 4. Results based on query three (Query3).

Qn	Strongly Agree	Agree	Average	Disagree	Strongly Disagree	Mean	S.D
q1	6	7	6	5	4	3.21	1.37
q2	5	6	8	5	4	3.11	1.31
q3	6	11	3	8	0	3.54	1.14
q4	6	9	5	7	1	3.43	1.20
q5	7	6	6	6	3	3.29	1.36

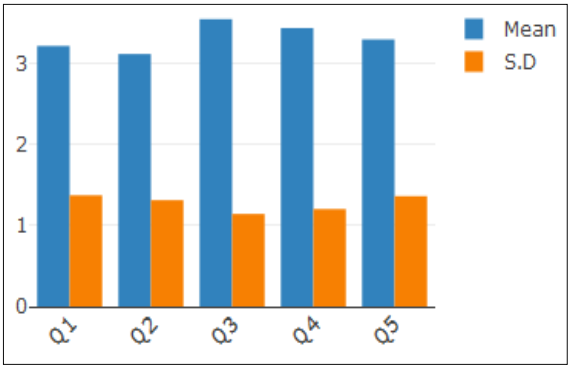

Figure 8. Graphical view of query three (Query3).

The data visualization algorithm is directly linked with the output of the queries. It works automatically according to the input and output and auto adjusts the data sources according to the end objectives. Furthermore, the algorithm includes the Plotly tool that is considered a great package for high data visualization. The data are treated from the relational and statistical side and finally visualized and reported according to the final decision.

The proposed work tackles the requirements and generates the results in both text and graphics formats to build a great environment for understanding and analysing the results. Based on the above results, the Query3 output achieved a formula (Formula-3).

The (Formula_3) denoted as (N, M1, M1, M1, M1, M1 and S) where:
M1= (select sum (Strong_agree) from code26).
M2= (select sum (Agree) from code26).
M3= (select sum (Average) from code26).
M4= (select sum (Disagree) from code26).
M5= (select sum (strong_disagree) from code26).
N: number of survey sheets.
Then
S = (M1+M2+M3+M$+M5).
S = (30+39+28+31+12) = 140.
N= 28.
*So (N * 5 =S) the formula_2 is True.*

Query4. To extend the statistical processes on the desired study, the system proposes to analyze the correlation between the study elements as shown in Table 5 and provides vast opportunities for the decision maker to choose the desired correlation according to his requirements and purpose as shown in Table 6.

Table 5. Results based on query four (Query4).

Criteria	Q1	Q2	Q3	Q4	Q5
Q1	1				
Q2	0.999	1			
Q3	0.787	0.755	1		
Q4	0.881	0.871	0.977	1	
Q5	0.981	0.974	0.851	0.939	1

Table 6. Results based on query four (Query4).

Creiteria_1	Creiteria_2	Correlation
Male	Female	0.91
Age < 30	Age > 40	0.92
Single	Married	0.90
Engineering	Management	0.87
Medical	Engineering	0.91

The biggest challenge in this research is that the statistical methods in the questionnaire systems cannot be executed using the normal formulas because each value has a specific weight and the weight occurs in each calculation. For example, if the decision maker wants to calculate the mean for the first question in Table 4 manually, the process for this step is shown below:

$let\ Xi = [6,7,6,5,4]$

$let\ Wi = [5,4,3,2,1]$

$n = [6 + 7 + 6 + 5 + 4]$

$$mean(xiwi) = \frac{\sum(xi * wi)}{n}$$

$$mean(xiwi) = \frac{((6*5) + (7*4) + (6*3) + (5*2) + (4*1))}{28}$$

$$mean(xiwi) = \frac{90}{28} = 3.21$$

Based on the above discussion, each statistical method is calculated separately and made suitable for the data analysis. The proposed system utilizing SABR uses R and SQL under the R environment to simulate data like a professional program.

Conclusion:

The approach used in this research is based on (SABR) algorithm that studies the technique of data transformation from the main sources to the desired locations based on the R environment and through the combination of two great data science languages, R programming language, and SQL. In a nutshell, the proposed technique allows the application of various statistical methods on the collected data. In this study, the problems of transforming data are resolved using (SABR) algorithm as much as possible. However, further research using the proposed technique needs to be carried out to prove the usefulness of its features and benefits in different fields. Furthermore, an example of a future work related to this study could be the way of combining the questionnaire systems and data mining techniques for the purpose of deep learning and data discovery. The other possible study is how to provide a suitable environment to run The DeLone and McLean Model of Information Systems Success based on R environment supported by SQL. Finally, it would be beneficial to investigate the proposed system in multidimensional and multi-valued databases.

Conflicts of Interest: None.

References:

1. Fetrina E, Rustamaji E, Nuraeni T, Durrachman Y. Inventory management information system development at BPRTIK KEMKOMINFO Jakarta. In2017 5th International Conference on Cyber and IT Service Management (CITSM) 2017 Aug 8 (pp. 1-4). IEEE.
2. Djiroun R, Boukhalfa K, Alimazighi Z, Atigui F, Bimonte S. A data cube design and construction methodology based on OLAP queries. In2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) 2016 Nov 29 (pp. 1-8). IEEE.
3. Chung TD, Ibrahim R, Hassan SM, Rosli NS. Fast approach for automatic data retrieval using R programming language. In 2nd IEEE International Symposium on Robotics and Manufacturing Automation (ROMA) 2016 Sep 25: 1-4. IEEE.
4. Zhang Y, Ordonez C, Cabrera W. Big data analytics integrating a parallel columnar DBMS and the R language. In Cluster, Cloud and Grid Computing (CCGrid), In 16th International Symposium on IEEE/ACM 2016 May 16:627-630. IEEE.
5. Myalapalli VK, Savarapu PR. High performance SQL. In 2014 Annual IEEE India Conference (INDICON) 2014 Dec :1-6. IEEE.
6. Mithani F, Machchhar S, Jasdanwala F. A novel approach for SQL query optimization. In IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016 Dec :1-4. IEEE.
7. Zhou Y, Goto Y, Cheng J. QSL: a specification language for e-questionnaire systems. In 5th IEEE International Conference on Software Engineering and Service Science (ICSESS) 2014 Jun 27 : 224-230. IEEE.
8. Yigzaw KY, Michalas A, Bellika JG. Secure and scalable statistical computation of questionnaire data in r. IEEE Access. 2016; 4:4635-4645.
9. Nakamura S, Ishii T, Akakura T. Item bank to estimate the answers of class evaluation questionnaire. In IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), 2016 Dec 7:150-153. IEEE.
10. Wu J, Zhao D, Lu L, Tian J, Xiang J. A Comparative Study of Knowledge Management on Undergraduate by Questionnaire. In IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C) 2017 Jul 25 :486-492. IEEE.
11. Xu, K., & Feng, J. An information flow based approach to finding informational relationships between questionnaires, 2012.

12. Wibawa J, Widjaja HA, Hidayanto AN. Integrating IS Success Model, SERVQUAL and Kano Model into QFD to improve Hospital Information System quality. In International Conference on Information Management and Technology (ICIMTech) 2016 Nov 16 :29-34. IEEE.

طريقة محسّنة لجلب وتحويل بيانات المسح بناءً على لغتي برمجة SQL و R

زليخة جمال الدين **فرات فالح حسن**

جامعة اوتارا، ماليزيا

الخلاصة:

ساهم تطوير نظم المعلومات (IS) في السنوات الأخيرة في تطوير أساليب مختلفة لجمع المعلومات لتقييم أداء (نظم المعلومات). يسمى الأسلوب الأكثر شيوعًا والمستخدم لجمع المعلومات بنظام المسح. ومع ذلك تعاني هذه الطريقة من خلل كبير. حيث يستهلك صناع القرار وقتًا كبيرًا لتحويل البيانات من أوراق المسح إلى البرامج التحليلية. وعلى هذا النحو، تقترح هذه الدراسة طريقة تسمى (خوارزمية الاستقصاء) تعتمد على لغة برمجةR (وهي لغة برمجة تستخدم اساسا في الحوسبة الاحصائية والرسومات) لتحويل البيانات من أوراق المسح داخل بيئات R عن طريق التعامل مع ترتيب البيانات كتنسيق علائقي. يوفر تنسيق البيانات R و تنسيق البيانات العلائقية فرصة ممتازة لإدارة وتحليل البيانات المتراكمة. علاوة على ذلك، تم تصميم نظام مسح قائم على لغة برمجة SQL و R لتحسين طرق إدارة أنظمة المسح من خلال تطبيق الميزات الكبيرة التي يتم تقديمها من خلال الجمع بين لغات علوم البيانات المتعددة. أثبتت التجارب تعزيزات المرونة والأدوات التقنية وميزات عرض البيانات المستخدمة لمعالجة البيانات التي تم جمعها من جوانب مختلفة ؛ لذلك ، يوضح النهج المقترح دراسة حالة بسيطة لتعزيز متطلبات التقييم للتقنية المقترحة. أخيرًا ، يمكن استخدام النتائج التي توصل اليها البحث لتحسين أساليب إدارة المعلومات في جوانب مختلفة مثل أنظمة المسح ونماذج البيانات الأخرى التي تحتوي على النماذج العلائقية وغير العلائقية باستخدام (خوارزمية المسح على أساس لغة البرمجة R) SABR. حيث أثبتت هذه الطريقة تحسين دقة البيانات التي تم جمعها ، وتقليل وقت معالجة البيانات وترتيب البيانات إلى النموذج الراغب.

الكلمات المفتاحية: تحويل البيانات, ليس فقط لغة الاستعلام الهيكلية, البرمجة بلغة أر ،لغة الاستعلام الهيكلية.