Received October 3, 2021, accepted November 3, 2021, date of publication November 10, 2021, date of current version January 12, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3127207

Automatic Generation of a Metric Report: A Case Study of Scientometric Analytics

AMAL BABOUR[1] AND JAVED I. KHAN[2], (Member, IEEE)

[1]Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Science, Kent State University, Kent, OH 44240, USA

Corresponding author: Amal Babour (ababor@kau.edu.sa)

ABSTRACT Automatic report generation is an emerging technology that mechanically generates documents in the form of a report consisting of text, tables, and figures about a specific topic. This paper proposes a model for automated generation of a scientometric research analysis report for any selected country. The scholarly database utilized in this study is Microsoft Academic Graph. Given the two-letter code of the selected country, starting study year, and ending study year data concepts to the employed database, the model extracts the datasets, including the scientific research information for the selected country in the specified period, and generates an in-depth analysis report about the country's research publications. The model consists of two main stages. The first stage extracts the datasets for the selected country from the utilized database using Azure Databricks and Azure Blob Storage services. The second stage utilizes a predefined scientometric research analysis report for generating a new report for the selected country. A case study on big data analysis for Saudi Arabian research publications was conducted. An evaluation was performed on the report within 10 evaluators to understand the practicability of the proposed model. They evaluated the report through four-point criteria on the user perceptions. The results indicated that the model was able to successfully create quite a pleasing scientific report in terms of factuality, coherency, sufficiency, and its ability to impart new knowledge for the readers.

INDEX TERMS Automated report generation, big data, coefficient collaboration, collaborative research, growth rate, scientometrics.

I. INTRODUCTION

Big Data analysis has gained enormous attention in recent years. It can be defined as a process of gathering data from different sources, organizing them in a meaningful way and then analyzing these data to uncover underlying facts and figures from that data collection [1].

Massive scientometric databases have recently created inspiration for much big data analytics and have been a treasure trove of insight ranging from scientific productivity and the science of collaboration to policy design. Scientometrics is a discipline which analyses scientific publications to explore the structure and growth of science. It is a scientific field that studies the evolution of science through quantitative measures of scientific information. It is important for measuring the growth of research, research quality, author productivity, obsolescence of publications, and collaborative research by country, all of which help to monitor the growth and pattern of research [2].

The scientometric research study can be made by analysis at a micro level (that involves authors analysis); a meso level (that includes institutions analysis); and/or a macro level (that includes countries analysis). The micro level shows the high active authors in a specific field and how they collaborate with each other through co-authorship. The meso and macro levels show deep insight about productive institutes and countries and how they collaborate with each other [3]. Together, they have been used to study the research trends in the past and predict the directions in patent and publications in the future.

For some researchers, writing such a scientific report that presents any single or combined level of analysis in an organized and professional manner is a challenging task. Text should be clear and interesting to the reader. Claims must be supported with reliable evidence. Figures and tables should be placed in the right place. Data should be presented clearly. Experimental details should be reported in enough details.

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Abid.

All of these concerns require a highly developed level of writing skills [4], [5]. With the rapid growth of information and the emergence of big data, a new challenge in the life of reporting data analysis needs to be developed.

To the best of the research team's knowledge, no existing work proposing a model for generating scientometric research analysis report had been proposed prior to this study. Accordingly, this paper introduces a model to generate an automated scientometric analysis report. The aim is to automatically generate a report that presents a comprehensive view of research by illustrating all major aspects of study collaboration on micro, meso, and macro levels, including a study of period-wise research output for each field of study and growth rate, authorship and collaboration pattern, major collaborative countries, and citation rates for any selected country. The main contributions of this paper are the following:

- designing a model for generating scientometric research analysis report.

- developing and evaluating the proposed model.

The rest of the paper is organized as follows. The related work is provided in Section 2. The proposed methodology is presented in Section 3. The case study is presented in Section 4. The evaluation results are described in Section 5. Finally, the conclusion and future work are presented in Section 6.

II. RELATED WORK

With the development of document automation technology, automatic report generation has found an increasingly wide utilization in many fields [6]–[10]. It can be used as a tool to lighten the load of writing frequent reports, especially the daily reports that carry the same patterns. It can also be utilized to present daily news, medical reports, students' feedback, and monitoring. In this section, several existing systems that generate automated generation reports in different fields are reviewed.

MIT SciGEN was an interesting experiment which attracted significant attention to the auto-generation of scientific papers. However, its goal was to generate seemingly random papers which had the appearance of a research paper in the field of computer science, but with no meaning to cleverly test the review process bogus conferences than generating high value papers [6].

Noh *et al.* [7] proposed an automated report generation system called Wise Issue Reporter (WIRE) that generates two types of well-summarized reports for emerging topics: Today's Briefing and Full Report. The first report provides five brief summaries for the five hot daily topics. It uses the Elasticsearch[1] as an IR engine to select the five topics from the topic pool and the neural multi-document summarization model (MDS) model [11] to generate an abstractive summary for each of the topics. The second report provides an in-depth report of a selected topic from the five daily hot topics. It uses the IR engine to return the related articles of the selected topic. Then, it applies the MDS model and the temporal-event summarization model (TES) [12] to generate two summaries that are combined to create the Full Report. The output of the system is a text report which consists of a number of paragraphs and an image related to the report topic selected from Google Image by using an image recommendation system.

Gkatzia and Hastie [8] presented a model for automatically generating a feedback report to students that describes their performance in a specified semester. The produced report is a feedback text summary based on trends in time series data. Its content can be formulated as a classification task based on a set of factors (marks, hours studied, understandability, difficulty, deadlines, health issues, personal issues, lectures attended and revision), where each factor can have a set of templates by different evaluators. A multi-label classifier (MLC) [13] is used for tackling the content selection and making a decision for all the templates. The selected content is presented as a text summary.

Mass *et al.* [9] created a Care2Report (C2R) system that generates automated medical consultation reports. The system combines the hardware of a camera, a microphone, and sensor technology with advanced software. The system transforms the medical dialogues, examinations, treatments and sensor data captured by the hardware into text by using speech and action recognition technology. It then extracts the concepts and relations from the text in a form of triples (subject, predicate, object) using Python-Frog and a FRED tool and adds the triples to prebuild an ontology of human anatomy, medical signs, and symptoms to populate a knowledge graph. Using the NaturalOWL system [14], it generates an automatic medical consultation report text from the knowledge graph.

Morita *et al.* [10] proposed a system called Movable-Beacon and Fixed-Scanner (MBFS) that monitors nursing home residents' activities and generates an automatically daily report for each resident. The system consists of four components: a BLE beacon, a scanner, a server, and a web application. The BLE beacon is carried by a resident. A scanner is installed in each target area where activities occur. The system determines the specific activities areas by the BLE signals observed by each scanner. The data for each resident is then sent from each scanner to the server database. Machine learning techniques were then used to classify the area of the residents based on their received data in the server. The system automatically generates a daily report for each resident via a web application which shows a combination between time windows of the time spent in all the nursing home areas.

On the other hand, several countries have been conducting scientometric analyses to evaluate their research productivity in a general or in a specific field at both a local and a global scale [15]–[20].

Jalal revealed [15] the international collaboration in publications between India and Bangladesh based on a Web of Science (WoS) database from 1991 to 2017. He focused on Scientometric tools such as the yearly growth of international collaboration in publications, preferred journals,

[1] Elasticsearch (https://www.elastic.co/)

authorship patterns, and the countries which collaborated the most between the two countries. Likewise, Nguyen *et al.* [16] explored the international collaboration in scientific research in a WoS database in Vietnam during the period from 2001 to 2015. They studied the collaborative research in terms of author collaborations, the number of publications in each field of study, collaborating countries, and citation rates.

Low *et al.* [17] investigated the pattern of research collaboration in scholarly publications in the field of clinical medicine in Malaysia during the period from 2001 to 2010. The bibliographic data were downloaded from the journal of clinical medicine covered in the ISI database during the study period. The authors studied the collaboration patterns in terms of journal impact factor, citation impact of domestic and international publications, and collaborating countries. Farooq *et al.* [18] explored the pattern of research in the field of energy indexed in the Scopus database in Pakistan from 1990 to 2016. Their study revealed the category of the publications, the geographical distribution of the research collaborations around the globe, the contributions in the productive journals, citations, major research areas, authorship patterns, active researchers, and the most productive institutes in the field of energy research in Pakistan. Zha *et al.* [19] studied the characteristic of research collaboration in the field of biotechnology in the Science Citation Index journals indexed in the WoS database in China from 1991 to 2014. The study explored the collaboration networks of the global academic institutions in the field of biotechnology, and collaboration types, collaboration degree, and collaboration fields of the Chinese academic institutions in biotechnology. Fakhree and Jouyban [20] compared scientific publications published by the researchers of 7 major Iranian medical universities from 2004 to 2009 and indexed in the Scopus database. Their comparison was in terms of the number of publications per year, the number of citations per year, the number of citations per year for each publication, H-indices, top ten authors, and top ten journals for the medical universities.

The work in this paper combines the automatic report generation and the scientometric analysis approaches in order to generate an automatic scientometric research analysis report for any selected country.

III. METHODOLOGY

In this paper, a model for an automatic report generation that gives a comprehensive insight about a country status in science publications during a specific period of time is proposed. The method used in the model reads several features from a report requester such as the country's name, classification of the country, and scope of the start and the end of publications. It reads the research publication data, including affiliations belonging to the selected country and performs an insightful scientometric analysis about the publications using a scholarly database. Then, it generates a high-level report, presenting information about the country's overall research status in the given period of time. The main content of the report is made up of text, figures, and tables illustrating details regarding the research performance of the selected country in the given time zone.

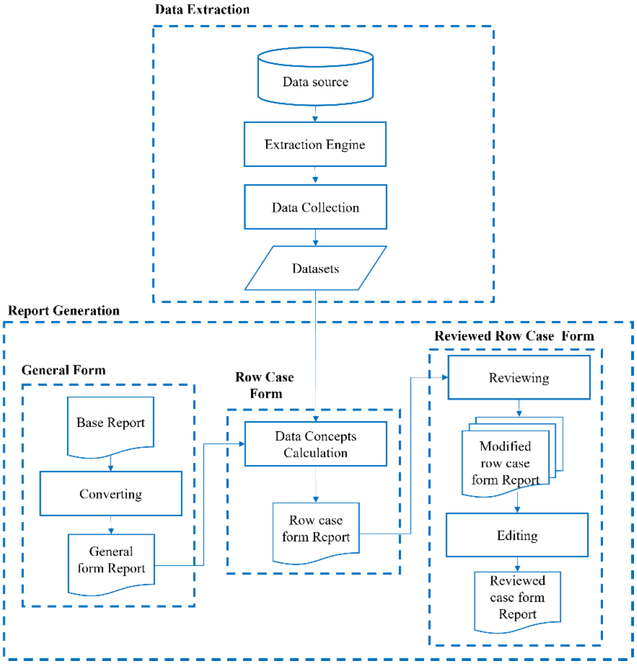


FIGURE 1. Generating automated report stages.

The proposed methodology generates an automated report similar to the one created for Vietnam [16] for any selected country. The report identifies the total publications released in each respective selected country, the number of publications in the fields of study, and the growth rate. It also presents the coefficient of collaboration, the collaboration status in each field of study, and the top collaborative countries. In addition, it shows the average citation classified by the fields of study and the collaborative status, and by each field of study and the first authors' affiliation. In this study, the Vietnam scientific research report is called the based report and the automated generated scientific research report is the case report.

The proposed methodology consists of two main stages: 1) the data extraction; and 2) the report generation; The overview of the two stages is shown in Fig.1.

A. DATA EXTRACTION

1) DATA SOURCE

The data repository employed in the proposed method is Microsoft Academic Graph (MAG) [21]. MAG is the largest free scholarly database licensed under ODC-BY 1.0 published by Microsoft. It provides information about papers, authors, journals, conferences, affiliations, and citations. It receives regular updates every 1 or 2 weeks, providing more than 250 million scientific publications as of January 2021 [22], [23]. It was chosen because it has the largest coverage of publications, including journals, conferences, books, and patents, when compared to WoS, Scopus, Dimensions, and CrossRef. In addition, it offers free and unrestricted access to the complete list of publications [24].

In this study, five datasets from MAG are utilized: Affiliation, PaperAuthorAffiliations, Papers, PaperFieldsofStudy, and FieldofStudy. The Affiliation dataset records institution-related information such as affiliation name and country. The PaperAuthorAffiliations dataset lists information about the authors and the affiliations of each paper such as paper id, author id, affiliation id and author sequence number. The Papers dataset consists of paper information such as title, digital object identifier (DOI), publication year, and publisher. The PaperFieldsofStudy dataset records information about the fields of study for each paper such as field of study id. The FieldofStudy dataset includes detail information about each field of study such as the field of study name and level. The database schema among the utilized datasets from MAG are presented in Fig. 2.

2) EXTRACTION ENGINE

Two Microsoft Azure services are utilized to extract the research publication data for the selected country: Azure Databricks and Azure Blob Storage. Azure Databricks is a cloud Apache Spark programming platform used to process big data and supports code written in Python, R Scala, Spark, and SQL. Azure Blob Storage is a cloud storage system optimized for storing massive amounts of unstructured data. It was used for the storage of MAG datasets and for the storage of the extracted datasets for the country's research publications. Microsoft Academic distributes the database for free through Microsof's Azure Storage, but they charge for storage of the data and any computation done on the Azure platform [22], [23]. When working with Databricks, the configuration of the Spark Cluster affects the performance of the algorithm and the executing time. For the scope of this paper, the suggested configuration of the clusters created on Databricks is presented in Table 1.

3) DATA COLLECTION

Given the two-letter code of country X, starting study year S, and ending study year E, find all the publications included affiliations belong to X between the period S and E. In this research, the code for extracting the data was written in Python 3.7. Fig. 3 presents the pseudocode for getting the paper ids for country X. The input of the algorithm consists of the *Affiliation* and *PaperAuthorAffiliations* datasets. *Iso3166Code* is a code published by the International Organization for Standardization (ISO) that defines two letter' codes for the names of countries. It returns a dataset *PaperIds_X* of paper ids that have at least one author belongs to an affiliation located in country X.

The pseudocode for getting the publication year and the number of citations for the country X papers is presented in Fig. 4. The input of the function is *PaperIds_X, S, E,* and the *Paper* dataset. The output of the algorithm is a dataset *Paper_X_info* which consists of id, year, and citation count of the country X's papers.

Fig. 5 presents the pseudocode for getting the field of study for country X's papers. The input of the algorithm consists of *PaperId_X*, *PaperFieldsOfStudy*, and *FieldOfStudy*. The algorithm in lines 1 and 2 receives the ids of all fields of study to which each paper belongs. To avoid classifying the paper into more than one field, the algorithm from lines 3 to 6 selects the field of the study name of level 0, as it is considered the abstract classification of the paper. The algorithm returns a dataset *Paper_X_fos* of country X's papers and their fields of study.

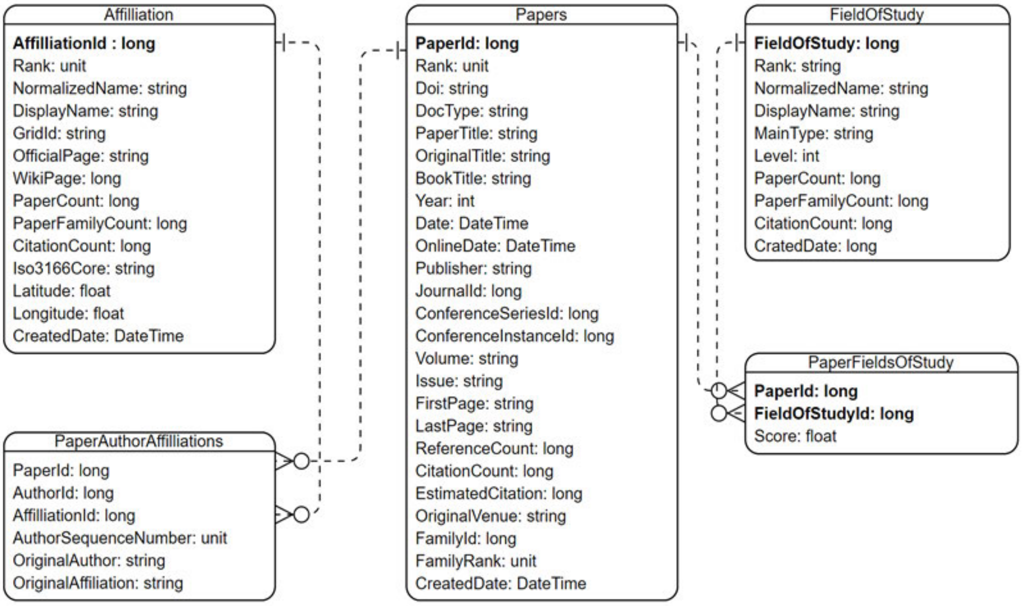


FIGURE 2. Subset of MAG database schema.

TABLE 1. Configuration of the created cluster on Databricks.

Databricks Runtime Version	9.0 (includes Apache Spark 3.1.2, Scala 2.12)
Worker Type	14 GB Memory, 4 Cores
Driver Type	14 GB Memory, 4 Cores
Number of minimum workers	2
Number of maximum workers	8

Getting the number of author participation in each paper is presented in Fig. 6. The input of the algorithm consists of the *PaperId_X* and *PaperAuthorAffiliations* datasets. The output is a dataset *Paper_X_author_count* of country 's papers and the number of authors for each.

According to the number of authors for each paper, the pseudocode in Fig.7 presents the classification of paper types. The algorithm input consists of the *Paper_X_author_count*, *PaperAuthorAffiliations* and *Affiliations* datasets. In line 5 and 6, the algorithm checks if the number of authors in a paper is 1. If this occurs, the paper is classified as a single author paper. From lines 7 to 12, if there is more than one author for a paper, the algorithm receives the affiliations for all authors participating in the paper and checks if all belong to country *X* affiliations. If this occurs, the paper is classified as a domestic paper. Otherwise, if at least one author does not belong to county *X* affiliations, the paper is classified as an international paper. The algorithm returns a dataset *Paper_X_type* of country *X*'s papers and their types.

The affiliation country of the paper' authors is considered one of the most important factors in the citation purposes. Fig. 8 presents the pseudocode for getting the country of the author's affiliation. The algorithm input is *PaperId_X* and *PaperAuthorAffiliations* and *Affiliations* datasets. The algorithm returns a dataset *Paper_X_authors_affiliation* of country X papers, along with the author sequence number and the affiliation country for each author.

By generating the *Paper_X_info*, *Paper_X_fos*, *Paper_X_author_count, Paper_X_typw*, and *Paper_X_authors_affiliation* datasets, the country X scientific research data is ready for further analysis.

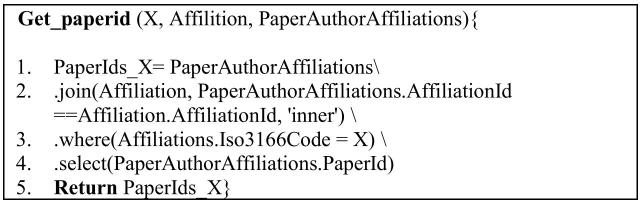
B. REPORT GENERATION

1) GENERAL FORM

The purpose of this stage is to convert the based report into a general scientific research report that can be used to present the scientific research data for any selected country. In this stage, all data concepts related to the country of Vietnam, whether quantitative or qualitative, and the used scholarly publication database are manually extracted and replaced with data concept variables in the based report. Fig. 9 shows an example of multiple types of data in the based report before and after replacing them with data concepts to generate the general form report.

2) ROW CASE FORM

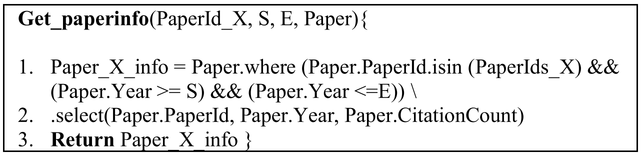
The data concepts assigned in the entire report are classified into three groups.



```
Get_paperid (X, Affilition, PaperAuthorAffiliations){

PaperIds_X= PaperAuthorAffiliations\
.join(Affiliation, PaperAuthorAffiliations.AffiliationId
==Affiliation.AffiliationId, 'inner') \
.where(Affiliations.Iso3166Code = X) \
.select(PaperAuthorAffiliations.PaperId)
Return PaperIds_X}
```

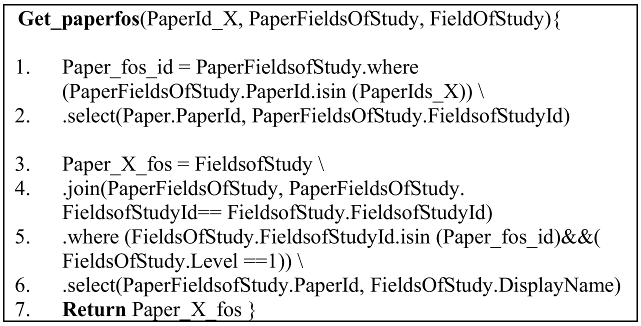
FIGURE 3. Getting country X paper ids pseudocode.



```
Get_paperinfo(PaperId_X, S, E, Paper){

Paper_X_info = Paper.where (Paper.PaperId.isin (PaperIds_X) &&
(Paper.Year >= S) && (Paper.Year <=E)) \
.select(Paper.PaperId, Paper.Year, Paper.CitationCount)
Return Paper_X_info }
```

FIGURE 4. Getting country X papers information pseudocode.

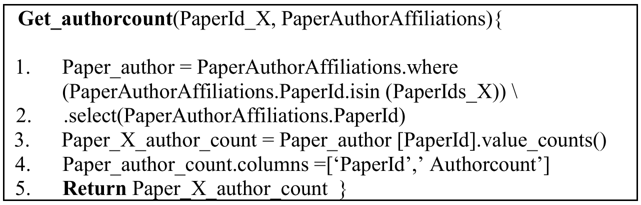


```
Get_paperfos(PaperId_X, PaperFieldsOfStudy, FieldOfStudy){

Paper_fos_id = PaperFieldsofStudy.where
(PaperFieldsOfStudy.PaperId.isin (PaperIds_X)) \
.select(Paper.PaperId, PaperFieldsOfStudy.FieldsofStudyId)

Paper_X_fos = FieldsofStudy \
.join(PaperFieldsOfStudy, PaperFieldsOfStudy.
FieldsofStudyId== FieldsofStudy.FieldsofStudyId)
.where (FieldsOfStudy.FieldsofStudyId.isin (Paper_fos_id)&&(
FieldsOfStudy.Level ==1)) \
.select(PaperFieldsofStudy.PaperId, FieldsOfStudy.DisplayName)
Return Paper_X_fos }
```

FIGURE 5. Getting country X papers field of study pseudocode.



```
Get_authorcount(PaperId_X, PaperAuthorAffiliations){

Paper_author = PaperAuthorAffiliations.where
(PaperAuthorAffiliations.PaperId.isin (PaperIds_X)) \
.select(PaperAuthorAffiliations.PaperId)
Paper_X_author_count = Paper_author [PaperId].value_counts()
Paper_author_count.columns =['PaperId',' Authorcount']
Return Paper_X_author_count  }
```

FIGURE 6. Getting country X papers authors count pseudocode.

Group 1

This group includes a set of data concepts given by the report designer.

Group 2

The data concepts in this group are given by the report requester.

Group 3

This group includes the data concepts that need to be calculated or concluded from the generated datasets. All types of data concepts are calculated automatically. Some data concepts have to be calculated before others. For example, the total number of publications for each study period must be calculated before calculating the rate of growth. Some data concepts need to be concluded from others. For example, it can be concluded from the growth rate values if it increased or decreased during the study period. Thus, data concepts in this group are classified into six sub-groups, according to their relations and the sequence of their calculations.

```
Get_papertype(Paper_X_author_count, PaperAuthorAffiliations,
Affiliation){

    Affiliations_X= Affilition.where (Affiliation.Iso3166Code = X) \
    .select(Affiliation.AffiliationId)
    Paper_X_type=[]
    for p in Paper_X_author_count:
        if p. Authorcount ==1:
            Paper_X_type.add(p.PaperId, 'single_authored_paper')
        else:
            author_affiliation= get_authors_affiliattions(p.PaperId,
    PaperAuthorAffiliations)
            if all(author_affiliation (is in(Affiliations_X)))
                Paper_X_type.add(p.PaperId, 'domestic_paper')
            else
                Paper_X_type.add(p.PaperId, 'International_paper')
    Return Paper_X_type }
```

FIGURE 7. Getting country X papers type pseudocode.

```
Get_authors_affiliation(PaperAuthorAffiliations, Affiliations,
PaperIds_X){

    paper_X_authors_affiliation = PaperAuthorAffiliations \
    .join(Affiliation, PaperAuthorAffiliations.AffiliationId ==
    Affiliation.AffiliationId, 'inner') \
    .where (PaperAuthorAffiliations.PaperId.isin (PaperIds_X))
    .select(PaperAuthorAffiliations.PaperId,
    PaperAuthorAffiliations.AuthorSequenceNumber,
    Affiliations.Iso3166Code)
    Return paper_X_authors_affiliation }
```

FIGURE 8. Getting country X papers authors affiliation pseudocode.

a: FIELD OF STUDY (FOS)

The data concepts in this group are about the number of publications for each field of study.

b: GROWTH RATE (GR)

This group includes data concepts about the publications' growth rate (GR) for each field of study. The rate of growth is estimated by Compound Annual Growth Rate (CAGR). Given that n is the number of the study period years, C is number of publications in the ending year, and Y is number of publications in the beginning year, CAGR can be calculated by (1) [25].

$$CAGR = \left(\frac{C}{Y}\right)^{\left(\frac{1}{n-1}\right)} - 1 \tag{1}$$

c: COEFFICIENT COLLABORATION (CC)

This group includes data concepts about the coefficient collaboration (CC) for each field of study in each study period. In this paper, CC is calculated by (2) [26], where P_j is the number of publications that have j authors in the research area, A is the greatest number of authors in the research area, and N is the total number of publications in the research area.

$$CC = 1 - \frac{\sum_{j=1}^{A}\left(\frac{1}{j}\right)P_j}{N} \tag{2}$$

d: PUBLICATION TYPE

This group includes data concepts about the publication types for each field of study (single authored paper, domestic paper, or international paper).

e: COLLABORATIVE PUBLICATIONS

The data concepts in this group include information about the collaborative publications with the international countries.

f: CITATION

This group covers information about the citation of domestic/international collaborations in each field of study.

Table 2 presents examples of the data concepts and their appropriate groups. The complete list of the data concepts is presented in the Appendix.

By using the extracted datasets for country *X,* the data concepts for group 3 are calculated. The sequence of calculating the Group 3 data concepts is shown in Fig. 10. The figures are constructed using a combination of the data concepts calculated in Group 3. Three figures were considered in the generated report. The first figure is a bar chart showing the percentage of publications of each field of study for the selected country. This can be formed by utilizing data concept values from the field of study group. The second figure is a network structure visualizing the international collaborations between the selected country and other countries. This figure can be drawn from data concept values in the collaborative publications group. The last figure is a set that shows the distribution of citations classified by both the field of study and collaboration types. This figure can be formed using data concepts from the citation group.

After finding the data concept values, the data concepts of each group and their values are saved in a.csv file. By using the MS Word "mail merge" feature available in MS Office and importing the.csv file, all the data concepts in the general form report are substituted by the assigned values generating the row case form report. The figures in the report are uploaded by the report requester.

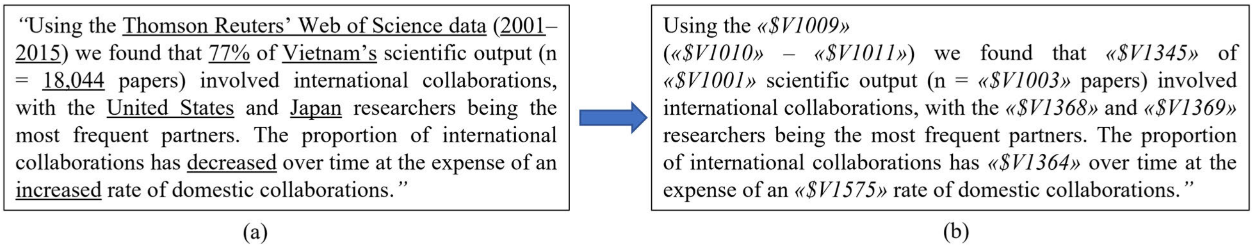
3) REVIEWED CASE FORM

In this stage, the row case form report is given to three reviewers to check the report's consistency. They are asked to classify each sentence in the row case report as relevant/irrelevant, correct/incorrect, and interesting/not interesting, and to specify the type of problems for the sentences classified as irrelevant, incorrect, and/or not interesting. They are also asked to suggest solutions for these problems and, if possible, create new sentences for the ones with problems.

After the reviewers complete their tasks, they send the suggested solutions and the newly created sentences to the editor. The editor reconciles among the new sentences, updates the row case form, and generates the reviewed case form.

IV. CASE STUDY

This section describes the case study using the proposed method to generate the reviewed case form. Given the data concepts by the report requestor, the country of Saudi Arabia (SA) was selected as a case study, and the study period ranged from 2001 to 2020. Using the MAG database and the suggested Azure services, a dataset of



"Using the Thomson Reuters' Web of Science data (2001–2015) we found that 77% of Vietnam's scientific output (n = 18,044 papers) involved international collaborations, with the United States and Japan researchers being the most frequent partners. The proportion of international collaborations has decreased over time at the expense of an increased rate of domestic collaborations."

(a)

Using the *«$V1009»* (*«$V1010»* – *«$V1011»*) we found that *«$V1345»* of *«$V1001»* scientific output (n = *«$V1003»* papers) involved international collaborations, with the *«$V1368»* and *«$V1369»* researchers being the most frequent partners. The proportion of international collaborations has *«$V1364»* over time at the expense of an *«$V1575»* rate of domestic collaborations."

(b)

FIGURE 9. Example of the data in the based report before (a) and after replacing them by data concepts (b).

181,130 publications were extracted. After removing records that have incomplete information such as missing year and field of study, 158,860 publications were considered in the statistical analysis.

All the data concepts for Group 3 were calculated and analyzed using Python 3.7 and MS Excel in Office 365. The data concepts' values for the three groups were inserted into the row case form using the "mail merge" feature from MS Word in Office 365. Table 3 shows some of the data concepts' values for the country of Saudi Arabia. The used reviewed row case form and all the case study data concepts values for the three groups are shown in the Appendix.

The generated row case form was evaluated by three reviewers. Among 200 sentences in the row case form report, an average of 55% of the sentences were found that had problems and needed to be adjusted. The problems in the sentences were divided into five types. The first type appeared in the sentences with information about the country of Vietnam and its history. The second type appeared in the sentences in which information was presented about the used

TABLE 2. Data concepts groups and descriptions.

Group	Data Concept	Data Concept description
Group 1	$V1009	The name of the scholarly database used in the study.
	$V1004	The number of the entire publications in the used database.
Group 2	$V1001	The selected country.
	$V1580	The classification of the selected country (developed/ developing).
	$V1010	Starting study year.
	$V1011	Ending study year.
Group 3.1	$V1266	Number of papers in field of study n.
	$V1083	Number of papers in field of study n for study period m.
	$V1040	Percentage of publications of the top five field of study.
	$V1362	The difference in publications between last and first study period.
Group 3.2	$V1163	Percentage of growth rate for field of study n.
	$V1601	The field of study has the strongest growth rate.
	$V1182	Percentage of the growth rate between the starting and the ending study year.
Group 3.3	$V1187	Coefficient collaboration in field of study n for study period m.
	$V1206	Coefficient collaboration in all field of study for study period m
	$V1597	The status of coefficient collaboration during all study periods.
Group 3.4	$V1286	Percentage of single authored paper in field of study n
	$V1306	Percentage of domestic paper in field of study n
	$V1326	Percentage of international paper in field of study n
	$V1887	The field of study has the highest international collaborations.
Group 3.5	$V1368	The first international collaborative country.
	$V1889	First international developed country has the highest number of collaborations.
	$V1388	Total number of published papers with the first international collaborative country in study period m
	$V1468	Total number of published papers with the first international collaborative country in all study period.
Group 3.6	$V1515	Percentage of citation of domestic collaboration in field of study n
	$V1534	Percentage of citation of international collaboration in field of study n
	$V1610	Field of study has the highest citation in the international publications.
	$V1624	Percentage of international paper has international first author in field of study n

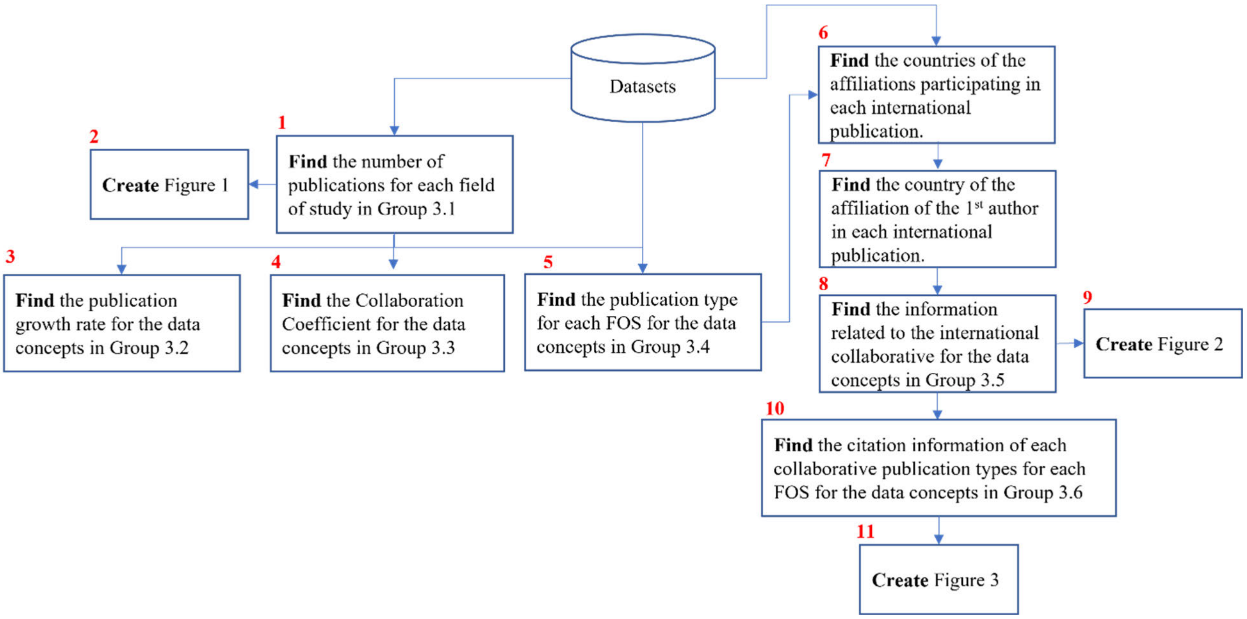

Datasets
1
Find the number of publications for each field of study in Group 3.1
2
Create Figure 1
3
Find the publication growth rate for the data concepts in Group 3.2
4
Find the Collaboration Coefficient for the data concepts in Group 3.3
5
Find the publication type for each FOS for the data concepts in Group 3.4
6
Find the countries of the affiliations participating in each international publication.
7
Find the country of the affiliation of the 1st author in each international publication.
8
Find the information related to the international collaborative for the data concepts in Group 3.5
9
Create Figure 2
10
Find the citation information of each collaborative publication types for each FOS for the data concepts in Group 3.6
11
Create Figure 3

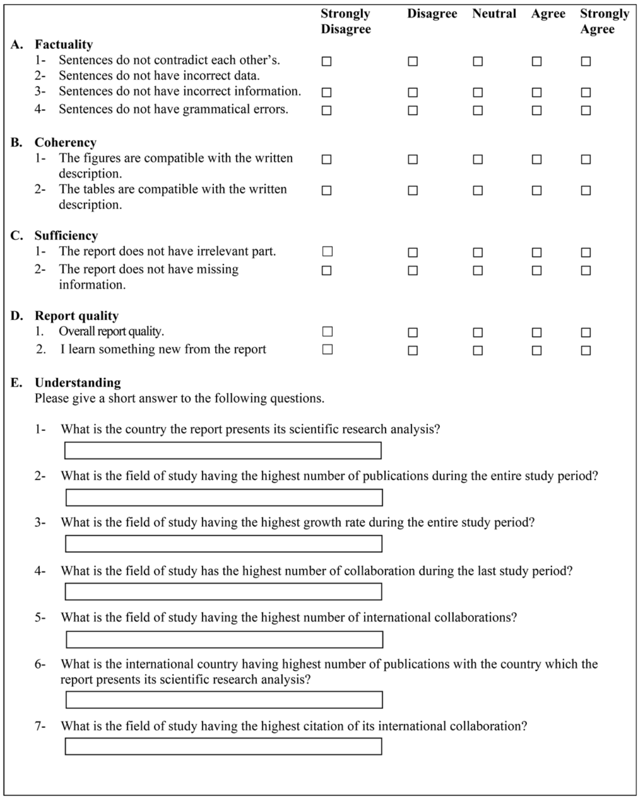
FIGURE 10. Sequence of calculation of groups values.

database. The third type appeared in the sentences which presented information about the type of publications used in the analysis. The fourth type of problem appeared in the sentences which contained deep information, where the last type appeared in the sentences that referred to the tools used for analysis. The reviewers suggested three types of solutions: adding data concepts for the information related to the country or the history of the country; rephrasing the sentences which have irrelevant or incorrect information; and/or removing the sentences. After the reviewers sent their solutions to the editor, the editor reconciled among the suggested solutions and generated the reviewed case form, which is the final generated report. Table 4 presents examples about the problem types and their solutions, where rendering sentence-1 presents the sentences in the row case form and rendering sentence-2 presents the suggested sentences by the reviewers to be in the reviewed case form. In the table, the problem is underlined in rendering sentence-1, and the solution is written in bold in rendering sentence-2.

V. EVALUATION

A. EVALUATION MECHANISM

In order to evaluate the generated report, the classical peer review was mimicked by an expert model used by most scientific publications, albeit slightly modified for scientific objectivity. (A typical review process does not assess the knowledge of the reader). An evaluation questionnaire using Qualtrics Survey API was designed and distributed online. The questionnaire inclusion criteria included evaluators who could read and understand English, who had technical writing skills, who were interested in scientometric research analysis, and who had completed the survey. A single response per evaluator was requested. The questionnaire asked the



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
A. Factuality					
1- Sentences do not contradict each other's.	□	□	□	□	□
2- Sentences do not have incorrect data.					
3- Sentences do not have incorrect information.	□	□	□	□	□
4- Sentences do not have grammatical errors.	□	□	□	□	□
B. Coherency					
1- The figures are compatible with the written description.	□	□	□	□	□
2- The tables are compatible with the written description.	□	□	□	□	□
C. Sufficiency					
1- The report does not have irrelevant part.	□	□	□	□	□
2- The report does not have missing information.	□	□	□	□	□
D. Report quality					
1. Overall report quality.	□	□	□	□	□
2. I learn something new from the report	□	□	□	□	□

E. Understanding
Please give a short answer to the following questions.

1- What is the country the report presents its scientific research analysis?

2- What is the field of study having the highest number of publications during the entire study period?

3- What is the field of study having the highest growth rate during the entire study period?

4- What is the field of study has the highest number of collaboration during the last study period?

5- What is the field of study having the highest number of international collaborations?

6- What is the international country having highest number of publications with the country which the report presents its scientific research analysis?

7- What is the field of study having the highest citation of its international collaboration?

FIGURE 11. Report evaluation questionnaire.

evaluators to evaluate the report on four key criteria: A. *Factuality* (F); B. *Coherency* (C); C. *Sufficiency* (S); D. *Overall Quality* (Q). Each of the criteria includes a set of question measurements. Factuality measures focus on the correctness and accuracy defects. Coherency measures check for logical

TABLE 3. Example of data concepts/values for Saudi Arabia country.

Group	Data Concept	Data concept value
Group 1	$V1009	Microsoft Academic Graph (MAG)
	$V1004	244,000,000
Group 2	$V1001	Saudi Arabia
	$V1580	Developing
	$V1010	2001
	$V1011	2020
Group 3.1	$V1266	28872
	$V1083	544
	$V1040	73%
	$V1362	84751
Group 3.2	$V1163	23.18
	$V1601	Political Science
	$V1182	17.16%
Group 3.4	$V1286	7.32%
	$V1306	22.70%
	$V1326	69.98%
	#V1887	biology
Group 3.5	$V1368	United States
	$V1889	United States
	$V1388	454
	$V1468	25158
Group 3.6	$V1515	23.45%
	$V1534	67.29%
	$V1610	Biology
	$V1624	81.40%

soundness and flow. Sufficiency measures focus on obvious superfluity or gap perceptions in the reading. Overall Quality measures check for delivering new knowledge to the readers. A fifth control criterion, User Understanding (U), was also included to evaluate whether the particular evaluator had read the report or not. Response styles included a five-point Likert scale ranging from ''Strongly disagree'' to ''Strongly agree'' which was used for criteria A, B, C, and D; and short answers for criterion E (as shown in Fig. 11). The evaluators were asked to read the reviewed case report and fill out the questionnaire after reading. For the measurements of the first four criteria, a score of 1 to 5 was given from strongly disagree to strongly agree. For the understanding, a score of 1 was given for a correct answer and a score of zero for an incorrect answer.

The score for U criterion for each participant was calculated by (3.A), where M is the number of measurements in U. $u_{i,m}$ denotes the score of responded i for answer of question m. The net understanding of evaluator i is the average score.

$$\boldsymbol{U_i = \sum\nolimits_{m=1}^{M} u_{i,m}}/\mathrm{M} \tag{3.A}$$

Understanding criterion was chosen as a checkpoint to assess the evaluators and used as a weight for F, C, S, and Q. These criteria are computed by (3.B, 3.C, 3.D, and 3.E), where P is the number of evaluators. $f_{i,m}$, $c_{i,m}$, $s_{i,m}$, or $q_{i,m}$ is the score of evaluator i for measurement m.

$$\boldsymbol{F = \frac{\sum_{m=1}^{4}\sum_{i=1}^{P} U_i * f_{i,m}}{\sum_{i=1}^{P} U_i * 4}} \tag{3.B}$$

$$\boldsymbol{C = \frac{\sum_{m=1}^{2}\sum_{i=1}^{P} U_i * c_{i,m}}{\sum_{i=1}^{P} U_i * 2}} \tag{3.C}$$

$$\boldsymbol{S = \frac{\sum_{m=1}^{2}\sum_{i=1}^{P} U_i * s_{i,m}}{\sum_{i=1}^{P} U_i * 2}} \tag{3.D}$$

$$\boldsymbol{Q = \frac{\sum_{m=1}^{2}\sum_{i=1}^{P} U_i * q_{i,m}}{\sum_{i=1}^{P} U_i} * 2} \tag{3.E}$$

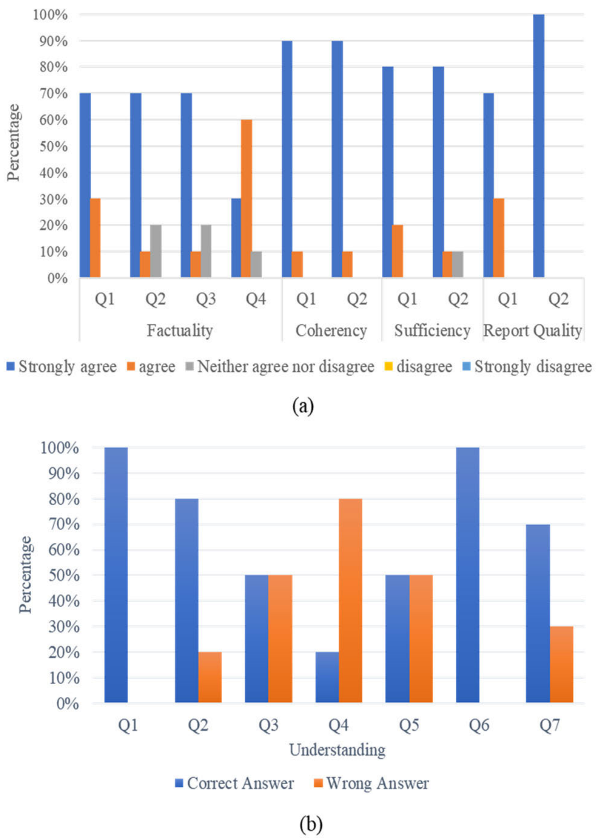
B. EVALUATION OF RESULTS

While typical scientific peer review uses an average of 3-4 evaluators [27], in this study a group of 10 domain experts (who regularly peer review papers) judged the reviewed case form generated by the proposed method. Fig.12 (a) plots the scores for F, C, S and Q for each measure. Fig.12 (b) plots the scores for U. Fig.13 plots the evaluation scores given by each evaluator for each criterion.

For F, C, S, and Q, it can be seen from Fig. 12 (a) that most evaluators 'strongly agreed' on the Coherency and Sufficiency of the report. When asked to evaluate

TABLE 4. Example of problem's types in in the row case form and their solutions in the reviewed case form.

Sentence [Rendering-1] (Row case form)	Problem Type	Suggested solution	Sentence [Rendering-2] (Reviewed case form)
The present study sought to examine the trend and impact of international collaboration in scientific research in Saudi Arabia during the period after the introduction of a reform policy and the normalization of relations with the United States.	Information about the interesting study period of Vietnam country	Adding data concepts to the interesting study period of the selected country	The present study sought to examine the trend and impact of international collaboration in scientific research in Saudi Arabia during **the last two decades that are permeated with King Abdullah Scholarship Programme (KASP).**
Prior to the "Doi Moi" policy and prior to the lifting of the United States trade embargo on Saudi Arabia (1995), Saudi Arabia scientists had had little contact with the Western world and did little research.	Information about the history of Vietnam country	Remove	
We downloaded the entire set of papers published in Microsoft Academic Graph (MAG) indexed journals during the period of 2001 and 2020.	Information about the publications type used in the analysis	Rephrase	We downloaded the entire set of **publications** published in Microsoft Academic Graph (MAG) **that include any of the Saudi Arabia affiliations in the paper affiliations** during the period of 2001 and 2020.
In general, papers with a US author or authors had attracted more citations than non-US authored papers, and this pattern is observed in virtually all research areas.	Deep information	Remove	
All analyses were conducted using the R statistical software version 3.13 (R Foundation for Statistical Computing, Vienna, Austria) on the Window platform.	Information about the tools used for analysis	Adding data concepts to the tools used for analysis	All analyses were conducted **using python 3.7 and Excel** workspace on the Window platform.



(a)

(b)

FIGURE 12. Survey results. (a) criteria (A–D). (b) criterion (E).

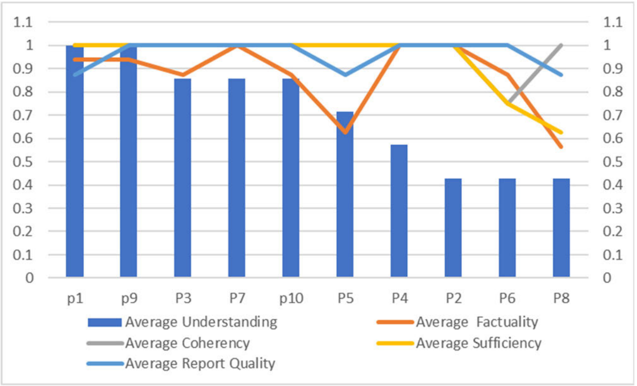


FIGURE 13. Participants' evaluation scores for the report.

contradictions (F.q1), incorrect data (F.q2), incorrect information (F.q3), and grammatical errors (F.q4) they were slightly less sure (slightly less 'Strongly Agree'). Also, 10% of the evaluators were not sure if the report had missing information (S.q2). It is interesting to note that when asked if they had learned any new information/insight (Q.q2) from the paper the scores were unanimously high. Gaining new insight is possibly the most important goal of big data analytics.

Descriptive summary statistics are presented in Table 5, which contains data on the weighted average, the standard deviation (SD), and the average for each report criteria. The weighted average for *F*, *C*, *S*, and *Q* for the 10 evaluators are 0.88, 0.99, 0.96, and 0.96 respectively. This seems to

TABLE 5. Report evaluation descriptive summary.

	U	P	F	C	S	Q
Weighted Average	-	10	0.88	0.99	0.96	0.96
SD	0.23	10	0.15	0.08	0.14	0.06
Average *(U>0)*	0.71	10	0.87	0.98	0.94	0.96
Average *(U>0.5)*	0.84	7	0.89	1	1	0.96
Average *(U>0.8)*	0.93	6	0.94	1	1	0.98

be overall high score. However, it should be noted that few standard deviations, especially for the U (.23), are high. It is possible that some of the evaluators had read the paper more thoroughly than the others. Because it was of interest to see how the final evaluation might have been influenced by the users' reading of the report, this study's researchers looked into all the scores for the three evaluator groupings: i) $U > 0$, which includes all evaluators; ii) $U > 0.5$ which includes evaluators who scored at least 0.5 in U; and iii) $U > 0.8$, which includes only evaluators who scored above.8 in U. The last group indicated the evaluators who possibly read the report most thoroughly. It is interesting to note that all of the *F, C, S,* and *Q* scores improved from the $U > 0$ grouping to the $U > .5$ grouping. It improved even more for the $U > .8$ grouping.

VI. CONCLUSION AND FUTURE WORK

The overall approach of this paper was to take creative work from human experts and multiply its usefulness and impact by using automation. It entails a human expert to 'design' the template paper. Computing was used for massive data analytics. Humans were also used to 'domain-edit' the final report. To show the effectiveness of the proposed model, an experimental human/computer co-authored report for Saudi Arabia was generated and evaluated. Preliminarily, it achieved a satisfactorily evaluation. The human/machine co-authored proto-type paper was rated very high for its ability to provide new learning amongst the readers. This is often the highest goal as well as the biggest challenge of big data analytics [28].

The proposed model can have academic implications on scientometric researchers by providing them with the structure and the main steps that can help multiply the impact and effectiveness of scientometric research analysis reports. For future work, an obvious next step is increasing the level of automation, possibly leading to a full paper. For this model, it means research in automating the current model's human steps and generating a further automated report. Moreover, a plan to generate an automated report comparing the research analysis amongst multiple countries is also suggested. Furthermore, developing a web-based service that can manage extracting the scholarly dataset and generating an automated updated report for any selected country or comparing the research analysis for multiple countries and make it available to the public will be addressed. The first results in this paper indicates that this idea is achievable.

APPENDIX

Please visit http://medianet.kent.edu/techreports/TR2021-09-01-ScientometricAnalysis/TR2021-09-01-ScientometricReport.html for supporting files, including the data concept definitions, the reviewed row case form, and the data concept values of the case study.

REFERENCES

[1] A. Chimariya, ''Streaming data analytics background, technologies, and outlook,'' M.S. thesis, Dept. Inf. Process. Sci., Univ. Oulu, Finland, 2018.

[2] M. S. Batcha and M. Ahmad, ''Publication trend in an Indian journal and a Pakistan journal: A comparative analysis using scientometric approach,'' 2021, *arXiv:2102.12914*.

[3] J. Li, F. Goerlandt, and G. Reniers, ''An overview of scientometric mapping for the safety science community: Methods, tools, and framework,'' *Saf. Sci.*, vol. 134, Feb. 2021, Art. no. 105093, doi: 10.1016/j.ssci.2020.105093.

[4] M. Derntl, ''Basics of research paper writing and publishing,'' *Int. J. Technol. Enhanced Learn.*, vol. 6, no. 2, p. 105, 2014, doi: 10.1504/IJTEL.2014.066856.

[5] C. Astin, C. Harvey, and S. Janusz, ''Writing about science for publication,'' *School Sci. Rev.*, vol. 97, no. 359, pp. 30–38, 2015.

[6] C. Labbé and D. Labbé, ''Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science?'' *Scientometrics*, vol. 94, no. 1, pp. 379–396, Jan. 2013.

[7] Y. Noh, Y. Shin, J. Park, A.-Y. Kim, S. J. Choi, H.-J. Song, S.-B. Park, and S. Park, ''WIRE: An automated report generation system using topical and temporal summarization,'' in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 2169–2172, doi: 10.1145/3397271.3401409.

[8] D. Gkatzia and H. Hastie, ''An ensemble method for content selection for data-to-text generation,'' *History*, vol. 78, nos. 51–74, pp. 11–78, 2015.

[9] L. Maas, M. Geurtsen, F. Nouwt, S. Schouten, R. Van De Water, S. Van Dulmen, F. Dalpiaz, K. Van Deemter, and S. Brinkkemper, ''The Care2Report system: Automated medical reporting as an integrated solution to reduce administrative burden in healthcare,'' presented at the 53rd Hawaii Int. Conf. Syst. Sci., Wailea, HI, USA, 2020. [Online]. Available: https://scholarspace.manoa.hawaii.edu/handle/10125/64184, doi: 10.24251/HICSS.2020.442.

[10] T. Morita, K. Taki, M. Fujimoto, H. Suwa, Y. Arakawa, and K. Yasumoto, ''Beacon-based time-spatial recognition toward automatic daily care reporting for nursing Homes,'' *J. Sensors*, vol. 2018, pp. 1–15, Aug. 2018, doi: 10.1155/2018/2625195.

[11] L. Lebanoff, K. Song, and F. Liu, ''Adapting the neural encoder-decoder framework from single to multi-document summarization,'' 2018, *arXiv:1808.06218*.

[12] J.-P. Ng, Y. Chen, M.-Y. Kan, and Z. Li, ''Exploiting timelines to enhance multi-document summarization,'' in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 923–933, doi: 10.3115/v1/P14-1087.

[13] D. Gkatzia, H. Hastie, and O. Lemon, ''Comparing multi-label classification with reinforcement learning for summarisation of time-series data,'' in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 1231–1240, doi: 10.3115/v1/P14-1116.

[14] I. Androutsopoulos, G. Lampouras, and D. Galanis, ''Generating natural language descriptions from OWL ontologies: The NaturalOWL system,'' *J. Artif. Intell. Res.*, vol. 48, pp. 671–715, Nov. 2013, doi: 10.1613/jair.4017.

[15] S. K. Jalal, ''Co-authorship and co-occurrences analysis using bibliometrix R-package: A casestudy of India and Bangladesh,'' *Ann. Library Inf. Stud.*, vol. 66, no. 2, pp. 57–64, 2019.

[16] T. V. Nguyen, T. P. Ho-Le, and U. V. Le, ''International collaboration in scientific research in vietnam: An analysis of patterns and impact,'' *Scientometrics*, vol. 110, no. 2, pp. 1035–1051, Feb. 2017, doi: 10.1007/s11192-016-2201-1.

[17] W. Y. Low, K. H. Ng, M. A. Kabir, A. P. Koh, and J. Sinnasamy, ''Trend and impact of international collaboration in clinical medicine papers published in Malaysia,'' *Scientometrics*, vol. 98, no. 2, pp. 1521–1533, Feb. 2014, doi: 10.1007/s11192-013-1121-6.

[18] M. Farooq, M. Asim, M. Imran, S. Imran, J. Ahmad, and M. R. Younis, ''Mapping past, current and future energy research trend in pakistan: A scientometric assessment,'' *Scientometrics*, vol. 117, no. 3, pp. 1733–1753, Dec. 2018, doi: 10.1007/s11192-018-2939-8.

[19] Y. Zhao, D. Li, M. Han, C. Li, and D. Li, ''Characteristics of research collaboration in biotechnology in China: Evidence from publications indexed in the SCIE,'' *Scientometrics*, vol. 107, no. 3, pp. 1373–1387, Jun. 2016, doi: 10.1007/s11192-016-1898-1.

[20] M. A. Abolghassemi Fakhree and A. Jouyban, ''Scientometric analysis of the major Iranian medical universities,'' *Scientometrics*, vol. 87, no. 1, pp. 205–220, Apr. 2011, doi: 10.1007/s11192-010-0336-z.

[21] M. Ayers, ''Quick review of Microsoft academic,'' *Issues Sci. Technol. Librarianship*, vol. 96, pp. 1–6, Jan. 2021, doi: 10.29173/istl2582.

[22] A. Salatino, A. Mannocci, and F. Osborne, ''Detection, analysis, and prediction of research topics with scientific knowledge graphs,'' 2021, *arXiv:2106.12875*.

[23] T. Ropinski, ''Combining interactive exploration and search for navigating academic citation data,'' M.S. thesis, Dept. Comput. Sci., Ulm Univ., Germany, 2018.

[24] A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E. D. López-Cózar, ''Google Scholar, Microsoft Academic, Scopus, Dimensions, web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations,'' *Scientometrics*, vol. 126, no. 1, pp. 871–906, 2021.

[25] D. G. Choi, H. Lee, and T.-K. Sung, ''Research profiling for ''Standardization and innovation,'' *Scientometrics*, vol. 88, no. 1, pp. 259–278, Jul. 2011, doi: 10.1007/s11192-011-0344-7.

[26] K. C. Garg and N. Kumar, ''Scientometric portrait of hari chand sharma an outstanding agricultural scientist,'' *DESIDOC J. Library Inf. Technol.*, vol. 39, no. 3, pp. 109–115, May 2019, doi: 10.14429/djlit.39.3.14071.

[27] A. Russo, ''Some ethical issues in the review process of machine learning conferences,'' 2021, *arXiv:2106.00810*.

[28] M. Birjali, A. Beni-Hssane, and M. Erritali, ''Learning with big data technology: The future of education,'' in *Proc. Int. Afro-Eur. Conf. Ind. Advancement*. Cham, Switzerland: Springer, 2016, pp. 209–217.

AMAL BABOUR received the Ph.D. degree in computer science from Kent State University, Kent, OH, USA, in 2017. She worked as a Visiting Assistant Research Professor with the Networking and Media Communications Research Laboratory, Computer Science Department, Kent State University. She is currently an Assistant Professor with the Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. She has participated as a Reviewer for multiple journals and conferences, such as *Journal of Medical Internet Research*, IEEE Access, *Applied Sciences*, ICISDM, and *Intelligent Human Computer Interaction*. Her research interests include artificial intelligence, machine learning, crowdsourcing, and information retrieval.

JAVED I. KHAN (Member, IEEE) received the B.Sc. degree from the Bangladesh University of Engineering and Technology (BUET) and the M.S. and Ph.D. degrees in computer science from the University of Hawai'i at Manoa. He is currently a Professor and the Chair of the Computer Science Department, Kent State University. He has a large experience in computer networks, in particular next-generation network architecture, cross-layer communication, semantic design and composition, complex community, and active and programmable networking. His current research interests include future internet architecture, software-defined networks, and video communications.