Palestinian Journal of Technology and Applied Sciences
No. (5)

Palestinian Journal of Technology and Applied Sciences

Annual Scientific Refereed Journal
No. (5) - January 2022

E - ISSN 2521- 411X
P - ISSN 2520 - 7431

E - ISSN 2521- 411X
P - ISSN 2520 - 7431

PALESTINIAN JOURNAL
OF TECHNOLOGY AND APPLIED SCIENCES (PJTAS)

Annual Scientific Refereed Journal
No. (5) - January 2022

PUBLISHER:
Deanship of Graduate Studies & Scientific Research
Al-Quds Open University

Ramallah & Al-Bireh \ Palestine
P.O. Box: 1804
Tel: +970 - 2- 2976240
+970 - 2- 2956073
Fax: +970 - 2 - 2963738
Email: sprgs@qou.edu

WEBSITE:
http://journals.qou.edu/index.php/pjtas

EMAIL
tas@qou.edu

DESIGN AND PRODUCTION:
Deanship of Graduate Studies & Scientific Research
Al-Quds Open University

E - ISSN 2521 - 411X
P - ISSN 2520 - 7431

المجلة الفلسطينية
للتكنولوجيا والعلوم التطبيقية

مجلة علمية محكمة سنوية
العدد (5) - كانون ثاني 2022م

الناشر:
عمادة الدراسات العليا والبحث العلمي
جامعة القدس المفتوحة

رام الله والبيرة / فلسطين
ص.ب 1804
هاتف: +970-2-2976240
+970-2-2956073
فاكس: +970-2-2963738
بريد إلكتروني: sprgs@qou.edu

الموقع الإلكتروني للمجلة:
http://journals.qou.edu/index.php/pjtas

البريد الإلكتروني للمجلة:
tas@qou.edu

تصميم واخراج فني:
عمادة الدراسات العليا والبحث العلمي
جامعة القدس المفتوحة

E - ISSN 2521 - 411X
P - ISSN 2520 - 7431

Palestinian Journal
of Technology & Applied Sciences (PJTAS)

Palestinian Journal
of Technology & Applied Sciences (PJTAS)

Vision

Achieving leadership, excellence and innovation in the field of open learning, community service, and scientific research, in addition to reinforcing the University leading role in establishing a Palestinian society built on knowledge and science.

Mission

To prepare qualified graduates equipped with competencies that enable them to address the needs of their community, and compete in both local and regional labor markets. Furthermore, The University seeks to promote students' innovative contributions in scientific research and human and technical capacity-building, through providing them with educational and training programs in accordance with the best practices of open and blended learning approach, as well as through fostering an educational environment that promotes scientific research in accordance with the latest standards of quality and excellence. The University strives to implement its mission within a framework of knowledge exchange and cooperation with the community institutions and experts.

Core Values

To achieve the University's vision, mission and goals, the University strives to practice and promote the following core values:

- Leadership and excellence.
- Patriotism and nationalism.
- Democracy in education and equal opportunities.
- Academic and intellectual freedom.
- Commitment to regulations and bylaws.
- Partnership with the community
- Participative management.
- Enforcing the pioneer role of women.
- Integrity and Transparency.
- Competitiveness.

The Journal

The Palestinian Journal of Technology and Applied Sciences is an annual scientific refereed journal, issued by the Deanship of Graduate Studies and Scientific Research. The first issue of the Journal was published in January 2018 after obtaining an International Standard Serial Number (E- ISSN: 2521-411X), (P– ISSN: 2520-7431).

The journal publishes original research papers and studies conducted by researchers and faculty staff at QOU and by their counterparts at local and overseas universities, in accordance with their academic specializations. The Journal also publishes reviews, scientific reports and translated research papers, provided that these papers have not been published in any conference book or in any other journal.

The Journal comprises the following topics:

Information and Communication Technology, Physics, Chemistry, Biology, Mathematics, Statistics, Biotechnology, Bioinformatics, Agriculture Sciences, Geology, Ecology, Nanotechnology , Mechatronics, Internet of things , Artificial Intelligence and Big Data.

Palestinian Journal
of Technology & Applied Sciences (PJTAS)

Publication and Documentation Guidelines

First: Requirements of preparing the research:

The research must include the following:

1. A cover page which should include the title of the research stated in English and Arabic, including the name of researcher/researchers, his/her title, and email.
2. Two abstracts (English and Arabic) around (150-200 word). The abstract should include no more than 6 key words.
3. Graphs and diagrams should be placed within the text, serially numbered, and their titles, comments or remarks should be placed underneath.
4. Tables should be placed within the text, serially numbered and titles should be written above the tables, whereas comments or any remarks should be written underneath the tables

Second: Submission Guidelines:

1. The Researcher should submit a letter addressing the Head of Editorial Board in which he/she requests his paper to be published in the Journal, specifying the specialization of his/her paper.
2. The researcher should submit his research via email to the Deanship of Scientific research (tas@qou.edu) in Microsoft Word Format, taking into Consideration that the page layout should be two columns.

   (Check the attached digital form on the website of the Journal)
3. The researcher should submit a written pledge that the paper has not been published nor submitted for publishing in any other periodical, and that it is not a chapter or a part of a published book.
4. The researcher should submit a short Curriculum Vitae (CV) in which she/he includes full name, workplace, academic rank, specific specialization and contact information (phone and mobile number, and e-mail address).
5. Complete copy of the data collection tools (questionnaire or other) if not included in the paper itself or the Annexes.
6. No indication shall be given regarding the name or the identity of the researcher in the research paper, in order to ensure the confidentiality of the arbitration process.

Third- Publication Guidelines:

The editorial board of the journal stresses the importance of the full compliance with the publication guidelines, taking into note that research papers that do not meet the guidelines will not be considered, and they will be returned to the researchers for modification to comply with the publication guidelines.

1. Papers are accepted in English only, and the language used should be well constructed and sound.
2. The researcher must submit his/her research via email (tas@qou.edu)in Microsoft Word format, taking into consideration the following:
   - Font type should be Times New Roman, and the researcher should use bold font size 14 for head titles, bold font size 13 for subtitles, font size 12 for the rest of the text, and font size 11 for tables and diagrams.
   - the text should be single-spaced
   - Margins: Should be set to: 2cm top, 2.5 cm bottom, 1.5 cm left and right.
3. The paper should not exceed 25 (A4) pages or (7000) words including figures and graphics, tables, endnotes, and references, while annexes are inserted after the list of references, though annexes are not published but rather inserted only for the purpose of arbitration.
4. The research has to be characterized by originality, neutrality, and scientific value.
5. The research should not be published or submitted to be published in other journals, and the researcher has to submit a written acknowledgment that the research has never been published or sent for publication in other journals during the completion of the arbitration process. In addition, the main researcher must acknowledge that he/she had read the publication guidelines and he/she is fully abided by them.
6. The research should not be a chapter or part of an already published book.
7. Neither the research nor part of it should be published elsewhere, unless the researcher obtains a written acknowledgement from the Deanship of Scientific Research.
8. The Journal preserves the right to request the researcher to omit, delete, or rephrase any part of his/her paper to suit the publication policy. The Journal has also the right to make any changes on the form/ design of the research.
9. The research must include two research abstracts, one in Arabic and another in English of (150-200) words. The abstract must underline the objectives of the paper, statement of the problem, methodology, and the main conclusions. The researcher is also to provide no more than six keywords at the end of the abstract which enable an easy access in the database.

11. The researcher has to indicate if his research is part of a master thesis or a doctoral dissertation as he/she should clarify this in the cover page, possibly inserted in the footnote.
12. The research papers submitted to the Deanship of Scientific Research will not be returned to the researchers whether accepted or declined.
13. In case the research does not comply with the publication guidelines, the deanship will send a declining letter to the researcher.
14. Researchers must commit to pay the expenses of the arbitration process, in case of withdrawal during the final evaluation process and publication procedures.
15. The researchers will be notified of the results and final decision of the editorial board within a period ranging from three to six months starting from the date of submitting the research.

Four- Documentation:

1. Footnotes should be written at the end of the paper as follows; if the reference is a book, it is cited in the following order, name of the author, title of the book or paper, name of the translator if any or reviser, place of publication, publisher, edition, year of publishing, volume, and page number. If the reference is a journal, it should be cited as follows, author, paper title, journal title, journal volume, date of publication and page number.
2. References and resources should be arranged at the end of the paper in accordance to the alphabetical order starting with the surname of author, followed by the name of the author, title of the book or paper, place of publishing, edition, year of publication, and volume. The list should not include any reference which is not mentioned in the body of the paper.
   - In case the resource is with no specified edition, the researcher writes (N.A)
   - In case the publishing company is in not available, the researcher writes (N.P)
   - In case there is no author, the researcher writes (N.A)
   - In case the publishing date is missing , the researcher writes (N.D)
3. In case the researcher decides to use APA style for documenting resources in the text, references must be placed immediately after the quote in the following order, surname of the author, year of publication, page number.
4. Opaque terms or expressions are to be explained in endnotes. List of endnotes should be placed before the list of references and resources

   Note: for more information about using APA style for documenting please check the following link:

   http://journals.qou.edu/recources/pdf/apa.pdf

Five: Peer Review & Publication Process:

All research papers are forwarded to a group of experts in the field to review and assess the submitted papers according to the known scientific standards. The paper is accepted after the researcher carries out the modifications requested. Opinions expressed in the research paper solely belong to their authors not the journal. The submitted papers are subject to initial assessment by the editorial board to decide about the eligibility of the research and whether it meets the publication guidelines. The editorial board has the right to decide if the paper is ineligible without providing the researcher with any justification.

The peer review process is implemented as follows:

1. The editorial board reviews the eligibility of the submitted research papers and their compliance with the publication guidelines to decide their eligibility to the peer review process.
2. The eligible research papers are forwarded to two specialized Referees of a similar rank or higher than the researcher. Those Referees are chosen by the editorial board in a confidential approach, they are specialized instructors who work at universities and research centers in Palestine and abroad.
3. Each referee must submit a report indicating the eligibility of the research for publication.
4. In case the results of the two referees were different, the research is forwarded to a third referee to settle the result and consequently his decision is considered definite.
5. The researcher is notified by the result of the editorial board within a period ranging from three to six months starting from the date of submission. Prior to that, the researcher has to carry out the modifications in case there are any.
6. The researcher will receive a copy of the journal in which his/her paper was published, as for researchers from abroad, a copy of the Journal volume will be sent to the liaison university office in Jordan and the researcher in this case will pay the shipping cost from Jordan to his/her place of residency.

Six: Scientific Research Ethics:

The researcher must:

1. Commit to high professional and academic standards during the whole process of conducting research papers, from submitting the research proposal, conducting the research, collecting data, analyzing and discussing the results, and to eventually publishing the paper. All must be conducted with integrity, neutralism and without distortion.

2. Acknowledge the efforts of all those who participated in conducting the research such as colleagues and students and list their names in the list of authors, as well as acknowledging the financial and morale support utilized in conducting the research.
3. Commit to state references soundly, to avoid plagiarism in the research.
4. Commit to avoid conducting research papers that harm humans or environment. The researcher must obtain in advance an approval from the University or the institutions he/she works at, or from a committee for scientific research ethics if there is any, when conducting any experiments on humans or the environment.
5. Obtain a written acknowledgement from the individual/individuals who are referred to in the research, and clarify to them the consequences of listing them in the research. The researcher has also to maintain confidentiality and commit to state the results of his/her research in the form of statistical data analysis to ensure the confidentiality of the participating individuals.

Seven: Intellectual Property Rights:

1. The editorial board confirms its commitment to the intellectual property rights
2. Researchers also have to commit to the intellectual property rights.
3. The research copyrights and publication are owned by the Journal once the researcher is notified about the approval of the paper. The scientific materials published or approved for publishing in the Journal should not be republished unless a written acknowledgment is obtained by the Deanship of Scientific Research.
4. Research papers should not be published or republished unless a written acknowledgement is obtained from the Deanship of Scientific Research.
5. The researcher has the right to accredit the research to himself, and to place his name on all the copies, editions and volumes published.
6. The author has the right to request the accreditation of the published papers to himself.

Contents

Higher Compression Rates for GSM 6.10 Standard Using Lossless Compression

معدل ضغط أعلى لبيانات معيار(جي إس إم GSM 6.10) باستخدام الضغط الأقل فاقدية

Islam Younis Amro
Associate Professor/ Al-Quds Open University / Palestine
iamro@qou.edu

اسلام يونس عمرو
أستاذ مشارك / جامعة القدس المفتوحة / فلسطين

Received: 18/10/2020, Accepted: 04/09/2021
DOI: https://doi.org/10.33977/2106-000-005-001
https://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 2020/10/18، تاريخ القبول: 2021/09/04
E-ISSN: 2521-411X
P-ISSN: 2520-7431

Abstract

This research aims at exploiting the lossless Hamming correction code compression algorithm (HCDC) to reduce the transmission data rate in the GSM 6.10 standard, which holds several similarities with modern adaptive multi-rate codec in coefficients calculations and excitation principles. The compression algorithms depend on the properties of the hamming codes where data bits can be calculated from the parity bits. In this research, we chose parity equals 3 and data bits equals 4. Several iterations were conducted over the compressed frame information to achieve even higher compression rates. The compression rate was implemented over the standard of GSM 6.10, which is a variation Code Exited Linear Prediction coding (CELP). Regarding the data samples selected to conduct the test, two males and two females' voice file samples at 8khz and quantized on 8-bit resolution were selected. The duration of the files varies from 4 to 6 seconds. Each sample was divided into 20ms frames; each frame was expressed using GSM6.10 with 260 bits of data included Linear perdition coefficients, pitch period, gain, peak magnitude value, grid position, and the sample amplitude. This shows that the 260 bits every 20ms form a data rate of 13kbps. The 260 bits were subjected to HCDC, and the data rate was reduced by 60%, reaching down to 5kbps on average. The results compared to the famous FLAC lossless audio compression, which showed 15% compression only. The research did not consider any quality testing since the compression is lossless. The research used standard ITU libraries to conduct the GSM6.10 data acquisition and open-source platforms for FLAC.

***Keywords**: Linear prediction coding, lossless compression, speech compression, source coding, cellular communication.*

الملخص

يهدف هذا البحث إلى توظيف خوارزمية ضغط البيانات المعتمدة على ترميز (هامينج) للتصحيح (HCDC) دون فقدان البيانات؛ وذلك لتقليل معدل بيانات الإرسال لمعيار (جي إس إم GSM 6.10) والذي يتم استخدام أسس الترميز الخاصة به في معظم الترميزات الحديثة من احتساب معاملات خطية، وطرق احتساب إشارة تفعيل الفلاتر الخطية بشكل عام، تعتمد خوارزمية الضغط المذكورة على توظيف خصائص ترميز (هامينج) بحيث يتم استخدام (البت) الخاص بحقل البيانات كأساس في احتساب (البت) الخاص (بالباريتي)، وعليه يتم ارسال (البتات) الخاصة بالبيانات، و توفير (البتات) الخاصة (بالباريتي). في هذا البحث، قمنا باعتماد مجموعة من (البتات) تتكون من سبعة حقول بحيث يكون عدد (البتات) الخاصة بالبيانات (4)، و عدد (البتات) الخاصة (بالباريتي) (3)، و تم أيضا إعادة ضغط البينات المضغوطة مرات عدة متتالية للحصول على معدل ضغط أعلى. تم تطبيق عملية الضغط على المعيار (GSM6.10) والذي يعتبر أحد أنواع ترميز التوقع الخطي المحفَّز خارجيا، تم استخدام عينات من أصوات ذكرين بالغين وانثيين بالغتين، وأخذت عينات الإشارة بتردد (8) كيلو هيرتز ورقمتنها بثماني بتات، مدة كل ملف عينة من أربع الى ست ثوانٍ، وتمّ تقطيع كل عينة إلى نوافذ طول النافذة عشرون ميلي ثانية، ومن ثم ترميز كل نافذة باستخدام (GSM6.10)، والتي بدورها تنتج (260) بتا كل عشرين ثانية. تشتمل هذه (البتات) على: معاملات توقع الخطة، تردد الصوتي الذبذبة (Pitch)، التحصيل (Gain)، قيمة القمة الأعلى للإشارة، موقع الشبكة (Grid position)، قيمة العينة. يتم تجمي ال (260) بتا، ومن ثم يتم ضغطهم مرات عدة باستخدام خوارزمية HCDC. تم تخفيض معدل الارسال من (13) كيلوبت في الثانية إلى قرابة (5) كيلوبت بالثانية في المعدل. تمت مقارنة هذه النتائج بخوارزمية (FLAC)، والتي حققت نسبة ضغط بمعدل (15%) فقط. وبما أن الضغط المستخدم هو ضغط لا يفضي إلى فقد البيانات (Lossless)، لم يتم التطرق إلى دراسة جودة الإشارة في هذه البحث. وتم استخدام المكتبة القياسية لترميز (GSM6.10)، والمتوفرة على موقع اتحاد الاتصالات (الدولب) إلى جانب مكتبة (FLAC) مفتوحة المصدر.

الكلمات المفتاحية: ضغط البيانات، ضغط الصوت، الشبكات الخليوية، معاملات التوقع الخطي، الترميز المصدري، ضغط بيانات بلا خسائر.

INTRODUCTION

Audio and speech compression might be considered the most diverse aspect in the data compression discipline. This is due to the diversity of its domains, data representation methods, and the high demand for high quality and lower data rate paradox. Not to forget, the complexity constraints over any algorithm are to be proposed (Wu et al. 2002).

The basic form of any signal is acquired after its quantization (Openhaim, 1997). This is the point where all digital compression algorithms start; a well-known followed track is the linear prediction coding compression approach due to its

low rate, good quality, and acceptable complexity. It became the hardcore of modern voice communication systems (Kain et al., 2001; Wah, 2005) and the raw data form for artificial intelligence applications on speech. (Wu et al., 2002, Lam et al., 2000).

As for the review of lossless audio compression standards and algorithms proposed by AbdulMuin et al. (2017); it shows that several compression approaches are used either in row PCM form or in other coding formats, mainly based on Huffman methods. Recent implantation was found in the study of Uttam, 2019, achieving lossless compression of audio by encoding its constituted components (LCAEC), which are based on Huffman and Burrows–Wheeler transform. On lossless audio compression based on heuristic methods based on neural networks found in Uttam, 2019, several hidden layers have been implemented in the proposed network for the present encoding framework based on deep learning process. Another lossless audio compression method is incorporated by the nature of channels of transmission and the types of data like in Takehiro et al., 2019, where the compression is considered in terms of video compression channel and is based on MPEG multichannel audio compression. Another statistical compression method found in Yanzhen et al., 2019. This method is based on pulse destitution modeling then generates a fixed codebook that enables AMR features. For spatial audio decoding and compression, extensive research was conducted by Menzies et al., 2017. The research considered decoding and compressing channel and scene objects to reduce processing complexity. In Luo et al., 2017, an auto encoder was exploited to detect the double compression for AMR. This research was useful in detecting several compressions for the same block when several transmission rates are used. Another research on statistical methods of auditory representation was found in Biesmans et al., 2017. Based on canonical correlation analysis, that emulates the auditory system signaling in EEG, brain is stimulated directly by passing the human auditory system. The importance of this research lies in how to generate and EEG signal from an audio signal. This is a new form of coding and compression.

This research exploits a new lossless compression algorithm based on the Hamming Correction Code Compression (HCDC) explained in Bahadili, 2007, in compressing speech/audio signals in its GSM 6.10 form. Similar work was conducted in Amro et al., 2011, using this compression algorithm over, and an experimental vocoder that exploits residual signal as excitation using Discrete Cosine Transform (DCT) with considerable compression ratio. The compression algorithm in this research addressed the linear prediction coefficients only without addressing the DCT excitation signal. The HCDC algorithm was also exploited in compressing audio signal based on Code excited linear prediction coding in Amro, 2013. In this research, both the excitation signal and linear prediction coefficients were addressed and achieved a good average compression rate.

Although several GSM 6.10 standards were promoted to Adaptive Multirate (AMR) codec, to enhance quality, in addition to the Adaptive Multirate Narrow-Band (AMR-NB) codec, which works in the telephony bandwidth in addition to the Global System for Mobile Communications (GSM) and Universal Mobile Telecommunications System (UMTS) systems. The founding principles of the GSM 6.10 are available in these codecs, such as the linear prediction coefficients calculation approach based on the Levinson Durban recursion and the quantization of the excitation signal, in addition to gaining value. These parameters are present in all generations of codecs and exploited in cellular communication (GSM 2020), making it easier for this research to prove the concept over GSM 6.10 with the possibility to generalize the results of the scale of different rates in the AMR in the future.

The following section discusses the GSM 6.10 encoding and decoding. We elaborate on the properties of the algorithm exploited and then mention the methodology and experiment design in the following section. The results are presented in the following section with comments and analysis. Then we finalize with a summary and conclusion.

The GSM 06.10 full rate

THE GSM 06.10 full rate coder is considered a hybrid code which is a form between waveform coders and vocoders. Waveforms coders consider the processing among physical

characteristics of the signal in the time domain, frequency domain, or any other transfer function domain. Vocoders have their own domain based on linear prediction coding (LPC). LPC works on the classification of speech signal as voiced or unvoiced. A voiced signal is formed from sound in which certain turbulence happens in vocal cords. This turbulence has a certain frequency which is called pitch. A pitch is a train on impulse with known frequency and gain, represented in the linear prediction domain. The frequency of this signal in the LPC domain is the pitch frequency for a given voice. The unvoiced signals in the LPC are incorporated with voices that do not include vocal tracts turbulence, like the letter S. This kind of letter has no certain frequency in the LPC domain. Thus, it is expressed as white noise. Both white noises and/or the pitch impulse train are synthesized with digital filer with certain order (10 minimum and usually 12). The filter is the linear synthesis digital filer, and its coefficients are calculated from time-domain parameters from the signal. This process synthesizes the spoken voice back. The quality of the output signal in terms of physical signal qualities (objective), such as signal-to-noise ratio (SNR) and Segmental Signal to noise ratio (SSNR), is considered low. However, it still can be heard and understood. That is why a special qualitative (subjective) technique is adapted based on voting. This quality assessment method is known as the Mean of Score (MOS), and it usually ranges from 0 to 5. However, 3.5 is the range of good and acceptable quality (Chu, 2003).

The GSM uses a compression approach that utilizes both waveform methods and LPC Methods. In GSM speech encoder, the encoder takes 13 bits as input as Pulse Code Modulation (PCM) signal from audio part of a mobile station or from the network or Public Switched Telephone Network (PSTN) via an 8bit / A-law to 13 (13bit* 8KHz=104Kbps) bit uniform PCM (Malvar, 2007). The encoded speech output is delivered to a channel encoder unit specified in GSM 05.03 (Hu et al., 2007).

On the receiving side, an inverted operation takes place as described in GSM 06.10. The process is based on a mapping between inputting 160 speech samples, each is 13-bit uniform PCM, then it is exploited to encode 260 bits, and from encoded blocks of 260 bits to generate an output of 160 reconstructed speech blocks. The rates are 8K samples per second, generating an encoded bitstream of 13kbps. This coding scheme is known as regular pulse excitation long-term prediction linear predictive coding.

GSM Full Rate Encoder

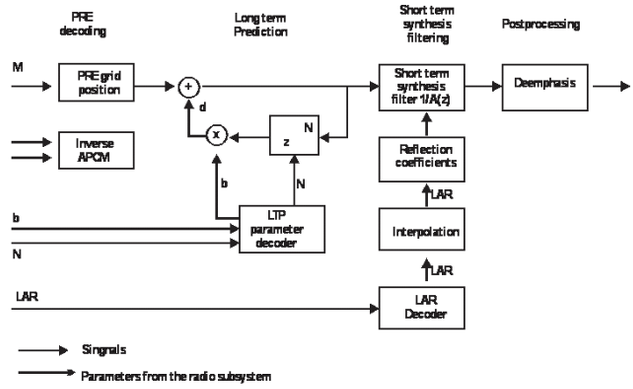


Figure 1 GSM Encoding

Figure 1 shows a detailed block diagram of GSM 06.10 Speech Encode. The speech input frame, made of 160 samples, is the first step to generate an offset-free signal, then a pre-emphasis filter is applied. Then160 samples were used to determine the short-term LPC coefficients through the LPC analysis. This process is conducted by calculating the Lavinoson Durban coefficient, then calculating the LPC residual signal for the short-term signal. Before transmission, the filter parameters, reflection coefficients, and gain are transferred to Log Area Ratios (LAR). The speech frames are then slitted into 4 sub-frames with 40 samples of short-term residual signal in each. Each sub-frame is processed as a block by the following functional components. Before processing sub-blocks of 40 short term residual samples, the parameters of the long term analysis filter, the Long Term Parameter (LTP), and the gain are estimated in the LTP analysis block, based on the current sub-block of the present and a stored sequence of the 120 previous short term residuals. Then by subtracting 40 estimates of the short-term residual signal from the short-term residual signal itself, where a block of 40 long-term residual signal samples is acquired. In the next stage, the block of 40 long-term residual signal is fed to the Regular Pulse Excitation (RPE) stage that performs a basic compression function analysis. Resulting from the RPE stage, the block of 40

input long-term residual signal samples is represented by one 4 sub-sequences candidates with 13 pulses each. The 13 RPE pulses are then encoded using Adaptive Pulse Code Modulation (APCM) with an estimation of sub-block amplitude which is transmitted to the decoder side as side information. The RPE values are also supplied to the local RPE decode-and-reconstruct module, which produces a block of 40 samples. These samples are the quantized versions of the long-term residual signal. Adding the quantized 40 samples of the long-term residual to the blocks of short-term residual signal previously encountered, the reconstructed short-term residual signal is acquired. The block containing the short-term residual signal is consequently obtained. Then the reconstructed signal is inputted in the analysis filter, which produces a new block of forty short-term residual signal estimates. These estimates are forwarded to the next sub-block to complete the feedback loop (ETSI, 2010).

GSM Full Rate Decoder

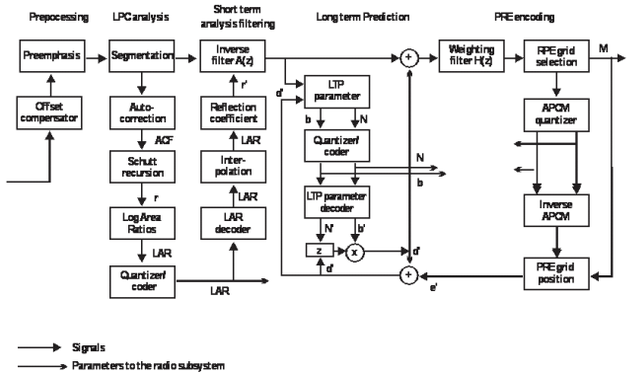


Figure 2 GSM Decoding

The GSM 06.10 Speech Decoder is shown in figure 2. As it can be seen, it includes similar stages to the feedback loop in the encoder. To ensure a zero-error transmission, the output must be the reconstructed short-term residual signal samples. These samples are inputted into a short-term synthesis filter. The next stage is the de-emphases filter in order to reconstruct the required speech signal. The GSM elaborated extensively on mapping input blocks of 160 samples in the original 13-bit uniform pulse code modulation format. This is done to encode 260 bits of blocks from encoded blocks of 260 bits of output blocks. This is obtained from 160 reconstructed speech samples. The average bit rate for the encoded stream is 13kbps obtained from 8000 samples per second. The bit allocation for the GSM full rate speech coding is seen in the table below and will be subjected to further compression using HCDC Algorithm. The frame length that is subjected in the process in 20 milliseconds.

Table 1 Bit allocation for GSM Full Rate Speech Coder (ETSI, 2010)

Parameter	No. per frame	Resolution	Total bits / frame
LPC	8	6,6,5,5,4,4,3,3	36
Pitch Period	4	7	28
Long Term Gain	4	2	8
Grid Position	4	2	8
Peak Magnitude	4	6	24
Sample Amplitude	4*13	3	156
Total			260

Hamming Correction Code Compression

Hamming Correction Code Compression (HCDC) is derived from hamming correction code. Let's consider the following set/ word of bits {*b0, b1, b2, b3, b4, b5, b6*}, re-expressing the set-in terms of its hamming version, we have {*p0, p1, d0, p2, d1, d2, d3*}, where number of parities=3 for a word of 7 bits length. In our research, we will transmit or save d bits only, and on the reception side, we will calculate the parity, so we can express 7 bits with 4 bits of data and save 3 bits. When we can do this process for the set of bits, we call it a valid word, which refers to the words' valid hamming calculation of data bits leads to the similar parity bits. If the word is invalid, this means that its data bit does not match its parity bits. In this case, we cannot compress it, and we have to transmit the word as is. We can compress valid words only; invalid words cannot be compressed since their actual bits don't match the ones calculated in hamming conditions. We mark valid words by 1 and invalid words by zero. This bit tells the decompressor what to do. In the case of a valid word, it means that we calculate the parity bits and place them in their right locations. In the case of an invalid word, we read 0 in the leading bit and read the whole word as is. Exhibited in Figure 3 is the compression algorithm, while Figure 4 exhibits the decompressor algorithm

1. *Initialization*
   - *Select* p
   - *Calculate* $n = 2^p - 1$
   - *Calculate* $d = n - p$
   - *Initialize* $b = 0$
2. *Read Binary Data*
   - *Read a Block of* n *bits length*
   - *[Add 1 to* b *]*
3. *Check block validity*
   - *If {Block = valid codeword} then*
     - *[Add 1 to* v *]*
     - *Write 1 followed by* d *block bits to the compressed file*
   - *Else {block= non-valid codeword}*
     - *[add 1 to* ω *]*
     - *Write 0 followed by* n *block bits to the compressed file*
   - *End if*
4. *if not end of data go to step 2*

Figure 3 HCDC Compressor

1. *Initialization*
   - *Select* p
   - *Calculate* $n = 2^p - 1$
   - *Calculate* $d = n - p$
   - *Initialize* $b = 0$
2. *Read Binary Data*
   - *Read the first bit* (h)
   - *[add 1 to* b *]*
3. *check for block validity*
   - *if {* $h = 1$ *}then*
     - *add [1 to* v *]*
     - *read* d *data bits*
     - *compute the hamming code for* d *write coded block to decompressed file*
   - *else {* $h = 0$ *}*
     - *[add 1 to* ω *]*
     - *Read block of* n *length*
     - *Write block* n *bits to he decompressed file*
   - *End if*
4. *if not end of data go to step 2*

Figure 4 HCDC deCompressor

Now we work on the evolution of its compression rate. The measuring references suggested in (Bahadili, 2008) are which represents the block size, i.e., the block we intend to analyze. The measuring references for Compression Rate suggested in the study of Bahadili, 2008 is, which represents the block size, the file to be compressed contains blocks, each is made of a number of bits, valid blocks count is expressed as and the invalid blocks are expressed as, the whole number of blocks can be expressed as:

$$b = \upsilon + \omega \quad (1)$$

This is a valid block led by 1 and an invalid one led by zero. So, the valid block is expressed by only its data bits excluding parity bits, the size of the valid block in the group is given by:

$$S_v = v(d+1) \quad (2)$$

For invalid blocks, the whole is used, so the size of the invalid blocks becomes

$$S_w = \omega(n+1) \quad (3)$$

And the size of the whole compressed file S_C becomes

$$S_c = nb + b - vp \quad (4)$$

The size of the compressed file in bits becomes

$$S_c = v(d+1) + \omega(n+1) \quad (5)$$

This can be written as

$$S_c = nb + b - vp \quad (6)$$

We know that the original file S_o is expressed as

$$C = \frac{nb}{nb+b-vp} \quad (7)$$

The compression ratio becomes

$$C = \frac{n}{n+1-rp} \quad (8)$$

expressing the ratio of valid blocks r as $r = \frac{v}{b}$.

The previous equation can be written as

$$C_k = \frac{S_o}{\prod_{i=1}^{k} C_i} \quad (9)$$

The algorithm can be iterated k times, where further compression can be achieved if the output of each phase is taken as an input for the next phase, the cumulative compression rate in this case the C_k, where k represents the number of iterations and C_i represent the compression on a given round, so if the code is to be compressed 8 times, then k is set to 8, and the compression rate becomes $\{C_1, C_2, \dots, C_{7,} C_8\}$.

MATERIALS AND METHODS

The experiment was carried out on several data sets. Signals were S1(female), S2(female), S3(male), and S4(male). The samples S were for adult native English-speaking males and females. For each signal, we used 8-bit resolution at 8KHz

sampling. The signal samples are segmented into a 20ms frame each, and the length of the samples ranges from 3 to 6 seconds each. The data to be compressed is obtained from Table 1 above, which includes the Bit allocation for the GSM Full Rate Speech Coder. For each 20ms of the sample signals, we will compress the 260 bits representing the GSM full rate speech coder information. The current data rate of the coder is 13Kbps. We will work on reducing this number in a lossless manner. In this case, there will be no need for quality detection. The performance of our work will then be compared for the FLAC algorithm since it is a wildly used lossless compression. The current transmission rate for the GSM algorithm used is 13Kbps obtained from 260 bits per sample over a window of 20ms. For each frame with data in table 1, we moved to the steps in figure 5, in order to evaluate the compression performance at parity =3 and at a different number of iterations. The selected number of iterations is 8 from practical experience. The iterations as mentioned above help enhance the compression ratio.

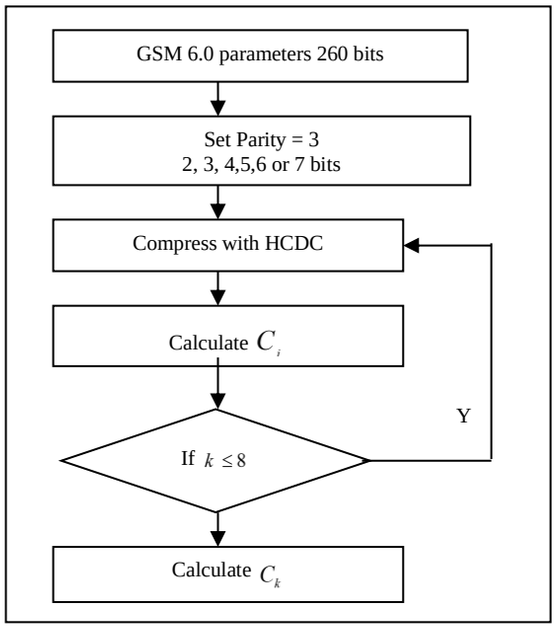


Figure 5 HCDC Experiment Design

The compression rate is to be calculated against the given parity = 3 on every count. The overall compression C_k is to be calculated for each sample accordingly and specified at the last iteration. The cumulative compression rate is then compared to the FLAC compression rate and the transmission rate. Then, the transmission rates are plotted together to see the average compression rate for the whole sample. This is calculated by averaging the rates for all frames within the sample for both HCDC and FLAC. Then performance notes are made.

RESULTS

For all the samples, compression was encountered only at parity=3. Table 2 below shows some of the best cases achieved with HCDC against frames at parity=3, the field Loop in the tables represents the compression turns, which iterates 8 times. The frame file size expressed the total number of bits in the frame. Valid Blokes represents the valid hamming codeword r as the valid blocks' ratio to the whole blocks in the file. Compression Ratio is C and computed by equation 8 above. Cumulative Compression is the and computed by equation 9, which represents the size rate between the original frame file and the current frame file size at the 8th iteration.

Table 2 One of the best cases achieved with HCDC against frames at parity=3

File Size	Total Block	Ratio	Valid	Invalid	Compressed	Ratio	Comm
60	37	0.46	17	20	228	1.14	1.14
228	32	0.5	16	16	192	1.19	1.36
192	27	0.44	12	15	168	1.14	1.55
168	24	0.46	11	13	148	1.14	1.76
148	21	0.43	9	12	132	1.12	1.98
132	18	0.44	8	10	112	1.18	2.33
112	16	0.31	5	11	108	1.04	2.42
108	15	0.27	4	11	104	1.04	2.51

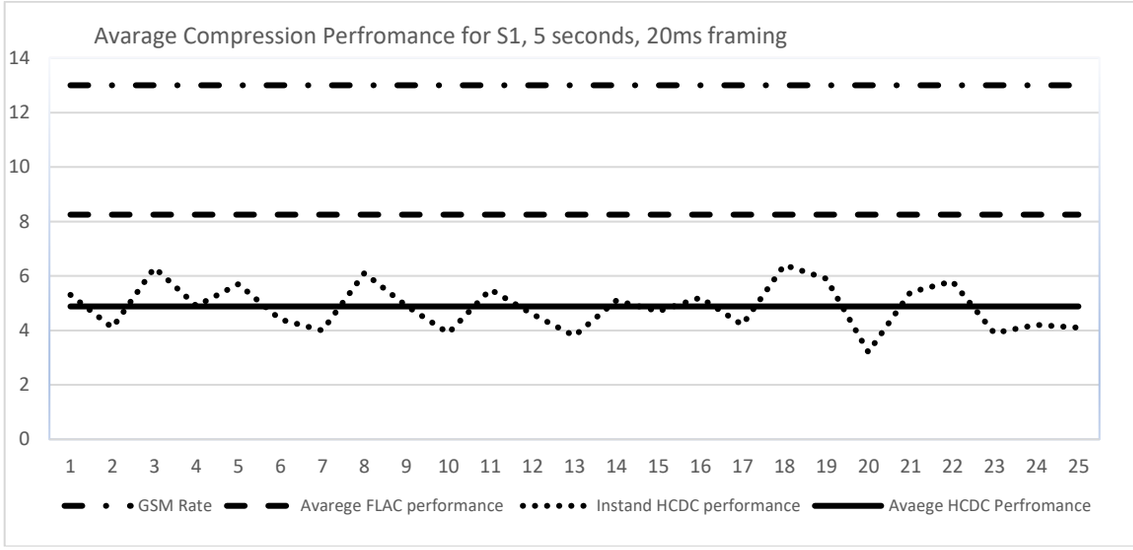


Figure 6 Average Transmission Rate for Sample Female 1

In Figure 6, we can see the algorithm has a very high potential of achieving lossless compression. The following table shows the frame information for the rest of the samples.

Table 3 Frame information for sample Male 2

File Size	Total Block	Ratio	Valid	Invalid	Compressed	Ratio	Comm
260	37	0.51	19	18	220	1.18	1.18
220	31	0.35	11	20	204	1.08	1.27
204	29	0.55	16	13	168	1.21	1.55
168	24	0.58	14	10	136	1.23	1.91
136	19	0.58	11	8	108	1.26	2.40
108	15	0.4	6	9	96	1.125	2.70
96	13	0.23	3	10	92	1.04	2.83
92	13	0.31	4	9	88	1.05	2.95

Table 4 Frame information for sample female 1

File Size	Total Block	Ratio	Valid	Invalid	Compresed	Ration	Comm
260	37	0.43	18	19	224	1.16	1.17
216	30	0.47	17	13	172	1.26	1.52
182	26	0.42	11	15	164	1.11	1.59
164	23	0.26	10	13	144	1.14	1.81
144	20	0.15	9	11	124	1.16	2.10
124	17	0.26	6	11	112	1.11	2.33
112	16		5	11	108	1.04	2.45
108	15		4	11	104	1.04	2.51

Table 5 Frame information for sample female 2

File Size	Total Block	Ratio	Valid	Invalid	Compresed	Ration	Comm
260	37	0.56	22	15	208	1.25	1.25
208	29	0.62	18	11	160	1.3	1.6
160	22	0.45	10	12	136	1.17	1.95
136	19	0.32	6	13	128	1.06	2.03
128	18	0.28	5	13	124	1.03	2.09
124	17	0.24	4	13	120	1.03	2.16
120	17	0.18	3	14	124	0.96	2.1
124	17	0.24	4	13	120	1.03	2.17

Average HCDC Compression performance and comparison

Table 6 Average performance of HCDC algorithm over given Samples

Sample	Duration in seconds	GSM kbps	FLAC Kbps	HCDC average Kbps	Reduction average (GSM to HCDC)
Male 1	5	13	8.25	4.86	62%
Male 2	6	13	7.89	5.21	59%
Female 2	5	13	9.14	4.79	63%
Female 2	6	13	9.21	4.88	62%

We can see from the table right above the general performance references regarding the HCDC compression. The FLAC has an average drop within 3 kbps. However, the challenge of FLAC since compression depends heavily on the nature of data. The file-based compression was used in this research, and the result was used as an average value in all of the cases. For the HCDC average Kbps, this is the average value of the broadcasted frames per file sample. As we can see, it achieved a very high compression rate with an average that exceeds 60% for all cases. In comparison to FLAC, it also achieved compression that exceeds FLAC but 40%. The HCDC is easier to implement and can give good performance in small blocks of data.

CONCLUSION

This paper exploits the Hamming Correction Code Compressor (HCDC) in compressing GSM full rate compression in a lossless manner. These parameters are calculated for every 20ms frame and then subjected to the lossless compressor. The parameters are the linear perdition coefficients, pitch period, gain, peak magnitude value, grid position, and sample amplitude. These parameters add up to 260 bits generated every 20ms. This information rate requires 13kbps to achieve the desired connection. This research implemented the HCDC compressor on the 260 every 20ms to achieve further lossless compression. We could reach data rates lower than 13kbps by 60%, reaching down to 5 kbps on average. The results were then compared to other lossless compression methods such as FLAC, and the algorithm we used showed better performance by 70% over FLAC. The research did not include any quality assessment due to the lossless nature of the algorithm.

References

- *AbdulMuin Fathiah, & Gunawan Teddy, & Kartiwi Mira, & Elsheikh A. (2017). A review of lossless audio compression standards and algorithms. Held in Malesia and Published in AIP Conference Proceedings 1883, 020006*
- *Amro Islam, & Abu Zitar Raed, & Bahadili Al-Bahadili (2011). Speech compression exploiting linear prediction coefficients codebook and hamming correction code algorithm. Springer International Journal of Speech Technology volume 14, Article number: 65*
- *Amro Islam (2013). Higher Compression Rates for Code Excited Linear Prediction Coding Using Lossless Compression. Presented in the Fifth International IEEE Conference on Computational Intelligence, Communication Systems and Networks. Madrid: Publisher IEEE*
- *Bahadili Hussein (2008). A novel lossless data compression scheme based on the error correcting Hamming codes. Elsevier international journal Computers and Mathematics with Applications. 56:143-14*
- *Bhatty P. Ninad, & Kosta P., & Kosta P. (2011). Proposed modifications in ETSI GSM 06.10 full rate speech codec and its overall evaluation of performance using MATLAB. International Journal of Speech Technology 14(3):157-165*
- *Biesmans Wouter, & Das Neetha, & Francart Tom, & Bertrand Alexander (2017). Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 25(5):402-412*
- *Chu H. (2003). Speech Coding Algorithms. New York: Wiley.*
- *European Telecommunications Standards Institute ETSI (2010). Digital cellular telecommunications system (Phase 2+) and Universal Mobile Telecommunications System (UMTS)and LTE and Network architecture V9.3.0. FRANCE:650 Route des Lucioles F-06921 Sophia Antipolis Cedex*
- *GSM Association (2020). Adaptive Multirate Wide Band Version 5.0. GSM Association Publications.*
- *HU YI, & Philipos C. Loizu (2007). Evaluation of Objective Quality Measures for Speech Enhancement. IEEE Transactions on Audio, Speech, and Language Processing 16(1):229-238*
- *Kain A., & Macon M. (2001). Design And Evaluation of A Voice Conversion Algorithm Based On Spectral Envelope Mapping And Residual Prediction. Centre For Spoken Language Understanding (CSLU), Oregon Graduate Institute, OR 97006, USA. 2000*
- *LAM Y., & Goodman J. (2000). A mathematical analysis of the DCT coefficient distributions for images. IEEE Transactions on Image Processing 9(10):1661-1666*
- *Lin Xiao, & Li Gang, & Li Zhengguo, & Chia Thien King, & Yoh Ai Ling (2001). A novel prediction scheme for lossless compression of audio waveform. Presented in IEEE International Conference on Multimedia and Expo: Publisher IEEE, Japan*
- *Luo Da, & Rui Yang, & Bin Li, & Jiwu Huang (2017). Detection of Double Compressed AMR Audio Using Stacked Autoencoder. IEEE Transactions on Information Forensics and Security. 12(2):432-444*
- *Malvar Henrique (2007). Lossless and Near-Lossless Audio Compression Using Integer-Reversible Modulated Lapped*

Transforms. Presented in Data Compression Conference DCC07: Publisher IEEE, USA

- *Menzies Dylan, & Filippo Maria Fazi (2017). Decoding and Compression of Channel and Scene Objects for Spatial Audio. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 25(11):2138-2151*
- *Openhaim et al. (1997). Signals and Systems. Second edition. NJ, USA: Prentice-Hall.*
- *Rao A.V., & Ahmadi S., & Linden J., & Gersho A. (2003). Pitch adaptive windows for improved excitation coding in low-rate CELP coders. IEEE Transactions on Speech and Audio Processing 11(6): 648-659*
- *Takehiro Sugimoto, & Shuichi Aoki, &Tomomi Hasegawa, & Tomoyasu Komori (2019). Advancement of 22.2 Multichannel Sound Broadcasting Based on MPEG-H 3-D Audio. IEEE Transactions on Broadcasting, PP (99):1*
- *Uttam Kr. Mondal (2019). Achieving lossless compression of audio by encoding its constituted components LCAEC. Springer and NASA Journal on Innovations in Systems and Software Engineering volume 15:75–85*
- *Wah Benjamin (2005). LSP-based multiple-description coding for real-time low bit-rate voice over IP. IEEE Transactions on Multimedia 7(1):167-178*
- *Wu Chou, & Bing Huang Juang (2002). Pattern Recognition in Speech and Language Processing.FL, United States: CRC Press, Inc.*
- *Yanzhen Ren; Hanyi Yang; Hongxia Wu; Weiping Tu; Lina Wang (2019). A Secure AMR Fixed Codebook Steganographic Scheme Based on Pulse Distribution Model. IEEE Transactions on Information Forensics and Security, 14(10):2649-2661*

Web-Based Market Information System for Farmers in Palestine

نظام معلومات السوق المستند إلى الويب للمزارعين في فلسطين

Yousef Saleh Abuzir
Professor/ Al-Quds Open University/ Palestine
yabuzir@qou.edu

يوسف صالح ابوزر
أستاذ دكتور/جامعة القدس المفتوحة/ فلسطين

Waleed Abdullah Awad
Associate Professor/ Al-Quds Open University / Palestine
wsalos@qou.edu

وليد عبدالله عوض
أستاذ مشارك/جامعة القدس المفتوحة/ فلسطين

Mohamad Hamdi Khdair
Lecturer / Al-Quds Open University/ Palestine
mkhdair@qou.edu

محمد حمدي خضير
محاضر/جامعة القدس المفتوحة/ فلسطين

Received: 17/01/2021, Accepted: 06/03/2021
DOI: https://doi.org/10.33977/2106-000-005-002
https://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 2021/01/17، تاريخ القبول: 2021/03/06
E-ISSN: 2521-411X
P-ISSN: 2520-7431

Abstract

The advances in information and communication technologies (ICTs) provide great tools for development in Palestine, especially utilizing the availability of internet access through web-based applications. One important sector to benefit from ICT tools is the agricultural sector. Farmers need to be empowered with technology to make the best use of the scarce resources of the farm. The main objective is to identify and analyze relevant material and information flows, production processes, and their interconnections and synergies in the agriculture sector. Market Information Systems for Farmers (MISF) is an information system used to gather, analyze, and disseminate information about agricultural yields, prices, and other information, such information is relevant to farmers, traders, dealers, and others involved in handling agricultural products. In this paper, we propose a system that will address some of the problems facing the farmer markets. This new approach will make the farmer and the buyers responsible for uploading their agricultural products, harvest data, and price information using the MISF website and the MISF application on their mobile devices. In this research, a descriptive approach was used to analyze the agricultural information marketing system. First, data related to our system was gathered using a literature review of research, reports, questionnaires, and site visits. We then used an Object-Oriented Approach to apply a system analysis and represent our solution using UML Notation, graph, figures, and tables. In this paper, we will try to address some of the problems facing the farmer markets. The proposed system will facilitate trade by creating a capacity for sellers to contact individual buyers. This system will provide information on what agricultural products are in demand by analyzing consumer consumption and market trends. The system will collect demographic details such as the types of crops grown, crop size, prices, cost, and maybe access to the type of irrigation, soil, and fertilizers as inputs from the farmers as well as other information about crops consumption. The data gathered by the proposed system can be used to advise farmers about needed crops and suggest ways to help them lower costs and improve productivity; this can be achieved using data mining techniques and maybe the Internet of Things (IoT). In general, the system will track farmers' daily activities, businesses and provide ongoing support in areas such as labor, costs, yields management, crops consumption, harvest management, market price discovery, and strong relation with buyers.

***Keywords**: Market Information Systems, Web Application, Farm Management, Farm software.*

الملخص

تُقدم التطورات في تكنولوجيا المعلومات والاتصالات (ICT) أدوات رائعة للتنمية في فلسطين، لا سيما الاستفادة من توفر الوصول إلى (الإنترنت) من خلال التطبيقات القائمة على (الويب). يعتبر القطاع الزراعي أحد القطاعات المهمة التي يمكن أن يستفيد من أدوات تكنولوجيا المعلومات والاتصالات. ويحتاج المزارعون إلى دعمهم بالتكنولوجيا لتحقيق أفضل استخدام للموارد النادرة للمزرعة. والهدف الرئيس هو المساعدة في تحديد المواد ذات الصلة بتدفق المعلومات وتحليلها، وعمليات الإنتاج، والترابط والتآزر في قطاع الزراعة. ونظم معلومات السوق للمزارعين (MISF) هو نظام معلوماتي سيتم استخدامه في جمع المعلومات حول المحاصيل الزراعية والأسعار وغيرها من المعلومات وتحليلها ونشرها، وهذه المعلومات ذات صلة بالمزارعين والتجار، والبائعين وغيرهم من المشاركين في التعامل مع المنتجات الزراعية. وفي هذا البحث، نقترح نظاما يعالج بعض المشكلات التي تواجه أسواق المزارعين. وسيجعل هذا النهج الجديد المزارع والمشترين مسؤولين عن تحميل منتجاتهم الزراعية، وبيانات الحصاد، ومعلومات الأسعار باستخدام التطبيق (MISF) على تطبيق (الويب) وأجهزتهم الخلوية. في هذا البحث تم استخدام المنهج الوصفي لتحليل نظام تسويق المعلومات الزراعية. أولاً، وتم جمع البيانات المتعلقة بنظامنا عن طريق مراجعة الأدبيات للأبحاث والتقارير والاستبيانات، وزيارة المواقع الزراعية. ثم استخدمنا المنهجية الكينونية لتحليل النظام ولتمثيل الحل الخاص بنظامنا باستخدام تدوين UML والرسم البياني والأشكال والجداول. في هذا البحث، وسنحاول معالجة بعض المشاكل التي تواجه أسواق المزارعين. وسيسهل النظام المقترح التجارة من خلال خلق قدرة للبائعين على الاتصال بالمشترين بشكل فردي. وسيوفر هذا النظام معلومات حول المنتجات الزراعية المطلوبة من خلال تحليل استهلاك المستهلك واتجاهات السوق. وسيقوم النظام بجمع التفاصيل الديموغرافية، مثل: أنواع المحاصيل المزروعة، وحجم المحاصيل، والأسعار والتكلفة، وربما الوصول إلى نوع الري والتربة والأسمدة كمدخلات من المزارعين، يضاق إلى ذلك معلومات أخرى حول استهلاك المحاصيل. ويمكن

استخدام البيانات التي تم جمعها بوساطة النظام المقترح لتقديم المشورة للمزارعين حول المحاصيل المطلوبة، واقتراح طرق لمساعدتهم على خفض التكاليف وتحسين الإنتاجية، ويمكن تحقيق ذلك باستخدام تقنيات استخراج البيانات، وربما إنترنت الأشياء (IoT). بشكل عام، وسيتتبع النظام الأنشطة اليومية للمزارعين، والأعمال التجارية، ويوفر الدعم المستمر في مجالات، مثل: العمالة والتكاليف وإدارة الغلة واستهلاك المحاصيل وإدارة الحصاد واكتشاف أسعار السوق والعلاقة القوية مع المشترين.

الكلمات المفتاحية: نظم معلومات السوق، تطبيقات الويب، إدارة المزارع، تطبيقات المزرعة.

INTRODUCTION

Accurate and easy to use Markets Information Systems for Farmers (MISF) are of fundamental importance for successful operational farm management. However, many farmers still do not use MISFs for various reasons, like lack of knowledge, absence of these systems, and the complexity of the available systems.

Developing a market information system that uses information and communication technology comes in handy for all agriculture sector stakeholders. The system can be used in the field of agriculture to provide efficient information management, flexible knowledge and information sharing, local and global communication, and production planning. It should be mentioned that researchers can use the data gathered by this system to improve the agricultural system. This ultimately results in an overall increase and improvement in the productivity in agriculture and thus the economy.

In recent years, new business models for agricultural markets have appeared. Under this perspective, there is a need for a new information system for urban markets to facilitate transactions. Both sides, consumers and farmers, require certain information from markets about agricultural products. For example, consumers may make requests about the exact information of agricultural products or their safety, while farmers may want to boast about their products. Under the considerations of such requirements at the markets, which may be conflicting, we will propose a new information system to assist in the negotiation between both parties.

The main objective of the study is to develop a market information system for farmers (MISF) and digitize it using Internet websites and smart devices. This step helps and strengthens the local agricultural system, improves productivity, improves lives, and provides jobs for farmers in Palestine. It also helps provide and create new markets and value chains, bring together a wide range of local and regional stakeholders, and strengthen relationships between farmers and trusted consumers (Abuzir, Awad, and Khdair, 2021). The market information system will play an important role in agro-industrialization and food supply chains.

Research Objectives

The main goals of this paper are:

- To study the development of Markets Information Systems for Farmers (MISF) using web-based and mobile applications in the context of a farmer application.
- To use technology to strengthen the local agricultural system and improve productivity for everyone in the agriculture value chain, including small farmers.
- To improve the lives and livelihoods of farmers in Palestine.
- To bring together a wide array of local and regional stakeholders to form a mutually beneficial value chain.
- To create access to new markets, value chains, and business models.
- To develop stronger relationships with trusted farmers and consumers.

These are the issues that were addressed in this paper:

- What problems can arise when developing Markets Information Systems for Farmers (MISF)?
- How can these problems be addressed by developing an Internet-based website and mobile applications?
- What problems and limits arise with the usage of the developed system?
- What benefits of MISF for the agricultural sector?
- What research possibilities are available using the data accumulated by the system?

This manuscript consists of seven sections. This section introduced the problem, problem statement, and the main objectives of the system, while the second section provides a literature review. The third section introduces the research methods, research instruments, and system structure. The fourth section discusses the system analysis, design and implementation. In section six, we discussed the results, while the last section provides the conclusion.

LITERATURE REVIEW

Syed Khizer (2017) conducted a study about the development of an online marketing information system for the agricultural sector of KSA. He emphasized the importance of the agriculture sector, which generates labor and capital and fulfill domestic demand to support growth in other sectors. Additionally, the agriculture sector plays a key role in ensuring national food security. Access to agricultural marketing information is an essential factor in promoting competitive markets, globalization, efficient marketing, market liberalization, and improving agricultural sector development. The majority of the stakeholders of this sector do not use agricultural marketing information. The stakeholders of the agriculture sector of Saudi Arabia need Agricultural Marketing Information System (AMIS).

Mawazo M. Magesa, Kisangiri Michael and Jesuk Ko (2014): This paper reviewed the agricultural market information services in developing countries. This study has explored the use of agricultural market information services in linking smallholder farmers to markets, especially in sub-Saharan developing countries. Origin of, the needs for, and the current status of agricultural market information services in developing countries are clearly presented. Lastly, the study explored the limitation of the success of most agricultural market information services in sub-Saharan developing countries.

C.G. Sorensen, S. Fountas, E., et al. (2010), in their paper presented a conceptual model of a future farm management information system (FMIS). The aim of this paper is to define and analyze the system boundaries and relevant decision processes for such a novel FMIS as a prerequisite for a dedicated information modeling. The boundaries and scope of the system are described in terms of actors and functionalities, where actors are entities interfacing with the system (e.g., managers, software, databases).

Mishra et al. (1999), Muhammad et al. (2004), Forster (2002), and Doye et al. (2000) covered the importance of farm management issues. The skillful and conceived management is one of the most important success factors for today's farms. Only when a farm is well managed can it generate the funds to finance its sustainable development and thereby, its survival in today's fast-changing environment. However, sophisticated management is a challenging and time-consuming task and must be organized as efficiently as possible.

Shepherd (2011) and David-Benz et al. (2011; 2016) indicated that the first-generation market information systems were mostly based on a single model, regardless of the market being studied, the type of product, and the country. Other systems often focused exclusively on price information, relied on project-based financing, and were imperatively implemented by public bodies, such as marketing boards and ministries (Rubio (2020); Nwafor et al. (2020); Muto (2009); Aker (2010); Belakeri et al. (2017); Chikuni et al. (2019); Roslin et al. (2020); Emeana et al. (2020)).

Several studies by David-Benz et al. (2011; 2012), (Galtier, 2014), and Mukhebi and Kundu (2014), showed the importance of the spread of mobile phones and the Internet, which paved the way to the rise of a new generation of Management Information Systems (MIS). The Information and Communications Technology (ICT) sector developments have made it possible to minimize the lag in transferring price data from collection points to Central Processing Units (CPUs) and disseminating information to the intended recipients. MIS that uses ICT has become known as the "second generation" MIS, or the 2GMIS "Second Generation of Management Information Systems.

Artificial Intelligence (AI) and Data Mining (Abuzir, 2018) are expected to report significant growth in the near future in the agricultural industry. Farmers can track their livestock in real-time by making use of AI. Dairy farms can now individually monitor the behavioral aspects of animals with AI solutions, including picture classification with body condition score and feeding patterns and facial recognition for

livestock. Furthermore, farmers use machine vision that allows them to identify facial features (Global, 2020).

MATERIALS AND METHODS

Method and Data Source

Based on our previous study (Abuzir, Awad, Khdair, 2021), as well as the main findings and recommendations of that study, it was concluded that ICT could play an important role in promoting and developing central markets in Palestine by organizing and saving time, effort, and money in all sectors. In this research, we implemented the recommendation by creating a web-based market information system and mobile application to benefit and for the use of the concerned parties.

In the first step, we reviewed historical and recent literature to understand and analyze Market Information Systems. Then we used a questionnaire to collect data from farmers and the other stakeholders in the following governorates Ramallah, Nablus, and Salfit. Then, we applied a system analysis to identify and analyze all the material and information flows, the production processes, and their interconnections and synergies.

Data related to the system were collected from different sources using various instruments and techniques as a literature review of journal articles, reports, legislation, case studies, on-site visits, and questionnaires for farmers and administrators.

We consequently gathered information about farmers containing all relevant data related to our study. Moreover, the collected data provided the basis for the development of the Market Information System, which describes all relevant factors of the system like input and output, reports, resources, production processes and activities, services, and administration.

In this research, a descriptive approach was used to analyze the agricultural information marketing system. Data were identified, grouped, and classified in order to answer problem questions and identify suggested solutions. We represented the results of the analysis and the solutions using UML notation. In addition, the solutions and the analysis are represented in various formats such as shapes, charts, and tables.

Proposed System Structure

The requirements analysis started with identifying and defining the scope and objective of the system for Market Information System for Farmers (MISF) based on a study of Abuzir, Awad, and Khdair (2021). So, the requirements for MISF were derived from that research. In their study, a survey was conducted to assess the user's expectations on future Information Technology, Internet Functions, and online services for farmers.

A questionnaire has been performed and analyzed, collecting feedback on the interpretation of the reality of the agricultural markets in Palestine, the levels of difficulties and problems facing the current distribution system, to what extent do the market information system contribute to achieving an efficient system, and to what extent are the technological requirements available for the market information system? These questions were studied and analyzed using SPSS to develop the conceptual architecture of MISF. A literature analysis was done on developing the agricultural system of the markets information system for farmers in Palestine. Table 1 presents a summary of the results of the different statistical methods used in this study to analyze the data collected by the questionnaire.

The results in Table 1 show that there is a general agreement among the different parties in the market (farmers, traders, specialists) that the current traditional market system is not efficient. It provides limited possibilities regarding marketing and information. Among the first core area, this item: "The local bodies that supervise the markets are effective and efficient," scored the lowest level of agreement at 43.4%, with a mean of 2.17, which is the lowest value for the first core area, thus indicating clearly the need for a new efficient system.

The second core area shows that the kind of difficulties facing the agricultural sector is related to the conventional system in use. All these problems can be solved using a well-designed information system. An example is a response to this item: Administrative costs for import (insurance, transportation, freight, fees, etc.) are reasonable, which scored the lowest level of agreement at 42.3%, with a mean of 2.11, which is the lowest value in the second core area.

Table 1: A Summary of the Results of Analysis of the Data Collected by the Questionnaire

Sub Question (Core Area)	Mean	Standard Deviation (SD)	Percent %	Attitude
Total Score of: The reality of the markets in Palestine	3.0714	0.37440	61.4%	**Medium**
Total Score of the: The levels of difficulties and problems facing the current distribution system	3.4286	0.36444	68.6%	**Medium**
Total Score of: To what extent does the market information system contribute to achieving an efficient system	3.9086	.77427	78.2%	**High**
Total Score of: To what extent are the technological requirements available for the market information system?	4.0980	.69622	82.0%	**High**

Another indicator is the answer to this item which scored the highest level of agreement: The ability to promote efficient agricultural transactions and contact agricultural supplies' companies, at 84.0%, showing an over-willing agreement to use the MISF.

Finally, the results for the fourth core area show that there is a high level of agreement for the role that is expected by technology is highly positive and will improve the overall performance of the market.

Besides the results of that study, the requirements analysis, the reference architecture for MISF was designed. The proposed system is shown in Figure.1. An approach that is user-friendly and fast to monitor and fulfill the user requests is implemented using a centralized server to store all relevant data. It includes various databases such as users, lands, agricultural supplies selling points, agriculture directorates, weather, research, crop, and farmers-related data that can all be stored at a single location on the server, thus making it available to all intended users. This data can be easily accessed by the end-users such as farmers, experts, consultants, researchers, etc., at any time from any location through computers or smartphone devices that are connected to the system. A generic user interface can be used that facilitates accessing the system for information.

The system was developed for the different stakeholders comprising the farmer, agricultural trader, agriculture production factory, Ministry of Agriculture, agriculture directorate, and agricultural supplies selling points.

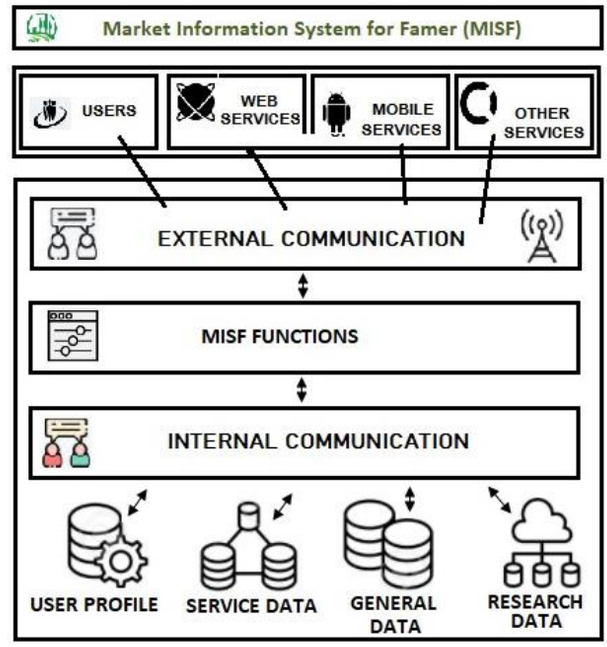


Figure.1. Proposed System

Information management (related to crop, soil, and production process) is easier as it will be managed by the service providers, a farmer, while other information as weather, wind type, wind speed, wind direction, and humidity are managed by the system. MISF obtains weather data from one of the most famous global sites specialized in weather forecasts via web API services to analyze and present it in a simple framework that helps the farmer and those working in the agricultural sector to understand the data and benefit from it. The system provides users with another information that is generated by the directorates of agriculture and the ministry of agriculture, such as planting seasons, dates for harvesting agricultural crops, dates for spraying pesticides and medicines, and fertilizers, announcements about the availability of seedlings, announcements from the ministry, farmers or agricultural instructions.

The effective use of information technology in the agricultural sector will bring positive changes by utilizing its key features.

The following benefits may serve as reasons as to why Information technology is necessary to be implemented in the agricultural sector:

1. Information management (related to crop, soil, weather, and production process) is easier as it will be managed by the service providers and promotes the circulation of agricultural products.
2. The system will reduce the long supply chains and complex links between farmers and consumers, making it difficult for the farmers to derive benefits and value from the markets.
3. Data availability at any time and at any location.
4. Technical issues will be reduced as the system handles them.

Figure 1 provides a high-level view of the Market Information System for Farmer architecture design. It describes the generic structure (concepts and relations) of our system. A Market Information System for Farm architecture structure is described according to the requirements and specifications that were defined in the next Section, "System Analysis and Design." The architecture comprises five main components:

- Users and Services
- Interface Communications
- MISF main Functions
- Internal communication
- Databases

Figure 1, summarizes the core of MISF and the framework for the MISF. It is an independent software running on the user's computer and smartphones with connectivity to the MISF database using a complete web-based MISF application. The MISF offers other services to users as the ministry of agriculture and moderator. The MISF stores users' profiles and data generated by services in their own format in their databases. The MISF, on the other hand, is an application framework that provides generic functionalities for service providers to offer different services to users. The MISF provides functionalities for adding crops, land, or services into a marketplace where other users can discover and use these services. Furthermore, the MISF provides a vertical communication (interface) enabler for communication between different services into the MISF based on the service usage.

In the next sections, an illustrative is provided to show how the different components of MISF fit together in practice.

SYSTEM ANALYSIS AND DESIGN

We used Object-Oriented Development Approach to developed MISF. In this approach, the system users are identified as "Actors," and the different functions for each user are considered "use cases". The system's input and output screens were designed and linked to the databases. The system provides the users with security options to protect their accounts. MISF was implemented to protect the privacy of the users by securing access to the system and its data. It preserves privacy for all system users by protecting their menus, privileges and accessing their data.

Design of the System

In our analysis, we used a Unified Modelling Language (UML) Use case Diagram as a preliminary step to create an overview of the system without giving more detail about the system. This diagram (Figure 2) normally consists of the overall application dataflow and functions or processes of the MISF. It contains all the users and their interaction with the system.

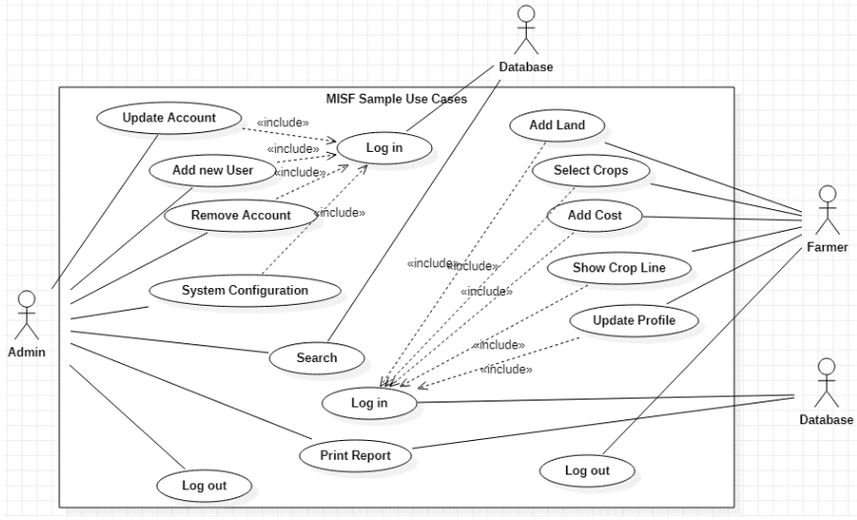


Figure 2 A sample Use Case Diagram for Farmer, Admin and Database Actors in MISF

A table containing all the database components that are part of the Market Information System for Farm reference architecture can be found in Table 2.

As shown in Table 2, we have 18 entities. The detailed data of MISF is stored in these tables. Each entity has different attributes and contains a primary key as a unique key, and may have a foreign key. The entity crops contain primary key cropID and fieldID as a foreign key. All the entities are normalized and reduce duplicity of records.

Crops (cropID, fieldID, cropPrice, cropAmount, cropUnit, cropDate, cropAdmin, cropBuy, data_in)

Based on our analysis of the system and the results of the survey, the design of the web interface is generated and includes two parts: A main screen serving the general public with general information for the different stakeholders and services listed on it, and the second part is a control panel protected by a username and password to enable administrators to carry out system related tasks. Listed here a screenshot of the main screen (Figure 3).

Table 2: Database Structure

Entity	Attributes
AccountInfo	accountInfoID, accountID, infoType, acountInfoValue, accountInfoTime,
AccountType	accountTypeID, accountAdmin, accountTypeName, accountTypeStatus accountTypeTime,
Accounts	accountID, accountType, accountName,accountLogINTime, accountStatus, accountIdCard, accountPass, cityID, addrID, accountTimeRegist, data_in,
Addr.	addrID, cityID, addrName, addrLat, addrLng
City	cityID, cityName
Crops	cropID, fieldID, cropPrice, cropAmount, cropUnit, cropDate, cropAdmin, cropBuy, data_in
Fertilizer	fertilizerID, fertilizerName, fertilizerTime, fertilizerAdmin, data_in
FieldCost	fieldCostID, fieldID, varDataID, fieldCostTime, accountID, fieldCostValue
Fields	fieldID, fieldName, accountID, varDataID, AgriType, fieldDate, fieldArea, AgriCost, fieldHarvest, fieldStatus, data_in
InfoType	infoTypeID, infoTypeName, infoTypeStatus
Msgs	msgID, msgSender, msgReceiver, msgBody, msgTime, msgStatus
News	newsID, newsType, newsTitle, newsBody, newsTime, newsImg, newsCity, newsStatus
Points	varDataID, accountID, cost,size, AddTime
RequestCrops	RcropID, cropID, RcropAmount, RTime, RacountID, FamerID, reuqestStatus, ResStatus
SystemVars	systemVarID, systemVarName, systemVarStatus, systemVarTime, systemVarAdmin
VarData	varDataID, varDataValue, varDataTime, varDataAdmin, systemVarID
Website PIC	PICID, PICName, PICCreateTime, PICSection, TextDET

Figure 3. The Homepage of the System.

Besides the functional requirements that are mentioned so far, there are several functional requirements to be addressed, such as user management, data security, or routing of information. These requirements are crucial for MISF.

System Implementation

This section describes the design and development of an agricultural market information system for the farmers in Palestine (MISF). Access to this system is provided through the Internet. The application layer protocol that is Hypertext Transfer Protocol (HTTP), is used to transmit all files (HTML files, image files, query results, etc.) on the World Wide Web. The users will be able to view the required information related to the different activities of farming.

The front-end of this system uses JQuery and HTML for its user interfaces, while the back-end uses a MySQL database to manage its data. The front-end and back-end of this system are connected using a MySQL driver. The data retrieval and update of this system are done using MySQL queries.

Since the database will require updating by non-computing proficient personals, the system provides easy access to the database for all types of data manipulation. Security of the database is ensured by the use of a password for updating purposes, which will be given to the different users of the system. The System provides the external user the ability to obtain summarized information in a preferred format. This can be produced for certain crop types for any given year.

The set of tables is created using the relational database using MySQL for the identified entities at the design stage. The uniqueness of the data fields in these tables is established using primary keys, while the relationships are maintained using foreign keys.

Figure 3 presents the home page for our market information system, MISF. Different functions or information are chosen through a hyperlink of the main page.

Making queries for MISF can be done using specified appropriate dialog, form, edit, and list boxes. Required query statements are constructed automatically by the system, and the users need not be aware of them. The following describes a sample query statement issued for the query to show the farmer's name. Edit box options are filled and passed to the query statement using variables of the SQL language. These variables are shown in italic font style.

SELECT *accountName* FROM `accounts` where *accountID*='12123'

Users can update the database related to their farm, crops, land and other information by login to their account by specifying the username and password (Figure 4) through the web browser.

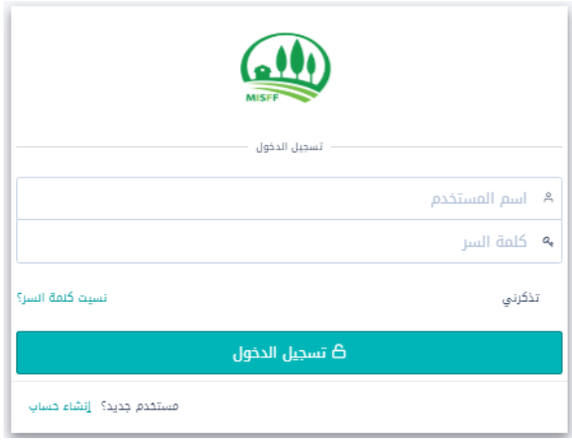


Figure 4. Login Screen for the user

Another sample is the farmer account (Figure 5). The farmer can define what he owns of agricultural lands and determine the types of crops planted on the lands. This is in addition to scheduling dates for picking crops. During the cultivation period, the operating expenses are added to reach the harvest period to help set a minimum cost price.

The software application on smart devices allows the system user to manage all the different operations, such as following up orders and adding the number and quantity of crops from his account. In this way, the merchant can search for and make an order of an offered crops. (Figure 6).

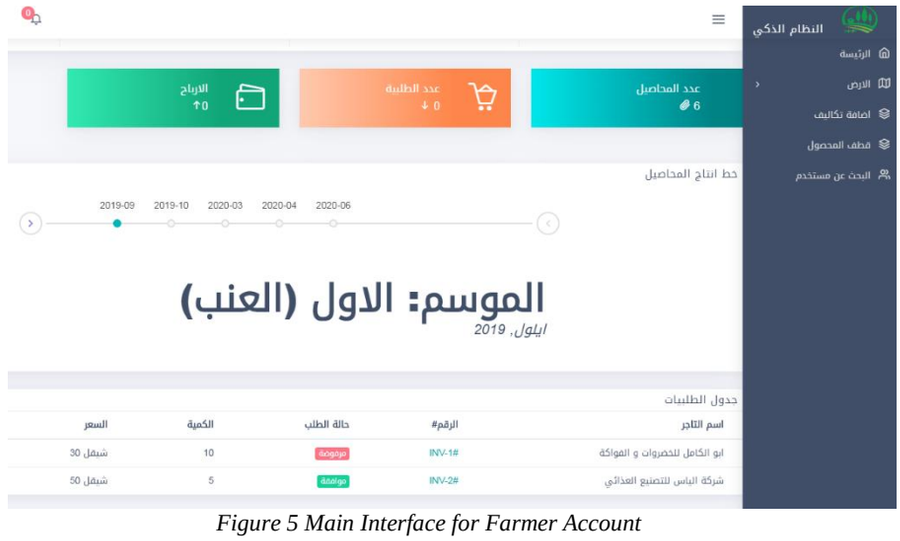


Figure 5 Main Interface for Farmer Account

Figure 6 Sample Interface for a farmer account

A static view of the runtime configuration of the processing nodes and the modules running on those nodes is illustrated in the deployment diagram. The deployment diagram in our framework shows the MISF hardware, the software that is installed on that hardware, and the middleware used to connect the various machines to each other. Figure 7 shows the deployment diagram.

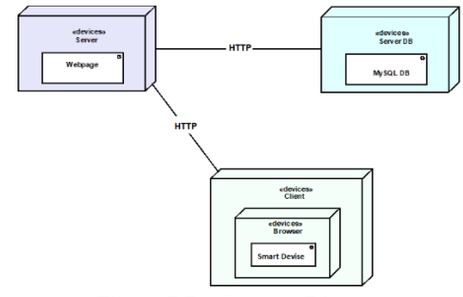


Figure 7 Deployment Diagram

RESULTS AND DISCUSSION

The proposed system was designed to provide reasonable, technology-based solutions for the difficulties facing agriculture in Palestine. One key design objective was efficiency; it was shown in the previous studies that the Palestinian agriculture market system suffers from many problems, all because the right information is not available when needed, and the current market mechanisms are outdated. The proposed system addressed these issues directly by providing a comprehensive, modern, and efficient solution to all the stakeholders in the market. The system stores data about crops, soil, irrigation, fertilizers, weather, pesticides, etc. Then this data is analyzed to advise farmers on the most suitable crops needed in the market and provide an easy and fast way to connect buyers and sellers. The data can also be used for future research and development projects concerning the agricultural sector in partnership with the Ministry of Agriculture.

Currently, this system is available online and runs its web server on the WEB. An administrator user maintains the system, supervising the different tasks of the system and the database. The ultimate objective is to allow this system to be used by as many users as possible in the agricultural sector. Having more users in the system will make its database more valuable and efficient.

The main stages for system development can be summarized as follows:

- Creating system database to support user types, Palestinian cities, product types, …
- Adding the data for Palestinian cities and villages to be used by the system.

The different types of user accounts have been identified and created with the right permissions for each account. A control panel was created to enable the system administrator to carry out the required admin tasks.

System protection was considered during the coding of the project to preserve the data and prevent unintended use of the system. In addition, user permissions were set to determine what kind of data and tasks each user can do according to his role.

The main outcomes of this study

- Software that helps improve farmer productivity
- Technology advises for farmers
- Developing stronger relationships between farmers and consumers
- Supporting the agriculture ecosystem in Palestine.
- Opportunities and benefits to a wide range of stakeholders, from small farmers to businesses and government.
- Ability to optimize market efficiency by connecting buyers and farmers.

This research has developed a proposed architecture for the MISF to contribute to the agriculture sector. This architecture can be used to map, plan, design, and implement a real farmer's market information system that meets the requirements set out in this study.

MISF suits the different stakeholders' needs, including an easy adaptation, user-friendliness, and accuracy in depicting the various production processes, management, and services.

The focus of MISF is to perform farm activities based on all farm transactions. Different users or stakeholders in the agricultural sectors can use the system. The application was successfully implemented using web technology and smart devices and tested where all different scenarios were recorded.

The evaluation of the proposed architecture contains two parts. First, the proposed architecture was verified based on the requirements. Second, a conceptual validation that maps the functionality of the proposed architecture was tested using two applications. These applications are based on web bases and smartphones.

CONCLUSION

This work is an initial study to show that the creation of a Market Information System for Farmers is feasible. The real benefits of this type of information system to the agricultural sector in Palestine can be seen when it becomes operational for all stakeholders: Farmer, agricultural trader, agriculture production factory, Ministry of Agriculture, agriculture directorate, and agricultural supplies selling points. This system will promote more hope for importers, exporters, and researchers who will access the updated information.

The main importance of the system is providing information on what agricultural

products are in demand by analyzing consumer consumption and market trends using Data Mining techniques. With this information, farmers would have a better idea of what crops to prioritize. This can also help stabilize the economic sustainability of farming by improving farm management. With the system at work, it will reduce oversupply and undersupply of certain agricultural products, and the stable supply-demand relationship will prevent the underpricing of agricultural products and help in stabilizing market prices.

ACKNOWLEDGEMENT

This study was supported by the Project Support Program for Research and Development Innovation by the Ministry of Higher Education in Palestine. We also appreciate the support of al-Quds Open University to carry out this work.

References

- *Abdallah W., Khdair M., Ayyash M ., Issa , A. IoT system to control greenhouse agriculture based on the needs of Palestinian farmers, International Conference on Future Networks and Distributed Systems, June 2018.*
- *Abuzir Y., Awad W., and Khdair M. (2021), Market Information System for Farmers, Palestinian Journal of Technology and Applied Sciences, under review (2021).*
- *Abuzir Y., Predict the Main Factors that Affect the Vegetable Production in Palestine Using WEKA Data Mining Tool, Palestinian Journal of Technology and Applied Sciences, pp 58-71, No 1 (2018).*
- *Aker, J. (2010). Information from markets near and far: mobile phones and agricultural markets in Niger. American Economic Journal: Applied Economics, 23.*
- *Belakeri P., Prasad C. K., Bajantri S., (2017) Trends of Mobile Applications in Farming, July 2017 International Journal of Current Microbiology and Applied Sciences 6(7):2499-2512 DOI: 10.20546/ijcmas.2017.607.295*
- *Chikuni, T.; Kilima, F. (2019) Smallholder farmers' market participation and mobile phone-based market information services in Lilongwe, Malawi. Electr. J. Inf. Syst. Dev. Ctries. 2019,*
- *COMCEC(2018) Improving Agricultural Market Performance: Developing Agricultural Market Information Systems, Comcec Coordination Office February 2018*
- *David-Benz H., Galtier F., Egg, J., Lançon, F. & Meijerink, G. (2011) "Market Information Systems: Using information to improve farmers' market power and farmers organizations'voice". Available at: http://www.esfim.org./wp-content/uploads/policy-brief7-english.pdf.*
- *David-Benz H., Egg J., Galtier F., Rakotoson, J., Shen, Y., & Kizito, A. (2012). Information systems on agricultural markets in sub-Saharan Africa: from the first to the second generation. (AFD, Ed.) Paris: Focales 14.*
- *David-Benz H., Andriandralambo N., Soanjara H., Chimirri C., Rahelizatovo N. & Rivolala B. (2016). Improving access to market information: a driver of change in marketing strategies for small producers? Paper prepared for the 149th European Association of Agricultural Economists Seminar on Structural change in agri-food chains: new relations between farm sector, food industry and retail sector. 27-28 October 2016: Rennes, France.*
- *Emeana E. M., Trenchard L., and Dehnen-Schmutz K. (2020), The Revolution of Mobile Phone-Enabled Services for Agricultural Development (m-Agri Services) in Africa: The Challenges for Sustainability, 8 January 2020.*
- *Galtier, F., David-Benz, H., Subervie, J., & Egg, J. (2014). Agricultural market information systems in developing countries: new models, new impacts. Cahiers Agricultures, 232-244.*
- *Global AI in Agriculture Market Research Report 2020-2030 (2020), November 2020, Research and Markets*
- *Goyal A., Information, Direct Access to Farmers, and Rural Market Performance in Central India July 2010*
- *Kashima T., Matsumoto S., Iseda H., Ishii H. (2012) A Proposal of Farmers Information System for Urban Markets. In: Watada J., Watanabe T., Phillips-Wren G., Howlett R., Jain L. (eds) Intelligent Decision Technologies. Smart Innovation, Systems and Technologies, vol 15. Springer, Berlin, Heidelberg.*
- *Khizer S. (2017), A Study on Development of Online Marketing Information System for Agricultural Sector of KSA, International Journal of Engineering and Computer Science, Volume 6 Issue 6 June 2017, Page No. 21703-21707.*
- *Kruize, J.W.; Wolfert, J.; Scholten, H.; Verdouw, C.N.; Kassahun, A.; Beulens, A.J.M. (2016). A reference architecture for Farm Software Ecosystems. Computers and Electronics in Agriculture, 125(), 12–28. doi:10.1016/j.compag.2016.04.011*
- *Magesa M. M., Michael K., and Ko J. (2010), Agricultural Market Information Services in Developing Countries: A Review, ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 3, No.9, May 2014.*
- *Mukhebi, A. & Kundu, J. (2014). Linking farmers to markets in Kenya: The evolving KACE model. Cahiers Agricultures, 23: 282-7. Available at: http://www.cahiersagricultures.fr/articles/cagri/pdf/2014/04/cagri2014234-5p282.pdf*
- *Muto, M. & Yamano T. (2009), The Impact of Mobile Phone Coverage Expansion on Market Participation: Panel Data Evidence from Uganda. World Development, 37: 1887-1896.*
- *Available at: http://www.sciencedirect.com/science/article/pii/S0305750X09000965*
- *Nwafor C. U. ,Ogundeji A. A., and Westhuizen C. (2020), Adoption of ICT-Based Information Sources and Market Participation among Smallholder Livestock Farmers in South Africa, Agriculture 2020, 10(2), 44; https://doi.org/10.3390/agriculture10020044*
- *Roslin N. A., Che'Ya l. N., Ismail M. R. (2020), Development of an Android Application for Smart Farming in Crop Management, August 2020, IOP Conference Series Earth and Environmental Science 540:012074 DOI: 10.1088/1755-1315/540/1/012074*
- *Rubio S.- V. and Rovira-M. F. (2020), From Smart Farming towards Agriculture 5.0:A Review on Crop Data Management, Agronomy 2020, 10, 207; doi:10.3390/agronomy10020207*
- *Shepherd A.W. (2011). Understanding and Using Market Information. Marketing Extension Guide No. 2 (updated version originally published in 2000). FAO Publication: Rome.*
- *Sorensen C.G.; Fountas S.; NashE.; PesonenL.; BochtisD.; PedersenS.M.; Basso B.; Blackmore S.B. (2010). Conceptual model of a future farm management information system. , 72(1), 37–47. doi:10.1016/j.compag.2010.02.003.*

Assessment of Genetic Relationships in Some Syrian Pistachio Cultivars and Genotypes, *Pistacia vera* L., Based on ISSR Markers

العلاقات الوراثية لبعض الأصناف والطرز المؤنثة من الفستق الحلبي في سورية اعتماداً على معلمات الـISSRs

Najwa Motaeb Alhajjar
Researcher / General Commission for Scientific Agricultural Research / Syria
najwa81hj@yahoo.com

نجوى متعب الحجار
باحث/ الهيئة العامة للبحوث العلمية الزراعية / سورية

Bayan Mohammad Muzher
Researcher / General Commission for Scientific Agricultural Research / Syria
bmuzher@hotmail.com

بيان محمد مزهر
باحث/ الهيئة العامة للبحوث العلمية الزراعية / سورية

Received: 07/02/2021, Accepted: 17/05/2021
DOI: https://doi.org/10.33977/2106-000-005-003
https://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 2021/02/07، تاريخ القبول: 2021/05/17
E-ISSN: 2521-411X
P-ISSN: 2520-7431

Abstract

It was clearly observed that the performance of the commercial pistachio genotypes was confusing within each variety according to the accredited substantial criteria of international descriptors. Therefore, the current work aimed to assess the genetic variation among 10 genotypes and cultivars, including 4 Ashouri, 2 Batouri, Ajami, Beid alhamam, and Ras alkharof across fifteen ISSR primers in Sweida Research Center 2019-2020. All of the used primers were polymorphic, which revealed a total of 148 bands, 141 of them were polymorphic (95.27%). The number of bands ranged from 6 to 15, with an average of 9.87 bands for each locus. Genetic similarity among all studied genotypes and cultivars ranged from 0.31 to 0.73. Depending on the UPGMA algorithm and the Dice equation, the cluster analysis divided the studied material into three main clusters. The first and second clusters comprised the following white genotypes: White Batouri, Batouri cultivar, Grahi, Beid alhamam, and Ras Alkharof that are analogous in many morphological characters, and the third cluster contained all other genotypes and cultivars: Ajami, White Ashouri, Ashouri cultivar, Ashouri Abureha, and Ashouri Mawardi. The current results demonstrated the efficiency of the ISSR technique by revealing the genetic variation among *P. vera* genotypes and cultivars and separating all of them into standing apart clusters according to their resembling appearance.

Keywords*: Pistacia vera* L., genotypes, ISSR, genetic similarity, clustering.

الملخص

يعتبر الخلط الوراثي ووجود طرز عدة ضمن الصنف الواحد من أهم التحديات التي تواجه زراعة الفستق الحلبي عند اعتماد المعايير الأساسية في توصيف الأصناف وفق المواصفات الدولية القياسية؛ لذا هدفت الدراسة إلى تقدير التباين الوراثي بين عشرة طرز وأصناف (4 طرز من الصنف (عاشوري)، 2 طراز من الصنف (باتوري) وبقية الأصناف هي: عجمي، بياضي، بيض الحمام، رأس الخروف) من خلال تطبيق (15) مركزاً في تقنية الـ ISSR في مركز البحوث العلمية الزراعية في محافظة السويداء (2019- 2020). كشفت كافة المركزات المستخدمة التعددية الشكلية حيث أعطت (148) حزمة، كان من ضمنها (141) حزمة متعددة شكلياً (95.27%). تراوح عدد الحزم من (6) إلى (15) حزمة، بمتوسط (9.87) حزمة لكل مركز، وتراوحت درجة التشابه الوراثي بين (0.31 و0.73) قسم التحليل العنقودي العينات المدروسة اعتماداً على طريقة (UPGMA) ومعادلة (Dice) إلى ثلاث مجموعات رئيسة، شملت المجموعتين الأولى والثانية الطرز والأصناف بيضاء اللون، والتي تبدي تشابهاً ظاهرياً فيما بينها في بعض الصفات، بينما ضمت المجموعة الثالثة بقية الطرز والأصناف. أثبتت النتائج كفاءة تقنية الـ (ISSR) في كشف التباين الوراثي بين أصناف وطرز الفستق الحلبي، وفصلها إلى مجموعات منفصلة بما يتوافق مع درجة التشابه المظهري لها.

الكلمات المفتاحية: الفستق الحلبي، الطرز الوراثية، ISSR، التشابه الوراثي، التحليل العنقودي.

INTRODUCTION

The leading world producers of pistachio are Iran, the USA, Turkey, and Syria (Fares *et al*., 2009). Within the genus *Pistacia*, *P. vera* L. is counted as the only comestible and vendible species (Al-Saghir and Porter, 2012). The most important economic cultivars in Syria are Ashouri, which covers over 75% of the pistachio acreage, and Batouri cultivar, which covers about 15% of the cultivated area. In contrast, the remaining acreages are cultivated with other local cultivars such as Olaimi, Bondokii, Nab-jamal, Ajami, and other marginal cultivars. The genetic variance of the *P. vera* L. species is huge, and the locally main cultivars are not pure. Many genotypes belong to the same basic name as Ashouri wardani, Ashouri Abushawka, Ashouri kafer, small and large or common Batouri, white Batouri, and red Batouri. This genetic assortment affects the specific behavior of each cultivar and genotype and creates many difficulties while certifying thorough credence for the normative staple characters. The same problem is presented in other produced countries as in Tunisia, where the most commonly cultivated variety is Matueur, which resembles the Syrian variety Ashouri (Ghorbel *et al*., 1998). This variety includes three main genotypes: Male precocious 25 A, male late 40 A, and female 11 D (Ghorbel and Kchouk, 1996). Relatedly, the inventory and identification of *Pistacia vera* L. in Algeria face taxonomic confusion problems (Kebour *et al*., 2012). Definitely, pistachio production fluctuates from one season to another due to the alternate bearing occurrence and climatic conditions. The agro-morphological

description of the pistachio is a leisurely growing tree. Nonetheless, its longevity exceeds 150 years. Great attention has to be directed towards preserving and evaluating pistachio genetic resources (Alhajjar *et al*., 2017). Furthermore, many genetic genotypes and marginal cultivars with basic characters are being neglected and face a serious risk of being lost. Works on pistachio breeding programs have been increased for the last few years (Alhajjar *et al*., 2016). Positively, there are good prospects for obtaining outstanding cultivars through crossing superior male and female cultivars from different species (Alhajjar *et al*., 2015). Therefore, more efforts have to be prearranged for genetic studies. Despite the revelation of several varieties, the morphological traits remain inconstant criteria that the same cultivar could be expressed in different characters according to the environmental conditions. Under these considerations, the precise description of the cultivars becomes very difficult, and the problem of varietal identification becomes complicated for improvement. For the last few decades, molecular markers have been applied on pistachio and its wild relatives to detect the DNA polymorphism, genetic diversity, and sex determination using either SSR (Alhajjar *et al*., 2017; Alhajjar and Muzher, 2017) or ISSR and RAPD techniques (Ehsanpour *et al*., 2008; Esfandiyari and Davarynejad, 2001; Kamiab *et al*., 2014). Hereafter, inter simple sequence repetition is a semi-arbitrary technique that seems to have the reproducibility of SSR markers for the cause of its longer length of their primers (Noroozi *et al*., 2009). Amplification in this method leads to multi-locus and exceedingly polymorphous outlines (Kafkas and Topaktas, 2003; Kafkas *et al*., 2006). However, the aspects of the recent investigation concern the determination of the genetic polymorphism of a collection of *Pistacia vera* L. Female genotypes based on genetic markers to assess the genetic diversity among some locally Syrian cultivars and genotypes.

MATERIALS AND METHODS

This investigation was carried out at the General Commission for Scientific Agricultural Research, Sweida Research Center in molecular biology laboratory during 2019- 2020.

Plant Material

The study was applied on 10 pistachio cultivars and genotypes, which have been planted in an experimental field since 1998, including 4 Ashouri genotypes, Ashouri Mawardi, Ashouri Abureha, White Ashouri, and common Ashouri, 2 Batouri genotypes (white Batouri, and common Batouri), Ajami, Beid alhamam, Ras alkharof, and Grahi cultivars.

METHODS

DNA Extraction

Samples of young leaves of all investigated genotypes and cultivars of *P. vera* were collected (a half gram of each sample), and DNA extraction was done by using the CTAB protocol (Porebski *et al*. (1997). DNA quantity and quality were estimated using a spectrophotometer (Eppendorf, Germany) by measuring the absorbencies at A260 and A280 nm.

Applying of ISSRs Primers

Fifteen ISSRs primers (Table1) were used, and the amplified reactions were done in a 25 μL volume containing 10X PCR buffer; 100 mM Tris-HCl (pH 8.4), 500 mM KCl. 2 mM of each of the dNTPs, 10 pmol primer, one unit of Taq DNA Polymerase enzyme (*Go taq*) and 50 ng of genomic template DNA. The cycling parameters were as follows: one cycle of 95° for 4 min 35 cycles of 94◦C for one min, annealing temperature for one min ranged between 38- 58°C according to GC/TA percentage of each primer, and 72◦C for one min, followed by 4 minutes at 72◦C for an extension. PCR products were injected in 1.0% agarose gel using gel documentation (VILBER LOORMOT Germany) and then were visualized after exposure to UV rays.

Genetic Analysis

The amplified bands were scored either as present or absent. The genetic similarity between any two genotypes was calculated from the bands across the 15 ISSR markers using the Dice similarity coefficient (Dice, 1945) using the PAST program. Polymorphism percentage was estimated according to the equation: the number of polymorphic bands / the total number of amplified bands × 100. A dendrogram was carried on using the UPGMA method.

Table 1 ISSR primers applied on female pistachio cultivars and genotypes and their repeat motifs

	Primer	Repeat Motif	Tm (GC%)
1	ISS2	(GA)5GC	49.17
2	ISS3	(CA)5 GT	45.75
3	ISS5	(GAA)5	50.17
4	ISS6	(AC)8 CG	66.78
5	ISS7	(AC)8TA	62.22
6	K11	(CA)6 AG	53.79
7	K25	(AG)8 G	63.50
8	K26	(AG)8T	61.09
9	K24A	(GA)8T	61.09
10	K24B	(CA)8T	61.09
11	UBC840	(GA)8TT	62.22
12	A2	(GA)6CC	62.22
13	A4	(AG)10T	68.88
14	A5	(CA)6GT	53.79
15	A6	(CT)10G	70.83

RESULTS AND DISCUSSION

ISSR banding patterns for assessing the polymorphism:

The 15 ISSR primers produced a various number of DNA fragments according to their sequence repeat motifs. The number of amplified fragments throughout all used primers ranged from 6 bands, ISS7 and K26, to 15 bands (ISS6 primer), giving an overall number of 148 bands, out of which 141 bands were polymorphic, and the polymorphism percentage was 95.27% as it is illustrated in Table 2. The recent results were in accordance with Baghizadeh and Dehghan (2018), who used 15 ISSR primers on 20 pistachio genotypes pertaining to four commercial cultivars. The number of total bands was 131 bands, and 124 of them were polymorphic with a polymorphism percentage of 94.6%. Noroozi et al. (2010) studied 31 pistachio cultivars and genotypes using three ISSR markers that amplified 28 bands, 13 of them were polymorphic, giving a polymorphism percentage of 46.42%.

The primer ISS6 amplified 15 bands as all of them were polymorphic (100%), followed by the primer UCB840, which amplified 13 bands, and similarly, all of them were polymorphic (100%). Besides, the primer K24A produced 13 bands, 12 of them were polymorphic 92.31% (Table 2). The primer K24B produced 12 bands (Figure1), where 10 were polymorphic (83.33%). The band size ranged between 209- 1208 bp. Tagizad et al. (2010) applied 10 ISSR primers on 19 pistachio cultivars. The percentage of polymorphism of the used primers ranged between 37- 92%, and the number of amplified bands was amplified 8- 12 bands for each primer. On the other hand, Turhan-Serttas and Ozan (2018) mentioned low bands size compared to our current results that ISSR primers detected 81 bands in a range of 161-188 bp only and polymorphism percentage was 96.3%.

Table 2 The number of total amplified and polymorphic bands, polymorphism percentage and band size (bp)

	No. of amplified bands	No. of polymorphic bands	Polymorphism %	Band size bp
ISS2	8	8	100	237-715
ISS3	9	8	88.89	405-1150
ISS5	10	10	100	249-660
ISS6	15	15	100	247-921
K25	7	7	100	322-1021
A4	9	9	100	350-663
A5	8	8	100	495-1138
A6	12	12	100	209-592
K11	13	12	92.31	212-974
ISS7	6	5	83.33	414-693
K26	6	6	100	416-895
K24A	13	12	92.31	248-567
K24B	12	10	83.33	372-1208
UBC840	13	13	100	225-1017
A2	7	6	85.71	298-818
Total	**148**	**141**	**95.27**	
Average	**9.87**	**9.4**		

Genetic Similarity

The percentage of genetic similarity ranged from 0.31 Beid alhamam and Grahi cultivars to 0.73 Ash. Mawardi and Ash. Abureha genotype, also between white Ashouri and the comparative Ashouri cultivar. Within Ashouri's genotypes, the average percentage of polymorphism was 0.638. The polymorphism percentage between white Batouri and the comparative Batouri cultivar was 0.64. Ghrahi cultivar occasionally missed up with Batouri cultivar with a genetic similarity of 0.61 with the reasonable Batouri as seen in Table 3. Compared to previous studies, Fares et al. (2009) referred to a high percentage of a coefficient similarity that reached 0.857 between Meknessy and Red Aleppo (Ashouri) cultivars and 0.750 between Kermezi and Kerman cultivars using ISSR markers. Mahmoodnia and Malekzadeh (2017) indicated that genetic similarity percentages ranged between 25- 78% across 12 ISSR primers carried out on 56 male and female pistachio genotypes. However, Amirebrahimi et al. (2017) referred to an adjacent genetic similarity among 56 male and female pistachio genotypes by 12 ISSR primers ranging between 0.25- 0.78.

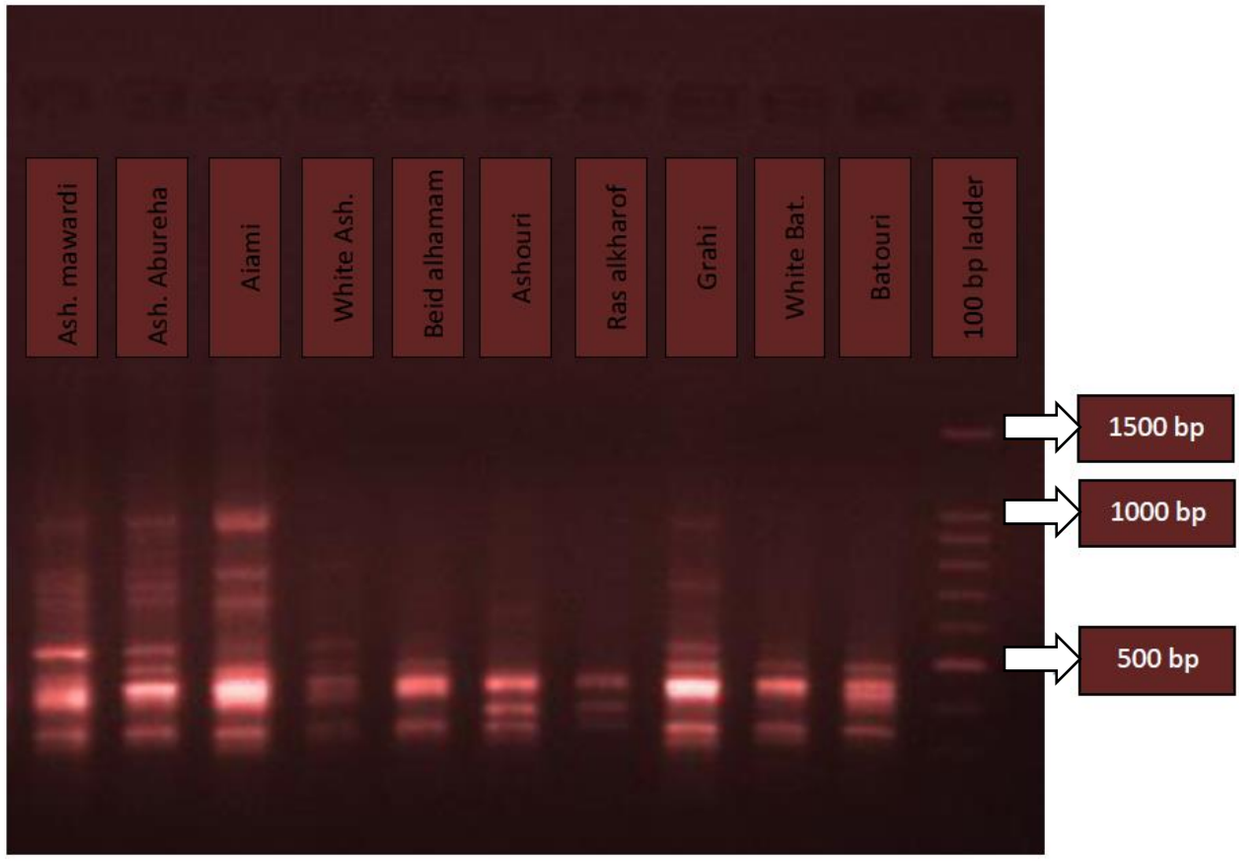


Figure 1 PCR amplified products using Primer K24B

Table 3 Genetic similarity amongst pistachio studied cultivars and genotypes

	Ash. mawardi	Ash. Abureha	Ajami	White Ash.	Beid alhamam	Ashouri	Ras alkharof	White Bat.	Batouri	Grahi
Ash. mawardi	1									
Ash. Abureha	0.73	1.00								
Ajami	0.55	0.56	1.00							
White Ash.	0.69	0.64	0.65	1.00						
Beid alhamam	0.40	0.47	0.43	0.53	1.00					
Ashouri	0.54	0.50	0.57	0.73	0.48	1.00				
Ras alkharof	0.42	0.37	0.45	0.56	0.53	0.61	1.00			
White Bat.	0.40	0.40	0.40	0.47	0.43	0.45	0.45	1.00		
Batouri	0.46	0.43	0.45	0.52	0.40	0.46	0.52	0.64	1.00	
Grahi	0.37	0.36	0.40	0.37	0.31	0.37	0.41	0.48	0.61	1.00

Cluster Analysis

Depending on the UPGMA algorithm and Dice equation, the cluster analysis divided the studied cultivars and genotypes into three main clusters. The first cluster comprised the white genotypes of the largest nut's size, and is divided into two main sub-clusters. The first sub-cluster encompassed white Batouri and comparative Batouri cultivar, while the second sub-cluster included Grahi cultivar. The second main cluster comprised Beid Al-Hamam and Ras Al-Kharof genotypes which are also analogous in many morphological characters and similarly have white nuts with genetic similarity 0.53, as seen in Figure 2. The third cluster was detached into 3 sub-clusters; the first sub-cluster contained Ajami cultivar, whereas all Ashouri genotypes were located in the second and third sub-clusters, white Ashouri and Ashouri cultivar were located together, and Ashouri abureha and Ashouri

mawardi were in another sub-cluster. Undeniably, the third cluster comprised all cultivars and genotypes of red hull nut (Ajami and all Ashouri genotypes) except the white Ashouri genotype due to its resemblance to Ashouri cultivar in most accredited parameters. Baghizadeh and Dehghan (2018) applied 15 ISSR markers on 20 pistachio samples to appraise genetic diversity. Their results improved that ISSR data clearly discriminated the cultivars in terms of their genetic characterization and divided the studied samples into four main clusters.

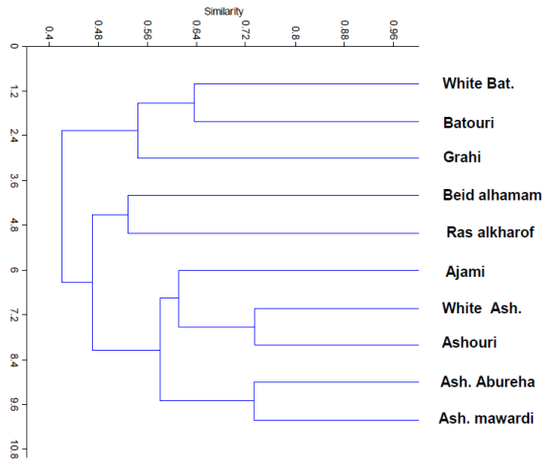


Figure 2 cluster analysis using UPGMA algorithm and Dice equation

Unique Bands

Excluding the primers A24A and A2, all other primers detected 33 positive, unique bands, as illustrated in Table 4. The highest number of unique bands was detected in the white Batouri genotype (9 positives). Followed by Ashouri Mawardi, 6 positives were detected by different primers. The genotypes Grahi and Beid alhamam were diagnosed by 5 positive unique bands. Expressly, in Grahi genotype, the primer ISS6 generated 3 unique bands (360, 678, and 796 bp). Both of the genotypes, white Ashouri and Ashouri Abu reha were not recognized by any unique bands. Five unique bands were scored by primer A6 with molecular weight (475 bp in Grahi, 335 bp in Ash. Mawardi, 300 bp in White Batouri, 215 bp Ajami, and 209 bp in Ras alkharof). Besides, five positive unique bands were recorded by primer ISS6 with molecular weight (921bp in White Batouri, 868 bp in Beid alhamam, and 796 – 678- 360 bp in Grahi).

Table 4 The over-all positive unique bands (bp) scored by ISSR primers in pistachio

Primer	Unique bands	Ash. Mw	Ash. Abur	Aja.	W. Ash.	Beid Alha.	Ash.	Ras Alkh.	W. Bat.	Bat.	Gra.
ISS2	2					569			716		
ISS3	4	1150 415				440	425				
ISS5	1					660					
ISS6	5					868			921		796 678 360
K25	3	861					424		1022		
A4	3	379		663					563		
A5	2			1029						1138	
A6	5	335		215				209	300		475
K11	1	558									
ISS7	1										513
K26	1								442		
K24A	-										
K24B	2								784 372		
UBC840	3					486		511	424		
A2	-										
Total	33	6	-	3	-	5	2	2	9	1	5

CONCLUSION

All the detected primers were effective in clarifying the genetic polymorphism and the unique bands. The cluster analysis classified all investigated cultivars and genotypes according to their resemblance. The current investigation persisted in the importance of molecular markers in identifying the genetic platform for each pistachio cultivar, mainly those more productive genotypes, to insight the knowledge of their genetic base and their relativeness in the aim to assess the genetic identification for all Syrian pistachio cultivars and genotypes.

References

- *Alhajjar, N.M., Hamed, F. & Muzher, B.M. (2017). Genetic similarity among pistachio (Pistacia vera L.) female genotypes and cultivars planted in Sweida province using SSR technique. Damascus Journal of Agricultural Science, 33(1), 239- 256.*
- *Alhajjar, N.M. & Muzher, B.M. (2017). Identification of male genotypes in Pistacia vera L. species using SSR markers. International Journal of Environment, 6(2), 1-12.*
- *Alhajjar, N.M., Muzher, B. M. & Hamed, F. (2016). Assessing genetic relationships between Pistacia vera L. Hybrids and their parents (P. vera × monoecious genotypes of Pistacia atlantica) using SSR markers. Jordan Journal of Agricultural Science, 12(1), 148- 157.*
- *Alhajjar, N. M., Muzher, B. M. & Hamed, F. (2015). The effect of pollen grains of Pistacia vera and Pistacia atlantica (unisexual and hermaphrodite) on quality parameters of Ashouri and Batouri pistachio cultivars. Jordan Journal of Agricultural Science, 11(1), 15- 25.*
- *Al-Saghir, M.G. & Porter, D.M. (2012). Taxonomic revision of the genus Pistacia L. (Anacardiaceae). American Journal of Plant Science. 3, 12- 32.*
- *Amirebrahimi, F.F., Meimand, M.M., Karimi, H.R., Malekzadeh, K. & Yajabadipour, A. (2017). Genetic diversity assessment of male and female pistachio genotypes based on ISSR markers. J Plant Mol Breed, 5(1), 31- 39.*
- *Baghizadeh, A. & Dehghan, E. (2018). Efficacy of SCoT and ISSR markers in assessment of genetic diversity in some Iranian pistachio (Pistacia vera L.) cultivars. PHJ., 1(1), 37- 43.*
- *Dice, L. R. (1945). Measures of the amount of ecological association between species. Ecology, 26(3), 297- 302.*
- *Ehsanpour, A.A., Tavassoli, M. & Arab, L. (2008). Sex determination of Pistaia vera L. using ISSR markers. Malays. Appl. Biol., 37(2), 25-28.*
- *Esfandiyari, B. & Davarynejad, G.H. (2011). Data to sex determination in Pistacia species using molecular markers. Euphytica. DOI 10.1007/s10681-011-0527-6.*
- *Fares, K., F. Guasmi, L. Touil, T. Triki & Ferchichi, A. (2009). Genetic diversity of pistachio tree using Inter-Simple Sequence Repeat markers ISSR supported by morphological and chemical markers. Biotechnology, 8(1), 24-34.*
- *Ghorbel, A., Ben Salem-Fnayou, A., Chatibi, A. & Twey, M. (1998). Genetic resources of Pistacia in Tunisia. In: Towards a comprehensive documentation and use of Pistacia genetic diversity in Central and West Asia, North Africa and Europ. Padulosi, S. and Hadj-Hassan, A. (eds.). IPGRI Report of the IPGRI Workshop, 14-17 December, (P: 62-71).*
- *Ghorbel, A. & Kchouk, M. L. (1996). Genetic resources of horticultural crop in Tunisia. Second Meeting of the WANA Working Group on Horticultural Crop. International Plant Genet.Resour. Inst., Aleppo, Syria.*
- *Kafkas, S. & Topaktas, M. (2003). Chromosome numbers of Four (Anacardiaceae) species. Journal of Horticultural Science Biotechnology., 78, 35–38.*
- *Kafkas, S., Ozkan, H. B., E., Acar, I., Atli, H.S. & Koyoncu, S. (2006). Detecting DNA polymorphism and genetic diversity in a wide germplasm: comparison of AFLP, ISSR, RAPD markers. American Society for Horticultural Science, 131, 522-529.*
- *Kamiab, F., Ebadi, A., Panahi, B. & Tajabadi, A. (2014). RAPD analysis for sex determination in Pistacia vera L. Journal of Nuts, 5(1), 51-55.*
- *Kebour, D., Boutekrabt, A. & Mefti, M. (2012). Using Inter Simple Sequence Repeat (ISSR) markers to study genetic polymorphism of pistachio (Pistacia vera L.) in Algeria, African Journal of Biotechnology, 11 (29), 7354- 7360.*
- *Mahmoodnia, M. & Malekzadeh, K. (2017). Genetic diversity assessment of male and female pistachio genotypes based on ISSR markers. J. Plant Mol. Breed, 5(1), 31-39.*
- *Noroozi, Sh., Baghizadeh, A. & Javaran, M.J. (2010). Study on genetic diversity of some Iranian pistachio (Pistacia vera L.) cultivars using RAPD, ISSR and SSR markers: Amplification study. Biotechnology, 4(3), 120-125.*
- *Noroozi, Sh., Amin, B. & Javaran, M. J. (2009). The genetic diversity of Iranian pistachio (Pistacia vera L.) cultivars revealed by ISSR markers. BioDiCon, 2(2), 50- 56.*
- *Porebski, S., Bailey, G.L. & Baum, B.R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components.Plant Molecular Biology Reporter, 15 (1): 8-15.*
- *Tagizad, A., Ahmadi, J., Haddad, R. & Zarrabi, M. (2010). A comparative analysis of ISSR and RAPD markers for studying genetic diversity in Iranian pistachio cultivars. Iranian Journal of Genetics and Plant Breeding, 1(1), 6-16.*
- *Turhan-Serttas, P. & Ozcan, T. (2018). Intraspecific variations studied by ISSR and IRAP markers in Mastic tree (Pistacia lentisecus L.) from Turkey, Trakya University. Journal of Natural Science, 19(2), 147-157.*

Risk Assessment Model for Cloud-Connected Networks with Case Study on an Academic Institution

نموذج تقييم المخاطر لشبكات الحاسوب الموصولة مع الخدمات السحابية للمؤسسات الاكاديمية

Islam Younis Amro
Associate Professor/ Al-Quds Open University / Palestine
iamro@qou.edu

اسلام يونس عمرو
أستاذ مشارك / جامعة القدس المفتوحة / فلسطين

Received: 07/09/2021, Accepted: 27/09/2021
DOI: https://doi.org/10.33977/2106-000-005-004
https://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 2021/09/07، تاريخ القبول: 2021/09/27
E-ISSN: 2521-411X
P-ISSN: 2520-7431

Abstract

The reliance on cloud services has increased recently, resulting in an abundance of networks connected to these services partially or fully. However, several risks emerge from this action that imposes new challenges. Organizations often maintain a range of services managed in its own local or expanded networks as well as services that could exist on the cloud services sites partially or totally. Organizations have to deal with two types of risks: The first relates to the internal information systems risk of the organization, and the second relates to risks that come with working with cloud services providers. Furthermore, organizations lack benchmarking and references on assessing information systems risks. Most organizations work with vulnerability management concepts rather than risk assessment and mitigation. In this paper, we reformulate strategic e-services in an educational institution as it works between local networks and cloud services at the same time to study the risks associated with them in a hybrid manner. These services are distributed over local network nodes and relevant cloud components. The local network components and nodes; represent hosts with known vulnerability values generated from commercial tools. These vulnerabilities are gathered into vectors with expected impacts and estimate assets value related to these services. Probabilities or risks are identified accordingly. The other component of the research considers analyzing the risk of the cloud services with the computational approach, but it deals with cloud standard components such as data management policies, internal cloud provider management, and internet security. Vulnerability in cloud providers is identified as the compromise of these components and their impact on business continuity. Using vulnerability concepts for both local network and cloud, we introduce a risk probability model for educational organization (e.g.: QOU) services where risks are estimated over Borda Count generated weights for both local network and cloud. Moreover, the overall risk is estimated independently for each component; local network and two clouds. The final step is to investigate the overall risk for the organization. It will be done by prioritizing these risks mutually and analyzing the value of each risk in terms of other risks. For this purpose, we use the analytic hierarchy process (AHP).

***Keywords**: Cloud Computing Risk Assessment, Vulnerability Management, Business Continuity, Borda Count., Analytic hierarchy process (AHP).*

الملخص

يزداد الاعتماد مؤخرا على الخدمات (السحابية) بحيث يتصل كثيرا من الشبكات بهذه الخدمات بشكل جزئي أو كلي. يصاحب هذه التغيرات جملة جديدة من المخاطر والتحديات تضاف إلى المخاطر القائمة في الشبكة. ففي كثير من الأحيان تحتفظ المنظمة بمجموعة من الخدمات التي تدار في الشبكات المحلية أو الموسعة الخاصة بها إلى جانب الخدمات التي يمكن أن تتواجد كليا أو جزئيا على مواقع الخدمات (السحابية). يضاف إلى هذه الإشكالية الضعف العام في المنظومات الإدارية للمخاطر في كثير من المنظمات بحيث تُغيب وثائق أساسية في التعامل مع المخاطر، مثل: استراتيجية نظم المعلومات، واستراتيجية أمن المعلومات، والأوراق المرجعية لمحددات أمن المعلومات، والمخاطر ذات العلاقة في المنظمات. أمّا ما يتم العمل عليه في المنظمات وبشكل يومي هو إدارة نقاط الضعف بحيث يتم التعامل مع نقاط الضعف الفنية في الشبكات المحلية والموسعة من خلال التقييم الدائم للتغيرات الدائمة على تقنيات البنى التحتية والتطبيقات، واقتراح الحلول الأمنية على مستوى العقدة في الشبكة، وحلها وبشكل آلي في بعض الأحيان وذلك بالاعتماد على تقنيات صناعية معدة لهذا الغرض، دون المرور على مفهوم المخاطر والتعامل معه. في هذا البحث نعمل على إعادة صياغة الخدمات الالكترونية الاستراتيجية في منظمة تعليمية والتي تعمل بشكل (هجين) بين الشبكات المحلية، والخدمات (السحابية) في الوقت نفسه؛ لدراسة المخاطر المرتبطة بها. وتم تقسيم الخدمات الاستراتيجية إلى مجموعات عمل تضم العقد المشاركة في بناء هذه الخدمة سواء كانت في الشبكة المحلية أو في المواقع (السحابية)، ثم تمّ احتساب الضعف لكل عقدة من العقد المشكلة للخدمة باستخدام أدوات تجارية متخصصة بهذا الشأن، وتم رسم مسار للعقد التي تشكل هذه الخدمة في الشبكة المحلية والتعبير عنه بمتجه يعرف بمتجه نقاط الضعف لخدمة استراتيجية معينة، وتم تقدير قيمة الأصول واحتمالية وقوع الخطأ، ودرجة تأثير الخطأ حين حدوثه. والخطوة التالية كانت اسناد أوزان هذه الخاطر، واحتساب قيمتها لكل خدمة استراتيجية، ومن ثم احتساب المخاطر الكلية للخدمات الاستراتيجية التي تعمل في المنظمة على شبكاتها المحلية. ومن ثم تمّ احتساب المخاطر المصاحبة للعمل مع الخدمات (السحابية) وهي ذات نوع وتأثير مختلف من حيث تسريب المعلومات أو فقدانها أو تعرضها للسرقة أو أي خلل يتسبب به مزود

الخدمة (السحابية). وهنا أودّ الإشارة إلى أن المخاطر المستخدمة في هذه الورقة هي ضمن المخاطر المعيارية، والمنصوص عليها في الأدبيات ذات العلاقة. وتمّ اتباع المنهجية ذاتها في احتساب مخاطر الخدمات (السحابية) ومن ثمَّ تمّ احتساب المخاطر الكلية حسب شدة الخطورة؛ وذلك حسب خوارزمية عملية التحليل الهرمي. بحيث تم الخروج برقم موحد لمخاطر المنظمة اعتمادا على الخوارزمية سالفة الذكر.وتمّ الاستناد إلى بيئة جامعة القدس المفتوحة في إعداد بيانات هذا البحث.

الكلمات المفتاحية: الشبكات السحابية، تقييم المخاطر، إدارة نقاط الضعف، استمرارية الاعمال، عداد بوردا، عملية التحليل الهرمي.

INTRODUCTION

General Prospect on Information Systems Risk Assessment

In the modern age of the fourth information systems revolution, an extensive dependency on information systems has become noticeable. One of the key issues related to the presence of information systems is the need for information systems security risk assessment. The key problem affecting information systems risk assessment arises from the lack of organizational benchmarks and references to assess an overall prospect for information systems risk. This leads to more contingency approaches in managing vulnerabilities-they can be assessed more easily than systems risk-as a substitute for system risk but not a replacement. Information system risk has a very broad concept that alludes to generic business risk and forms an essential compound of business risk matrices. This explains how organizations usually have very good knowledge, skills, and plans to manage vulnerability on an information systems level. However, they still have a less mature explanation and methodologies on transforming vulnerabilities management into information systems risk assessment and part of business risk over an organization. Information systems risk is concerned with issues of vulnerabilities but exceeds those concepts to risk identification, analysis, prioritization in terms of impact, probabilities, dependencies, time, and other avalanches. The outcome of this process is subjected to risk mitigation plans and so on (Metzenger et al., 2007). Based on ISO/IEX IS13335X and ISO 27001 families, some are actively modified and updated while some are withdrawn since a key common concept of information systems seems timeless. These concepts are assets, threats, and vulnerabilities. An asset is defined as anything tangible or intangible within an organization and has value. Each asset presence, absence, or malfunction has a certain impact on an organization which is very important to understand when risk is being assessed. The process of tracking the impact is defined as the impact assessment. Another key concept is threat. Threats are a set of actions and/or events that may cause harm. The last concept is vulnerability, which refers to the weakness in protecting this asset. The combination of these three elements forms the foundation of information security risk management. Risk management entails two main phases; the first is to identify the risk, and the second is to manage it. Risk identification entails the process of assessing assets and their values, their impact on the system cycle, and their vulnerabilities. Managing risk is related to defining and implementing mitigation plans that would avoid risks or define operational alternatives then adopting them if the risk is being actualized. The International Organization for Standardization ISO developed a wide set of procedures and concepts in this regard under the ISO 2700(1:5) family (ISO 2018). However, the problem arises from several standards, approaches, concepts, and even understanding of the risk assessment, as in Lonita et al. (2014). Moreover, the strengths and weaknesses of each approach are difficult to track. From the researcher's point of view, the key problem of all approaches comes from answering two questions: 1- How to integrate the information security risks as a business risk for non-technical people and 2- How to calculate the framework parameters regardless of the type of the framework. The inner details of each approach and standards are different; therefore, the comparison between the approaches can be fascinating. We should take into account the purpose of information risk assessment as a part of the organizational risk assessment. Regarding the first problem on the integration of technical terms into business terms, the techniques of calculating the risk assessment parameters may vary from simple questionnaires to Heuristic calculation methodologies, as in Andersen (2014). These parameters include threats, impacts, and even vulnerabilities. The need for Heuristic methods arises from the lack of

benchmarks, references, and clear organizational assessment. The importance of Andersen's work arises from combining business and technical risks on one computational model, which reflects the tight relationship between technical risks and business risks on the computational level. This has presented sufficient information to non-technical people, according to Andersen. Business risk assessment for cloud computing was addressed (Bernardo, 2013), where a computational model was developed to assess information systems risks over the cloud for non-technical people. Khidzir (2010) pointed out that the investigation worked with the outsourced services and risks related to them, namely, risk identification, analysis, treatment plans, implementations, monitoring, and control. Moreover, regarding technical issues, the research suggested business Service Levels Agreements (SLA) rather than infrastructures problems. Extensive work on parameters calculations found in reference (Amin et al., 2013) considers the impact of organizational structure combined with information security tools and technology-based security systems in fault-tolerant control on risk calculations. The analysis considers the service-oriented architecture (SOA) as a reference. Amin (2013) also suggested that the risk assessment might also depend on the technical architecture and showed good incorporation between business and technical terms. Maule et al. (2009) presented in a study a specific risk model for SOA. Furthermore, the research found that this model is very similar to the traditional risk model based on risk probability and asset value. From our perspective, the real value of this research is that it focuses on the business components of SOA. Xiaojun et al. (2011) introduced another risk assessment of a Web service case based on SOA of multiple applications. Asosheh et al. (2009) found a very clear incorporation between technical and business terms. This research represents a new quantitative method for assessing the overall information security risk in a real business environment. The new method is based on Microsoft and Callio Secura methods, which are common and practical methods. The advantage of this approach is that the organization can determine its business risks and return on security investments. Kassou (2012) introduced a maturity model of SOA risk assessment. In contrast, this research introduces the principles of a new tool that supports the organization's SOA security maturity assessment called SOASMM (SOA Security Maturity Model). This model is defined by combining information security best practice methods into a service-oriented architecture paradigm using controversial methods and mapping models. Saleem et al. (2015) considered integration between business risk analysis and IT Security Risk. He showed the classification between services according to strategic importance and considered these issues accordingly in assessing the organization's risk. In reference to the second point: How to calculate the framework parameters regardless of the type of the framework, we can conclude the following. All of the preceding techniques, such as ISO/IEC 13335-2, ISO/IEC IS 17799, and ISO 270001, would still require a method for quantitative risk assessment, estimating the values of assessing values, risk impact, with a series of questionnaires included in security plans for organizations. Unfortunately, a wide range of organizations lacks detailed information security strategies and sometimes mitigate on purpose. These strategies are usually acquired from broader strategies such as information systems strategy, which in its turn reflects the broader organizational strategy. Butting all of these cascaded strategic documents is exploited to calculate the overall organizational risks, technical and non-technical. A wide range of methodologies is used to project the organizational strategies. Most of these methods are computational, but some are empirical. For an organization that has not developed these concepts maturely, the systems' risk is minimized into technical vulnerability management. These vulnerabilities are quantified and obtained from specific systems that analyze the security status of these assets. Other problems have appeared, such as specifying the asset's value, risk probability, and risk impact. Then sorting out these values and how these values are going to be expressed in business terms. Furthermore, several approaches have been developed addressing the exploitation of vulnerability value, asset value, risk impact, and the probability of the occurrence of the risk. To translate these calculations into business terms, Andersen (2010) of IBM and Asosheh et al. (2009) used probabilistic approaches. These two interrelated works subjected the parameters to a probabilistic model and projected the overall risk

within an organization based on technical information. The researcher used a multistage approach in analyzing the systems and then expected the overall risk based on a specific estimation model. The weights produced from an adaptive hierarchical process were optimized using a heuristic neural network method made by Xi et al. (2010). This issue entailed substantial calculations for the weights of risk assets. However, we do not believe risk assessment should go through due to the dynamic nature of risks. Xi et al. (2010) had the same concerns with large calculations as in Xiao et al. (2010). Another approach exploited fuzzy logic and inference systems to identify the risk parameters and protect them from given vulnerability systems, as in Jinxing et al. (2020) study. Another exploitation of fuzzy logic and Bayesian networks for estimating the overall risk was based on known vulnerability values found in the study of Zang et al. (2018). Relatively simpler approaches were found in Riaz et al. (2019) study; it exploited simpler fishbone methods in investigating business risks on software development. The weights produced from an adaptive hierarchical process were optimized using a heuristic neural network method made by Xi et al. (2010). This issue entailed substantial calculations for the weights of risk assets; still, we do not believe risk assessment should go through due to the dynamic nature of risks. Xi et al. (2010) have the same concerns with large calculations as in Xiao et al. (2010) study. Furthermore, other approaches exploited fuzzy logic and inference system to identify the risk parameters and protect them from given vulnerability systems, as in Jinxing et al.'s (2020) work. Another exploitation of fuzzy logic and Bayesian networks for estimating the overall risk based on known vulnerability values was found in the study of Zang et al. (2018). Relatively simpler approaches were found in Riaz et al.'s (2019) study since it exploited simpler fishbone methods in investigating business risks on software development. Another approach based on calculating Risk and Borda Calculations was exhibited in Amro's (2015) study.

From the previous literature review, we can conclude that some issues need to be dealt with. First, scientists have to do extensive work relating to business information systems risk methodologically, where technical terms do not consume business terms. In addition, there are several models identified to quantify business risk related to system architecture and software services type. Furthermore, several numerical methods vary in complication to estimate the business risk value based on given technical information. Regardless of any organization's situation, there are three documents -Business Strategy, IT Strategy, and Security Strategy- which should be referenced to build a proper risk assessment and containment plan, as in known frameworks or Information Security Management System (ISMS). These documents are essential to assess the risks related to business assets, Assets Values, and related impacts on business. Unfortunately, many businesses lack either an IT strategy or a security strategy and sometimes both. We still need organizational references to figure out how much our assets are worth. Even though technical knowledge about vulnerabilities is available, the risk model's calculation must be quantified on business. Unfortunately, many organizations do not have an IT strategy or security strategy, or both. We still need organizational references to assess the values of our assets.

Information Security Risk Identification for Cloud Services

The core issues of IT sourcing services were addressed by Moona et al. (2018). The core of the information security risk for the outsourced managed services running on clouds is related to the nature of the service provider company. Theoretically, the information of the served company will be processed by the serving company. There is a potential of exposure of sensitive information for the served company by the serving company. Unauthorized access to sensitive information and leaked information to a third party can be possible. The information security risk is divided into subjective risk and objective risk. The subjective occurs when the contractor takes advantage of services running on his cloud to achieve certain benefits and uses the client's data for other risks. However, the objective risk occurs under the condition when someone leak the information and the contractor lacks experience and level, even though he has realized the importance of security and taken certain measures. This can be addressed using a

powerful information security system. Key security risks form for managed services contains the following issues:

Data Protection Protocol

The clients in these cases should establish a very clear data protocol where clients define all the types of the implemented process. In transfer, processing, transmitting, storing, etc., this protocol has a contractual value and should be very controlling to the service provider. This is an essential step to reduce the information security risks. Moreover, this would include defining very clear security technologies, communication technologies, how to move and store data, the kind of protocols, levels of security, conditions on future subcontracts, and the cause of breaking the contract.

Network Security

Network security contains the subsequent contents: the hardware and software of the network system. The data within the system should be protected against damage, modification, and leakage for infrequent or vicious reasons: The system can normally operate constantly and reliably, and the network service will not be interrupted. Network security means the data security on the network in essence. Hackers aim to illegally obtain, peep, modify or damage sensitive information by using various technologies. The contractor should utilize the foremost advanced technology to extend firewall and antivirus systems in the network, such as invasion detection and vulnerability scanning to the network as well as set storage limits to guarantee the safety of the network, host machine system, and application system. Moreover, contractors should make a powerful disaster recovery plan and data backup to guarantee the client's information security.

Internal Management

An early survey on information security affairs by Gartner-collective information technology marketing research company-found that over 70% of faults are caused within corporate. The survey and research made in two departments by Abdulwahes et al. (2014) verified that almost all affairs related to security occur within the organization. These security risks/violations include using the organization's resources for other purposes, such as sharing the password with colleagues and external persons and plugging incorrect or forged information in the system and computer procedure. Moreover, the organization should implement information security education and career training for its staff, improving their knowledge of the significance of security knowledge and ensuring the client's info security. Second, each confidential staff passes security authentication, signs the safety and confidentiality agreement, and understands concrete security measures. Third, the organization should perfect the principles and regulations and ensure that the division of labor is explicit and the responsibilities are clear yet strictly controlling the confidential scope. Fourth, the organization should perfect the network supervision and management mechanism and forestall any security accidents caused by internal employees, particularly confidential staff and external interference, to maintain the client's information security. Fifth, organizations should provide clear administrative management measures such as door access, internal and external monitoring systems, and server protection.

Regulations

The information security protection does not depend on the contractor alone, but it requires the government's provision of a decent information security environment such as legal support towards dispute in outsourcing managed services and explicit specification for the defense of property. Furthermore, enhancing the public knowledge awareness of security and perfecting belongings protection and interrelated law. China's legislation of information security protection is comparatively backward; there was no law protecting the individual and organizations' information security until 2010. In this year, DOC, Industrialization and Informationization Department issued several regulations about Information Protection of Outsourcing Managed Service Contracted by domestic companies, to complete relative law as soon as possible. Moreover, the protection executive strength for holding is weak. Chinese people have weak awareness of private information protection and belongings protection because China lacks laws within the field for an extended time. Although a

series of rules and regulations have been made in recent years, changing people's concepts requires a process, which also causes information security risk towards clients. Therefore, education, publicity, and execution efforts should be enhanced. Third, the industry entry threshold should be set positively to guide the contracting enterprises to attain ISO27001 Information Security Management System (ISMS) authentication. The full information security condition of the corporate should pass the assessment of some institutions. The safety and reputation of the contractor company should be assessed to confirm the grade. Targeted protection measures should be applied maximally to reduce the data security risk for the client.

Supervision Mechanism

Enhancing supervision and management is an essential means for effectively finishing the enterprise's execution. During the execution of the contract, the contractor should establish a regularly formal communication system, find information security risk in time, and establish corresponding preventive measures to reduce information security risk and guarantee the client's information security via control. The client must participate in planning and processing and consider his role as a supervisor. The corporate might form the supervision and management team internally or consider hiring a third-party supervising institution to search out the matter in time, take measures and reduce risk. The corporate should realize visualization of its internal operation and might respond quickly when the client monitors the qualitative process, and thus the objectivity of assessment will further improve.

Determination of Danger Elements

Based on the International Information Security Management Practice Norms ISO/IEC 17799 and Information Security Technology and Knowledge Security Risk Assessment Standards GB/T20984-2007\. Five risks exist within the IT Outsourcing Managed Service Security, which concluded betting on three fundamental elements: assets, threat, and vulnerability. By taking the knowledge safety features of IT Outsourcing Managed Service into consideration, the concrete content of every risk is demonstrated in Table 1.

Table 1 Risk Concerns of IT Outsourcing Managed Cloud Service

The data protection agreement	methods, scope, degree, intellectual property ownership, liability for breach of contract, safety measures, etc., of data protection
Internal management	System construction, educational training, information Access control and maintenance, prevention of the malicious staff to tamper with the information emergency measures.
Internet security	Including data protection of the internet, the host system and application system, and antivirus measures
Supervising Mechanism	Communicate and exchange ideas, clients participate in the supervision, establish supervision institution, and visualize the internal operation.
Law and policy	The construction of laws and regulations, intellectual property protection, set industry entry threshold, and evaluate the information security Protection level of the contractors.

This paper addresses building a risk assessment model for a network that has a series running locally and other services and services components running over clouds. The Local Network has vulnerability values only, without referencing documents essential for calculating risk values and impacts. The following section explains our problem, relates it to the literature review and discusses the research problem and methodology. After that, we discuss the proposed Network Service-Based Risk Assessment Model, which combines the local area network and cloud service. It explains the roadmap for building the model components through several steps. First, we build the testing environment, which is the network we based our simulation on, then we work on the Vulnerability Calculation Model for local networks and clouds. After that, we explain our method- Risk Probability and Risk Impact Estimation- then we work on the Determination of the Risk Rank Reference. Later, we determine the risk rank and then calculate the Risk Weight Estimation, which will be used in the Overall Risk Calculation. Finally, we write a final flow chart summary for all the steps on how to exploit this approach for similar networks. In section 4, we implement our model into a testing environment as a case study, go over the steps in section 3 and generate the risk of an educational organization.

Then we conclude our research with a finalization of the results.

RESEARCH PROBLEM AND SOLUTION METHODOLOGY

Three documents: Business Strategy, IT Strategy, and Security Strategy should be referenced to build a risk assessment and containment plan, as in known frameworks or Information Security Management System (ISMS). These documents are essential to assess the values related to risk: the Business Assists, Asset Values, and the related impacts on business. Unfortunately, many businesses lack either an IT strategy or a security strategy, or both. Although technical knowledge about vulnerabilities is available, we still need organizational references to figure out how much our assets are worth. In addition, the risk model's calculation must be quantified.

1. Without an IT or security strategy, how can you construct a network services risk assessment model?
2. How to put together a composition that provides strategic services by combining business strategies, information system components, Cloud services, and infrastructure components.
3. Introduce a more user-friendly adaptive approach to calculating risk doe both locally hosted and managed services.
4. How to build a risk assessment model that is aware of cloud-based services.
5. In light of the preceding circumstances, how can risk be assessed for both cloud and network risks?

We adapted these concepts in expressing business strategies in terms of information systems services and infrastructure services, which is not an SOA. Instead, we used a combination of infrastructure components and information systems resources to measure its vulnerability in expressing them as services and then reflecting these services on business strategies. In addition, we took into consideration the risk problems that appear in services running over clouds. This research extends the works conducted by Amro (2015) to include services running on cloud connected to the network topology. This is conducted by computing the risk values for the local network then the risk for each cloud. A final resultant risk is obtained from the three elements local network, Cloud A, Cloud B, using the AHP method explained in Moona et al. (2018). Several multi-criteria of decision-making methods can be used in resultant risk assessment, as in Maček et al. (2020). However, we used AHP for its relevant simplicity.

Network Service-Based Risk Assessment Model Testing environment

Suppose we have a computer network for an organization, as represented in Figure 1. This figure suggests a topology-based representation for the network, with one broadcasting domain around its central switch and protected behind a firewall. The network can be accessed through two router ports; internal and external. These routers represent a separation point between the routing and broadcasting domains. The organization's network is connected to two clouds, cloud A and cloud B.

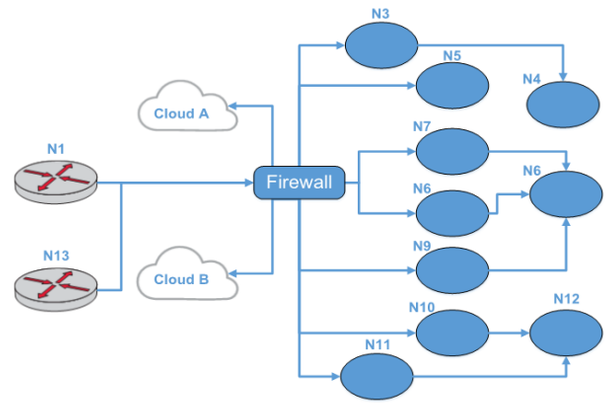


Figure 1 Computer network for Testing Organization

The network has several hosted services running on or through the Nodes (N); these are the insourced network services. Each node represents the hosting machine(s) for the provided services over the network. Each node N is associated with a vulnerability vector V which is calculated using standard tools and quantized from 0 to 5. In this research, Qualys Vulnerability Assessment Tool is used. The vulnerability number for each node represents the average of vulnerabilities for this host. The problem with this number is that it comes with a high-dimensional vector that varies in its norm after each scan. The rating of each vulnerability is given according to Qualys Standard, which is beyond the scope of this paper. However, for the nodes 1 - 12 in Figure 1, the vulnerabilities were 2, 3, 4, 1, 3, 4, 3, 2, 4, 3, 4, and 3, respectively. On the other hand, the risks

incorporated with clouds A and B are different in nature since they are outsourced services. Both circumstances and conditions are different. The sources of risk in IT outsourced managed services are listed in Table 1. The resultant risk for the organization is a summation of both internal and external risks. Suppose that we have the following strategic elements, and we would like to investigate and assess their risk. These strategic services are running on the mentioned network in Figure 1. Table 2 shows these services. Table 2 maps service elements with corresponding nodes, i.e., service path scenarios based on the network predefined access plan. In Addition, for elements, we clarify that CA stands for Cloud A and CB for Cloud B.

Table 2 Service Path Access Scenarios

Element	Service Elements	Related Nodes
1	E-learning	N1,N12,N2,N3,N4,CA
2	MAIL	N1,N12,N2,N5,CA
3	Registration and Student portal	N1,N2,N12N7,N6,CA
4	HR portal	N1,N12,N2,N8,N6,CA
5	Financial system	N1,N2,N9,N6,CB
6	Journals portal	N1,N2,N10,N12,CB
7	Library portal	N1,N12N2,N11,N12,CB
8	Infrastructure	All Nodes

We need to incorporate Tables 2 and 3 by mapping service elements into a higher level for business-related purposes since risks are addressed on a higher level of the servers and other connectivity issues. Table 3 maps the major risk items that we have identified in this study S1 to S6 with the service elements. It is worth mentioning that this issue is network-scenario specific, and it might vary from one network to another.

Table 3 Service Elements Incorporation with Risk Element

Risk Item	Service Elements
Student electronic Services (S1)	Mail, e-Learning, registration and student portal, Library portal
Academic Systems (S2)	Mail, e-learning, registration and student portal, Library portal, Journal System
Human Resource Systems(S3)	Mail, HR Portal.
Financial Systems (S4)	Financial system, HR Portal
Research Systems(S5)	Library portal, Journal System
Infrastructure Components(S6)	All Service Elements in table 1.

Vulnerability Estimation Model

Vulnerability Assessment software works on the network node level, which does not express the business risk level. Figure 1 shows If N is a node in a network configuration. If we rewrite services running on the network nodes as shown in Table 2 in terms of network nodes in Figure 1, the services are classified into service elements E and are expressed in Table 2. The vulnerability for the network node N, expressed as, is the weighted average of all vulnerabilities of node N. Accordingly, each node is expressed by the vulnerability value and expresses the number of nodes' participation in the constitution of service element E expressed in Table 2. The resultant value for the vulnerability service element E expressed as and calculated by taking the maximum vulnerability value obtained from the above process for the nodes N1 to Ni constituting the element E, formally can be expressed as:

$$V_E = Max\,(V_{N_1}, V_{N_2}, ..., V_{N_i}) \quad (1)$$

The use of the maximum in Equation 1 is justified by the need to obtain the extreme value for the risk. Other approaches may use weighted averages, but we do not prefer to use them since they might only drop the vulnerability value for calculation. The next step is to incorporate node risks with risk elements that are forming the services to obtain the service vulnerability. The element vulnerability is mapped to the total risk items vulnerability Vs using the same logic in building Equation 1. Formally, is written as:

$$V_S = Max\,(V_{E_1}, V_{E_2}, ..., V_{E_j}) \quad (2)$$

where *j* represents the service element of component *E*, which forms risk item S. Equations 1 and 2 make it possible to write vulnerabilities on an organizational level in our work. The values calculated for V_S were 3, 4, 4, 4, 3, and 4, respectively.

The Estimation of Risk Probability and Risk Impact

We suggest that the risk probability P and risk impact I are ranked in 5 levels: very low, low, medium, high, very high, which express the frequency of vulnerabilities encountered and the risk probability. The value for the service reflects

the impact of risk. Probability and Impact are then expressed in 2D matrices exploited in the retrieval of the quantified value of P and I value. The risk probability P is then quantified by threats encountered for T times. In addition, it is expressed in Equations 3.

$$P = f_1(V, T) \tag{3}$$

$$T = (t_1, t_2, \dots\dots, t_i, \dots t_m) \quad , \quad 1 \leq i \leq m$$

$$f_1 = \alpha t + \beta v_S$$

$$\alpha = \begin{cases} 2, t \leq 3 \\ 3, 3 < t < 5 \\ 4, t = 5 \end{cases}$$

$$\beta = \begin{cases} 1, v \leq 3 \\ 2, 3 < v < 5 \\ 3, v = 5 \end{cases} \tag{4}$$

Alpha (α) and beta (β) are important to quantify P over the interval assumed. We selected the values of α and β so the higher the vulnerability, the higher the values for *P*. The impact *I* expresses the impact of the risk in accordance of asset value, these terms are expressed in Equations 5 and Equations 6:

$$I = f_2(V, A) \tag{5}$$

$$f_2 = \varphi a + \phi v$$

$$\phi = \begin{cases} 1, a \leq 2 \\ 2.5, 2 < a < 5 \\ 3, a = 5 \end{cases}$$

$$\varphi = \begin{cases} 2, v \leq 2 \\ 3, 2 < v < 5 \\ 4, v = 5 \end{cases} \tag{6}$$

$$V = (v_{S1}, v_{S2}, \dots\dots, vj, \dots . v_m), 1 \leq j \leq n$$

Estimation of the Reference of the Risk Rank

Table 4 below expresses the risk quantification by combining numerical and description levels; the first column presents the risk probability level. The impact has several levels and may vary from very low (-L) to medium (M) for the first row and from medium (M) to very high (+H) in the fifth row. Table 4 demonstrates a fine resolution between risk probability levels and risk impact levels.

Table 4 Relationship Between Risk Probability and Risk Impact Levels

Risk probability	**Risk Impact levels**				
levels	1	2	3	4	5
1	0.5 –L	1 -L	1.5 L	2.5 M	3 M
2	1 –L	1.5 -L	2 -L	2.5 M	3.5 H
3	1.5 L	1.5	3 M	3 M	4 H
4	2.5 M	3 M	3 M	3.5 H	4.5 +H
5	3 M	3.5 H	4 H	4.5 + H	5 + H

Risk Weight Estimation

In order to translate values from qualitative to quantitative, we need to define and determine risk weights; we exploited Borda count to achieve that. If total risk factors set of *N*, and *i* is a specific risk of set *N* with a criterion of k, then the value for risk in *N* can be expressed as:

$$b_i = \sum_{k=1}^{n} (\mathrm{N} - r_{ik}) \tag{7}$$

With total risk value expressed as:

$$B = \sum_{i=1}^{N} b_i \tag{8}$$

The weight for given risk RW_i expressed as:

$$RW_i = b_i / B \tag{9}$$

Overall Risk Calculation

Upon completion of the resultant risk-judging matrix, the overall security risk rank is expressed in equation 10, as:

$$RRT = \sum_{i=1}^{k} (\mathrm{RR}_i \, \mathrm{x} \mathrm{RW}_i) \tag{10}$$

CASE Implantation

The implementation goes through the steps as seen in Figure 2.

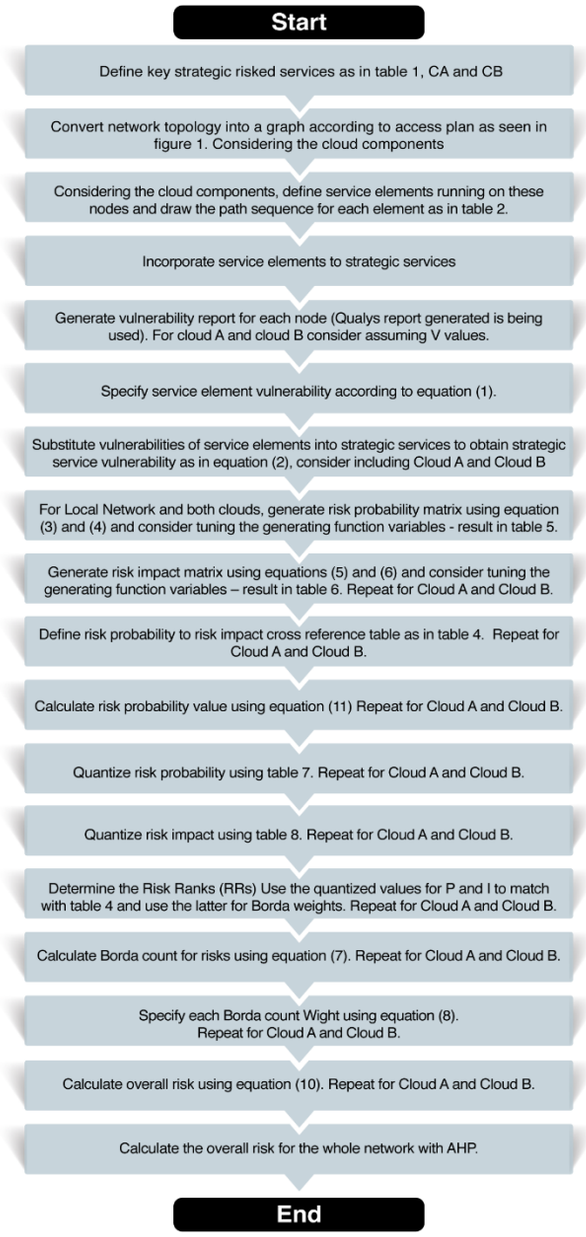


Figure 2 Experiment Case Implementation

As seen in Figure 2, we have implemented the previously mentioned steps to construct the general risk matrix seen in Table 9 for the local network, Table 10 for cloud A, and Table 11 for Cloud B. The steps from 1 to 6 have been previously implanted and explained. The resulting risk for six strategic services is 3, 4, 4, 4, 3, and 4, respectively. Then we implement step 7 to generate the risk probability and step 9 to generate the risk impact matrix. The matrices are shown in Tables 5 and 6, respectively. For Table 5 representing P, we assume the T values to be 5, 2, 2, 1, 3, and 4, respectively. For Table 6 representing I, we assume A to be 3, 3, 5, 2, 2, and 5, respectively. We assume the resulting risk for 6 strategic services running on a local network to be 4, 3, 4, 4, 4, and 4, respectively. For Cloud A V values are 3, 3, 2, 1, 1 and for cloud B 2, 2, 3, 1, 1. Note that the value for both clouds are obtained from Table 3. Then we implement step 7 to generate the risk probability and step 9 to generate the risk impact matrix. The matrices are shown in Tables 5 and 6, respectively. For Table 5 representing P; we assume the T values for the local network to be 5, 2, 2, 1, 4, and 4. And for the T value for Cloud A 2, 2, 3, 1, 3 and for Cloud B 2, 1, 1, 2, 1 For Table 6 represents I for services running on the local network; we assume A to be 3, 3, 5, 2, 2, and 5. And for cloud A 1, 3, 2, 1, 2 and for cloud B 1, 2, 2, 1, 3.

Table 5 Risk Probability Matrix

$P = f1(V,T)$	V				
	1	2	3	4	5
T	1	3	4	5	10	12
	2	5	6	7	12	14
	3	7	8	9	14	16
	4	13	14	15	20	22
	5	16	17	18	23	25

Table 6 Risk Impact Matrix

$I = f2(V,A)$	V				
	1	2	3	4	5
A	1	3	5	10	13	16
	2	4	6	11	14	17
	3	9.5	11.5	16.5	19.5	22.5
	4	12	14	19	22	25
	5	14.5	16.5	21.5	24.5	27.5

Using table 5; the risk probability for given values for V and T were 23, 7, 12, 20, and 20, and for cloud A, the risk probability is Cloud A 7, 7, 8, 3, 7 Cloud B6, 4, 7, 5, 3. The impact of these vulnerabilities were 19.5, 16.5, 24.5, 14, 11, and 27.5. I for cloud A was 3, 16.5, 6, 3, 4. I for cloud B 5, 6, 11, 3, 3. For step 10 we need to specify the risk probability value; this was achieved in equation 11:

$$r = \frac{Risk \quad Probability}{TotalRisk} \tag{11}$$

Total risks are 25 from Table 5, and thus the value of *r* becomes 23/25 and so on. These values were the local network 0.92, 0.28, 0.48, 0.8, 0.36, and 0.84. The total risk for cloud A 0.28, 0.28, 0.32, 0.12, 0.25 and for cloud B 0.24, 0.16, 0.28, 0.2, 0.12. In steps 12 and 13, we quantize *R* values, and *I* values using Tables 7 and 8. The quantization

in both tables is done by finding the interval *P* and *I*, the quantization values for *P*are 5, 2, 3, 4, 2, and 4. For *I*, the quantization values are 4, 4, 5, 4, 3, and 5

Table 7 Risk Probability Quantization

Probability P	1—5	6—11	12—16	17—21	22—25
P Level	**1**	2	3	4	5

Table 8 Risk Impact Level Quantization

Impact I	**1-5.5**	**6—11**	**12—15.5**	**16—22.5**	**23—27.5**
Impact level	**1**	**2**	**3**	**4**	**5**

In step 14, we use the quantized values of *P* and *I* to refine the risk rank. This was done by substituting *P* and *I* into Table 4. The values of risk rank (RR) were 4.5H, 3M, 4H, 3.5H, 1.5L, and 3.5H, as seen in Table 9. The implementation of steps seen in case implantation shows the result for the Local Network with an overall Risk of Value of 3.445 and the overall risk for Cloud as seen in Table 10 with a value of 1.8. For cloud B, the overall risk was 1.53.

Table 9 General Risk Matrix for The Local Network

Service (Risk)	P%	Quantized I	Quantized P	RISK RANK R	Quantized value Rank	Borda P criterion r_{11}	Borda I criterion r_{12}	b_i	risk Wight RW	Overall risk
S1	92	4	5	4.5	H	0	0	9	0.26	0.9
S2	28	4	3	3	M	1	1	4	0.13	1.29
S3	48	5	2	4	H	0	0	8	0.23	0.29
S4	80	4	2	3.5	H	0	1	6	0.17	0.6
S5	36	3	4	1.5	L	1	1	1	0.03	0.045
S6	84	5	5	3.5	H	1	0	6	0.17	0.6
Total								34		3.445

Table 10 General Risk Matrix for Cloud A

Service (Risk)	P%	Quantized I	Quantized P	RR	Quantized value Rank	Borda P criterion r_{11}	Borda I criterion r_{12}	b_i	b_i Wight	0verall risk
CA1	28	1	2	1	L	1	0	1	0.03	0.03
CA2	28	3	2	2	L	1	1	2	0.16	0.32
CA3	32	1	2	2	L	1	0	3	0.25	0.5
CA4	12	1	1	2	L	1	0	3	0.25	0.5
CA5	25	1	2	2	L	0	1	3	0.25	0.5
Total								12		1.8

Table 11 GENERAL RISK MATRIX FOR CLOUD B

Service (Risk)	P%	Quantized I	Quantized P	RR	Quantized value Rank	Borda P criterion r_{11}	Borda I criterion r_{12}	b_i	b_i Wight	Risk Wight RW
CB1	0.24	1	2	1.5	L	0	1	2	2/11=0.18	0.27
CB2	0.16	2	1	1.5	L	1	0	2	0.18	0.27
CB3	0.28	3	2	2	L	1	0	3	0.27	0.54
CB4	0.20	1	2	1.5	L	1	0	2	0.18	0.27
CB5	0.12	1	1	1	L	0	1	2	0.18	0.18
Total								11		1.53

Table 10 and 11, concerning the cloud value CA1 and CB1, represent the values acquired from Table 1 and the data protection agreement. CA2 and CB2 represent internal management risks, while CA3 and CB3 represent internet security. The fourth row of the two tables represents the supervision mechanisms, and the fifth row represents the law and policy. The first column of Tables 9, 10, and 11 represent the strategic element of service we are analyzing. The second column P% represented the risk probability value obtained from equation 10. The Quantized Impact I is the third column and is obtained from Quantizing impact vector using Table 8, while the fourth column Quantized P is obtained from Quantizing probability vector using Table 7. The Quantized Risk value is obtained from Table 4. Table 4 also plays an important role in quantizing both risk impact and probability. The fifth and sixth columns are dedicated to Borda P criterion r_{11} concerning the probability of risk and Borda I criterion r_{12} concerning the impact of risk. Since we are working with two Borda parameters, the impact and the probability has two criteria. These values are set to maximize or minimize the effect of either impact or probability in the final stages of assessment. Column b_i is the Borda count for that element obtained from equation 7. The following column is b_i Wight and is obtained from equation 9. The last column is the calculated completion of the resultant risk-judging matrix. The overall security risk rank is expressed in equation 10. We have the result for the Local Network with an overall Risk of Value of 3.445 and the overall risk for Cloud A seen in Table 10 with a value of 1.8. For Cloud B, the overall risk was 1.53.

Estimating Resultant Risk Using AHP

From the previous section, we find that the overall local network risk is 3.445, where cloud A is 1.8 and Cloud B is 1.53. Let us assume the following:

- Local Network with a value of 3.445 is two times riskier than Cloud A with a 1.8 value; accordingly, Cloud A is 1/3 risky from the local network.
- Local Network with a value of 3.445 is three times riskier than Cloud B with a 1.53 value; accordingly, Cloud A is 1/2 risky from the local network.
- Cloud A and Cloud B are within the same risk margin; therefore, their risk has equal impact and is set to 1.

Based on this assumption, we generate the AHP matrix in Table 11.

Table 12 AHP Priority Matrix

	Local Net.	Cloud A	Cloud B	Operta Criteria	Result	Wight
Local Net.	1	1/2	1/3	$(1 \times 1/2 \times 1/3)^{1/3}$	=0.5505	0.1692
Cloud A	2	1	1	$(2 \times 1/2 \times 1)^{1/3}$	=1.2599	0.3874
Cloud B	3	1	1	$(3 \times 1 \times 1)^{1/3}$	=1.4423	0.4434
				Sum =	3.2525	

We have the following risks with the following weights:

Table 13 AHP Result at Organizational Risk

	Network Risk	Wight	$RRT = \sum_{i=1}^{k} (\mathrm{RR}_i \mathrm{xRW}_i)$
Local Net.	3.445	0.1692	2.067
Cloud A	1.8	0.3874	0.57
Cloud B	1.53	0.4434	0.79
			3.429

The resultant risk for the whole network in terms of cloud services is equal to 3.429.

CONCLUSION

Recently, networks have considered partial or total migration of their services to clouds. This move, which produces new obstacles, presents several risks. Many of the networks run on multiple network connections or wide-area networks of organizational ownership. An organization has two sorts of risks to cope with; firstly, the risk of the organization's internal information systems, and secondly, the risk involved in dealing with cloud service provider companies. Another issue is the lack of benchmarking and references in the information system of risk assessment for enterprises.

Most organizations, rather than risk assessments and mitigation, are working with vulnerability management ideas. In this study, we conceive strategic services for information systems that function simultaneously and hybrid through local network and cloud services spread through local network nodes and cloud components. Regarding local network components and nodes that represent hosts, known vulnerability values created by commercial tools are identified. These vulnerabilities are collected in vectors with anticipated effects and an evaluation of the value of assets associated with such services. Probabilities or risks are therefore recognized.

The other part of the research investigates the computer approach to analyze the potential of cloud services. It addresses common cloud components such as data management policies, internal cloud provider administration, and internet security. The vulnerability of these components and their influence on business continuity in cloud providers is determined. We have presented a risk probability model for an educational organization, using vulnerability ideas for both local and cloud networks. Risks are calculated for both local and cloud-created weights via Borda Count, and the overall risk has been evaluated separately for each component; local network and two clouds. Finally, the organization's entire risk should be assessed jointly by priorities, and each risk should be analyzed in relation to other risks. For this aim, we employ analytical hierarchy (AHP).

References

- *Amin Saurabh, Galina A., Schwartz, & Alefiya Hussain (2013). In quest of benchmarking security risks to cyber-physical systems. IEEE Network Transaction. 27(1)19 - 24*
- *Amro I. (2015). A Network Service-Based Risk Assessment Model with Case Study on an Educational Organization. Palestinian Journal for Open Learning and e-Learning. Volume 15*
- *Andersen A. (2010). Firm objectives, IT alignment, and information securit. IBM Journal of Research and Development.54(3):5.1-5.7*

- *Asosheh A., & Dehmoubed A., & Khani A. (2009). A new quantitative approach for information security risk assessment. Presented in 2nd IEEE International Conference on Computer Science and Information Technology pp. 222-227. China*
- *Bernardo D. (2013). Utilizing Security Risk Approach in Managing Cloud Computing Services. Presented in IEEE 16th International Conference on Network-Based Information Systems. South Korea*
- *George R. (2014). Systems Engineering Guide. The MITRE Corporation. Produced by MITRE Corporate Communications and Public Affairs. USA*
- *International Organization for Standardization ISO (20018) The ISO 27001 standard on information security matters, http:// www.27000.org/*
- *Jianxing Y. , C. Haicheng, W. Shibo & F. Haizhao (2020).A Novel Risk Matrix Approach Based on Cloud Model for Risk Assessment Under Uncertainty. in IEEE Access, vol. 9, pp. 27884-27896, 2021.*
- *Kassou M., & Kjiri L. (2012). SOASMM: A novel service-oriented architecture Security Maturity Model. Presented in IEEE International Conference on Multimedia Computing and Systems, 2012, pp. 912-918. Morroco*
- *Khidzir Nik Zulkarnaen, & Azlinah Mohamed, & Noor Habibah Hj Arshad (2010). Information Security Risk Management: An Empirical Study on the Difficulties and Practices in ICT Outsourcing. Presented in IEEE Second International Conference on Network Applications, Protocols and Services. Malysia.*
- *Lonita D., & Hertel P., & Pieters W., & Wieringa R. (2014). Current Established Risk Assessment Methodologies and Tools. ICT Section, Delft University of Technology. Netherlands*
- *Mačekl Davor, & Magdalenić1 Ivan, & Nina Begičević Ređep1 (2020). A Systematic Literature Review on the Application of Multicriteria Decision Making Methods for Information Security Risk Assessment.International Journal of Safety and Security Engineering Vol. 10, No. 2, pp. 161-174*
- *Maule R. W., & Lewis W. C. (2009). Risk Management Framework for Service-Oriented Architecture. Presented 2009 IEEE International Conference on Web Services. Proceeding pages: 1000-1005. USA*
- *Metzger Louis, & Bender Lisa (2007). MITRE Systems Engineering (SE) Competency Model Version 1.13. The MITRE Corporation. Bedford, MA 01730*
- *Moona Jewook, & Chanwoo Lee, & Sangho Park, & Yanghoon Kimc (2018). Mathematical model-based security management framework for future ICT outsourcing project.Discrete Applied Mathematics The Journal of Combinatorial Algorithms, Informatics, and Computational Sciences.Volume 241: 67-77*
- *Riaz M. T., M. Shah Jahan, K. S. Arif and W. Haider Butt (2019). Risk Assessment on Software Development using Fishbone Analysis. 2019 International Conference on Data and Software Engineering (ICoDSE), 2019, pp. 1-6,*
- *Saleem M, & Jaafar J., & Hassan F.Model Driven Security framework for definition of security requirements for SOA based applications. Presented in IEEE International Conference on Computer Applications and Industrial Electronics, 2010, pp. 266-270, Malasya.*
- *Xi G., H. Ruimin, P. Yongjun, B. Hao and L. Haitao (2010). The Comprehensive Assessment Method for Community Risk Based on AHP and Neural Network Presented in 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing, 2010, pp. 410-413, doi: 10.1109/NSWCTC.2010.230.*
- *Xiao B. and J. Ran. (2010). Risk Evaluation of Network Security Based on NLPCA-RBF Neural Network," in Multimedia Information Networking and Security, International Conference on, Nanjing, Jiangsu China, 2010 pp. 398-402.*
- *Xiaojun Wu and Cong Li (2011). Research and design of one security model for service-oriented multi-application architecture. Presented in IEEE International Conference on Computer Science and Service System (CSSS), pp. 3990-3993. China*
- *Zhang, Q. C. Zhou, Y. Tian, N. Xiong, Y. Qin and B. Hu (2018). A Fuzzy Probability Bayesian Network Approach for Dynamic Cybersecurity Risk Assessment in Industrial Control Systems. in IEEE Transactions on Industrial Informatics, vol. 14, no. 6, pp. 2497-2506*

Unsupervised Machine Learning Method for Researchers' Profiles Matching

طريقة التعلم الآلي غير الخاضع للإشراف لمطابقة ملفات تعريف الباحثين

Thabit Sulaiman Sabbah
Assistant Professor/ Al-Quds Open University/ Palestine
tazazmeh@qou.edu

ثابت سليمان صبّاح
أستاذ مساعد/ جامعة القدس المفتوحة / فلسطين

Received: 21/09/2021, Accepted: 11/10/2021
DOI: https://doi.org/10.33977/2106-000-005-005
https://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 2021/09/21، تاريخ القبول: 2021/10/11
E-ISSN: 2521-411X
P-ISSN: 2520-7431

Abstract

Researcher Profiles Matching is an initial and important step of effective research teams' formation. The researchers' wide, multidisciplinary, and changeable research interests complicate the process of profile matching using traditional methods and affect its performance. This research aims to solve the problem of Profile matching in Scientific Research and Scholarly Work by employing unsupervised machine learning methods. The K-mean clustering method is utilized to categorize researcher profiles based on the statistical analysis of their publication titles, and the correlation-based similarity is employed for profile matching within the categories. The proposed method is implemented, tested, and evaluated using an extracted dataset from Google Scholar. The profile matching results and the clustering quality test result show that the designed task was achieved, in addition to high similarity values of publications within the categories and low correlation values among the clusters. Moreover, the clustering results' analysis can reveal interesting and enlightening information about the scholarly work, which may help the researchers, research management departments, as well as policies and decision-makers in their scholarly work associated tasks.

***Keywords**: Researcher Profiles Matching, Unsupervised Machine Learning, Correlation-based Similarity, K-mean algorithm, Google Scholar.*

الملخص

مطابقة ملفات تعريف الباحثين هي خطوة أولية ومهمة لتشكيل الفرق البحثية الفعالة. إن الاهتمامات البحثية الواسعة ومتعددة التخصصات والمتغيرة للباحثين تُعَقِّد عملية مطابقة الملفات التعريفية باستخدام الأساليب التقليدية، وتؤثر على أدائها. يهدف هذا البحث إلى حل مشكلة مطابقة الملفات الشخصية في مجال البحث العملي، والعمل البحثي من خلال توظيف طرق تعلم الآلة غير الخاضعة للإشراف. واستخدمت طريقة التصنيف (ك-متوسطات) لتصنيف ملفات تعريف الباحثين اعتمادا على التحليل الإحصائي لعناوين أبحاثهم، ووظِف التشابه المبني على الارتباط لمطابقة ملفات التعريف ضمن الفئات. وتم بناء الطريقة المقترحة، وفحصها، ثم قُيِّمت باستخدام مجموعة بيانات مستخلصة من موقع الباحث العلمي ل(جوجل). وأظهرت نتائج مطابقة الملفات الشخصية، وفحص جودة التصنيف أن المهمة المصممة قد تمّ إنجازها، يضاف إلى ذلك ظهور قيم تشابه عالية للأبحاث داخل الفئة وقيم ارتباط متدنية بين الفئات. ويمكن لتحليل نتائج التصنيف أن تكشف معلومات مضيئةً ومهمةً حول العمل البحثي، والتي من شأنها أن تساعد الباحثين، ودوائر إدارة البحث، وصُناع السياسات والقرارات في مهامهم المرتبطة بالعمل البحثي.

الكلمات المفتاحية: مطابقة ملفات تعريف الباحثين، تعلم الآلة غير الخاضع للإشراف، التشابه المعتمد على الارتباط، خوارزمية ك-متوسطات، الباحث العلمي.

INTRODUCTION

Researcher profiles matching is a special case of the general known problem of User Profile matching, which has been tackled in several works over the years. It is a part of the team formation process encouraged by mast organizations to carry out complex tasks (Sun et al., 2009). Many other profits and benefits can be brought to the organization because of effective teams. However, the environment in which the team will be formulated, the task to be accomplished, and many other factors affect the formation process and criticality. Some of these factors are related to the team size, distribution (Milojević, 2014), available data about users (Nurgaliev et al., 2020), and such as the case of team formulation in complex networks and large communities (Sun et al., 2009). On the other hand, from individuals' (researchers) perspectives, researcher profile matching helps in finding potential research collaborators, expertized researchers in a certain domain, expanding network opportunities (Tran et al., 2020), and improving profile building skills (Li et al., 2019).

A research team is defined as a "group of researchers collaborating to produce scientific results, which are primarily communicated in the form of research articles" (Milojević, 2014). A research team may consist of some core researchers and many other researchers who may change over time. Hence there are many works focused on studying the statistical measures of a team such as size, median, and mean, assuming that teams are unchangeable, while fewer studies consider the changeability of teams (Milojević, 2014).

Therefore, several models are proposed in these studies for different cases, aims, bases, and domains such as the Agent-based model (Sun et al., 2009), supervised ML model (Nurgaliev et al., 2020), and others. This work presents the unsupervised machine learning clustering method for researcher profile matching based on researchers' publications metadata available on Google Scholar, such as researcher interests, and publication titles. The rest of this article contains sections about related works, proposed method, methodology, results discussions, and conclusion.

LITERATURE REVIEW

As mentioned earlier, Profiles Matching was a well-known problem that was studied from different aspects over the years in many domains. However, fewer studies were found in the domain of Scientific Research (i.e., matching researchers' profiles to find potential research collaborators and expertized researchers in a joint domain). Therefore, this section summarized the existing works on profile matching and the unsupervised machine learning clustering method utilized by this study.

Profile Matching Works

Profile Matching Algorithm (PMA) was employed in many fields such as business, social networks, and others, following a brief summarization of some studies from different domains.

In the business domain, Sugiarto et al. (2021) described the use of PMA in the context of a decision support system that could help shorten the required time for choosing business partners or potential colleagues in companies. However, the study focused on analyzing input factors of the PMA and the GAP calculations and weightings. The study concluded that the application of PMA based on predetermined conditions could accelerate model calculation and select prospective partners' processes.

Nurgaliev et al. (2020) proposed PMA that dealt with a set of linked nodes from various social networks based on inadequate user profile data such as username and relationship. The proposed framework included two individual algorithms and a combination of them. The proposed User identity linkage (UIL) algorithm aimed to determine mathematically whether any two users on different social networks are the same person in reality. The proposed algorithms were tested on datasets from VK social network and Instagram; the experiments showed relatively high recall and accuracy results.

Eze et al. (2020) presented a configurable PMA in the domain of health community care management. The work aimed to associate common data from various stakeholders to support the process in the domain. Eze et al. (2020) focused on the performance of PMA utilization in the cloud-hosted case study. They tested the proposed model within a pilot project for supporting interoperability between Community Support Service (CSS) provider agencies and the Regional Health Authority (RHA) in Canada. The Proposed PMA consisted of many modules such as feature identification, standardization, match weight summarization, decision, and global identifier generation. The first run of the system was conducted based on about 145,000 user-profiles and took about 35 minutes; however, the sequent daily runs performed the task incrementally and required less than 5 minutes per day.

Similarly, Li et al. (2019) applied the PMA to find the match users' profiles under the condition of restricted data access of users' profile data such as profiles with privacy policies. The proposed method in Li et al. (2019) utilized the public data such as username and display name and accomplished the matching task through a three-step approach, including feature extraction, a two-stage classification framework, and a relationship elimination algorithm. Experimental results on real social networks datasets showed excellent performance and concluded the possibility of applying PMA based on small and public online user profile data.

Paembonan et al. (2018) employed the PMA for drug substitution to facilitate the process of drug substitution in cases of drug lack or exhaustion. The K-means method was utilized to categorize the medicines' profiles to accomplish the task of new medicine recommendations, where the Selection Matching method was employed to control the substitute. The proposed method was tested and evaluated. The authors reported the accuracy of the proposed method was 93.5%.

Earlier, many works have been presented in the field of User Profile Matching, such as (Garcia, 2016; Pizzi and Ukkonen, 2008; Sun et al., 2009; Wassermann and Zimmermann, 2011).

Nevertheless, none of these works was in the field of Scientific Research or Researchers Profile matching and applying any unsupervised machine learning clustering techniques. Although the work of (Paembonan et al., 2018) utilized the k-means algorithm, the work does not explain much about utilizing K-means with PMA. Therefore, this work tried to accomplish the process of PM in the field of Scientific Research by applying some unsupervised machine learning clustering techniques. The following subsection illustrated the principles of unsupervised clustering methods and described the k-means clustering method.

Unsupervised Clustering Methods

Clustering was defined as "the unsupervised classification of data objects into groups or clusters" (Santos et al., 2013). The term "unsupervised" indicated that the process was done under the condition of missing ground-truth labels of classified objects. Therefore, unsupervised clustering methods must first notice any patterns in the data objects being clustered and then group similar objects in a category such that the objects in a group were the most similar to each other. This process of clustering was unlike supervised learning (known as supervised classification), where human experts usually provided the ground-truth labels of the training data. These unsupervised clustering advantages were included but not limited to a slight workload to audit and formulate training data, and superior independence in identifying and utilizing hidden patterns that "experts" had not observed. However, the cost of such benefits included the need for more amount of data for training to achieve acceptable performance which indicated extra storage and computational necessities, as well as the possibility of such method to consider some anomalies or artifacts found in training data as bases of clustering (Delua, 2021). Many methods and techniques were used for clustering such as hierarchical clustering (Franklin, 2005), and k-means which was one of the popular and simplest unsupervised machine learning algorithms (Garbade, 2018).

K-Means Clustering Algorithm

Andrews and Fox (2007) considered this algorithm as the most regular and simple algorithm used for clustering. The algorithm aimed to group the nearest data objects to each other onto smaller sets. A key point for the algorithm was the determination of the number of clusters. After this determination, the algorithm spread the data objects into the determined number of clusters based on objects' features, reflecting the likeness of the data objects (Jain et al., 1999). As mentioned earlier, this clustering method was employed in many fields such as "Topic Detection." For example, Li et al. (2010) performed a study in which the k-means algorithm was employed on top of the Vector Space Model (VSM) representation to detect topics among a corpus. Similarly, Zhang and Li (2011) proposed the k-means clustering method for topic detection in a large-scale dataset. The K-means algorithm was performed by applying the following steps:

1. Determine the number of clusters (the value of k).
2. Randomly select k data objects as preliminary cluster centers (in some implementations, the first K data objects were selected for this step).
3. Calculate the *distance* between the defined cluster centers and the remaining data objects, and assign each data object to a cluster center based on the nearness of the cluster center.
4. For each defined cluster, calculate the mean and update the cluster center to become the calculated mean.
5. If no change occurred to any cluster center values, then STOP, otherwise repeat steps 3-5.

Nevertheless, the k-means clustering method had some downsides, such as its sensitivity to the initial selection of cluster centers, as well as its sensitivity to outliers and noise, and the non-predefined number of clusters. These drawbacks might constitute inaccuracy (Sharma and Gupta, 2012) or unwanted solutions (Jain et al., 1999). However, several techniques were proposed in the literature to overcome these problems. For example, Ray and Turi (1999) recommended the validity measure to determine the k number. Some other works were planned to solve the problem of finding the preliminary cluster centers using different principles, such as Erisoglu et al. (2011), Deelers and Auwatanamongkol (2007), and Redmond and Heneghan (2007).

The distance calculation mentioned in step 3 of the k-mean algorithm differed according to the domain of application. For example, in case that

the data points to be clustered were point 2D or 3D Cartesian coordinate system, the regular distance law between points in such coordinate system and be applied. However, when applying the k-means algorithm in other domains, such as text clustering where the data points represent the documents, the Euclidian distance or the Cosine similarity could be applied. In this research, the algorithm was applied to multi-dimensional feature space. Therefore, the Euclidian Distance Law was applied. The Euclidian distance between two documents represented in a high dimensional feature space was defined as follows:

Let the two data points (i.e., documents) to be A and B, where A and B were vectors of n features such that: A= $\{a_0, a_1, a_2, \ldots, a_n\}$ and B = $\{b_0, b_1, b_2, \ldots, b_n\}$, then the Euclidian distance D between these two data points was calculated according to equation (1).

$$D(A,B) = \sqrt{\sum_{i=0}^{n}(a_i - b_i)^2} \quad (1)$$

The next section explained the proposed method for Researcher Profile Matching.

MATERIALS AND METHODS

This work proposed an Unsupervised Machine Learning Clustering Method for Researcher Profile Matching. The proposed method was based on the analysis of user-profiles data from Google Scholar (GS) Search Engine. A researcher profile on GS contained many informative data portions such as interests, count and distribution of publications over the years, h-index, i10th index, count of citations, and publications list. Nevertheless, some of these elements might be missing or incomplete or not updated in some user profiles. Therefore, some of these elements were utilized in this work for profile matching, especially the publication list, which reflected researcher interests. The next subsections showed the details of the proposed matching method and the dataset used in this work.

Proposed Matching Method

Figure 1 demonstrates the proposed method steps and processes followed by a brief description of the shown steps, where each numbered bounded area was considered as one step, and the method consists of five steps.

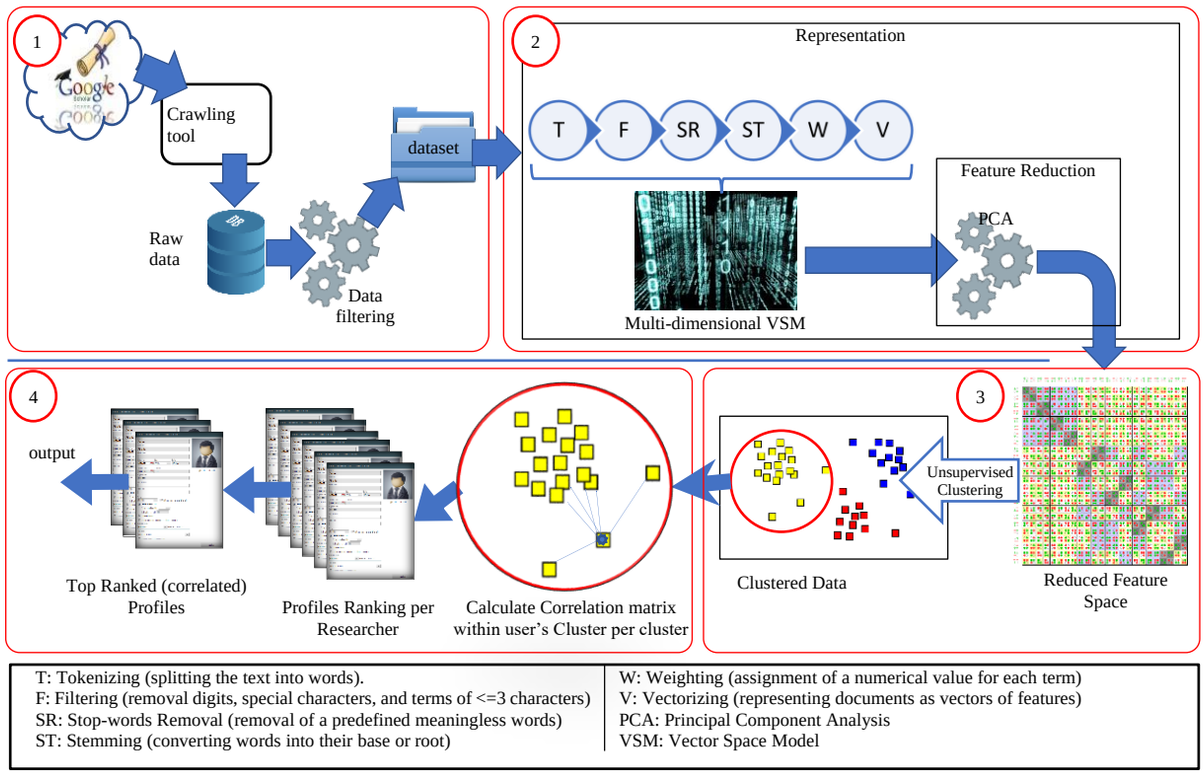


Figure 1 Proposed Researcher Profile Matching Method Steps and Processes

Step 1: This step was devoted to Dataset generation. Dataset description was shown later in the section. In this step, a crawling tool was developed to download hundreds of researcher profiles from GS. These profiles were stored on a local database in HTML format, and then it was processed, filtered, and prepared as the final dataset. The researcher profile on GS contained many portions of data; the distribution of these data chunks on the researcher's profile page was as shown in Figure 2.

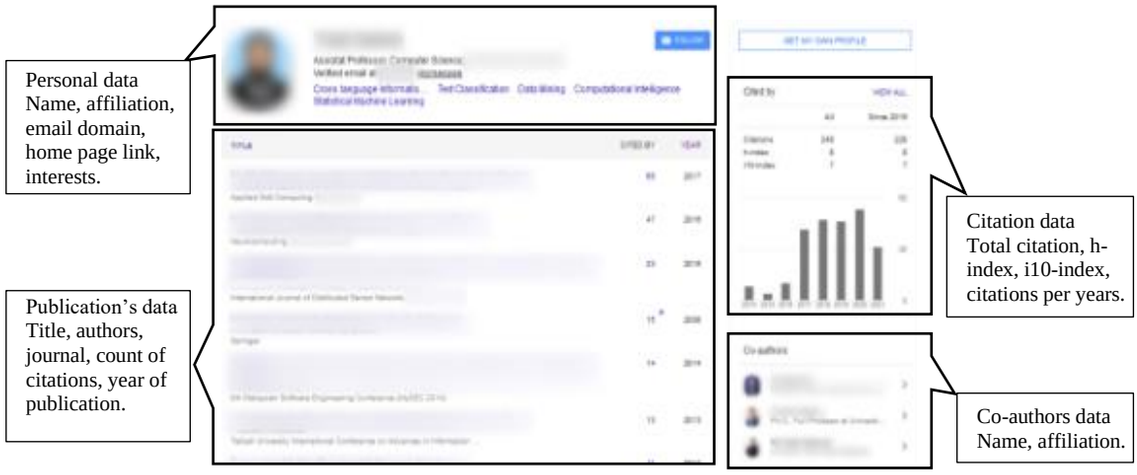


Figure 2 Distribution and Data Chunks on Researcher's Profile Page on GS

However, as mentioned earlier, some of these data chunks might be missing, incomplete, or not updated in some user profiles. Moreover, some of the researcher's publication lists might contain multi-lingual titles. The list consisted of hundreds of publications. Therefore, this step included a filtering process for the publications within the last five years, in which the titles in the English language were considered in the dataset.

Step 2: The dataset was presented numerically to be suitable for the Machine learning methods. The Vector Space Model (VSM) representation was considered in this work. A series of preprocesses tasks were performed for each textual data for each instance in the dataset to achieve this representation. These tasks are: ***Tokenizing, Filtering, Stop-words Removal, Stemming, Weighting, and Vectorizing***. A brief description of these tasks is presented at the bottom of Figure 1. However, regarding **Weighting**, which was the process of assigning a numerical value for each word (term or feature) per dataset instance. This numerical value of a term (known as term weight) represented the importance of that term in that instance. In literature, there were many weighting techniques such as the binary, the Term Frequency (TF), the Term Frequency-Inverse Document Frequency (TF-IDF), and many more (Sabbah et al., 2017). However, this work utilized the Term Occurrence (TO) method that considered the count of term appearance as the term weight. This technique of term weighting, i.e., TO did not consider the normalization of weighing such as the TF and TF-IDF techniques; moreover, it did not show any kind of semantic proximity such as the Term Co-occurrence weighting method. The choice of TO weighting technique in this research was based on the nature of the processed text (i.e., Publication Titles), which was assumed to be clear, specific, and direct to the point.

Vectorizing: In this process, each data sample was represented as a vector of features, where the features of the vector included all the features (terms) contained by the dataset. The vectors were finally collected in one matrix. The rows represented the data samples, the columns represented the features, and the matrix's cells' values represented the weights.

Feature Reduction

The generated VSM based on text vectorization was known as multi-dimensional, in which the count of features was large. For example, during our experiments, the count of features based on the unigram vectorization of publication titles and publication summaries was more than 450,000 features, i.e., unique single word, which was out of our capability to

manipulate due to lack of computational capacity). Therefore, we restricted the textual analysis in this work to publication titles where the count of features in the generated feature space was about 25000 features, which was huge. Therefore, the Principal Component Analyses (PCA) dimensionality reduction method was employed to reduce the dimensionality, reducing computational cost and time.

Step 3: K-means clustering - which was an unsupervised machine method- was a learning method applied to categorize the data samples into clusters or categories where the categories represented the research fields or research topics reflected from publication titles. However, there was a wide range of research fields or topics that could be identified. Thus, the determination of clusters count- that represented the K value in the K-means algorithm- was not an easy task. To do so, the lists of research fields were studied from different sources, as follows:

Table 1 Count of Research Fields from Different Online Sources

List Source	Count of Research fields
Wikipedia: (https://en.wikipedia.org/wiki/Outline_of_academic_disciplines)	1000
Digital Commons Network™: (*https://network.bepress.com*)	1280
Web of Science (WoS): (https://images.webofknowledge.com/images/help/WOS/contents.html)	258
Higher Education Statistics Agency (HESA), UK: (*https://www.hesa.ac.uk/support/documentation/jacs*)	165
Japan Society for the Promotion of Science (JSPS), Japan: (*https://www.jsps.go.jp/english/index.html*)	323

Table 1 showed that the count of research fields was not standard and differed from one source to another, and the count was not enclosed in a small range. Therefore, it was a challenge to determine the count of research fields (i.e., clusters). Nevertheless, there were several computational based techniques to automatically determine the best value of (K), such as the Distortion Analysis (known as Elbow Curve Method) (Yuan and Yang, 2019), Davies-Bouldin Index (Petrovic, 2006), and Calinski-Harabasz Index (Wang and Xu, 2019) and more. Hence, in this study, the results of these techniques were analyzed to determine the best value of K (i.e., clusters count). However, the application of these methods was time-consuming, as the algorithm was required to run numerous times based on various values of K for each technique, which was applicable only for small datasets and K values. However, in our case, the potential value of K was as high as expected by the common sense shown in Table 1, and the dataset size was as big as shown in the dataset subsection. Hence, a sample dataset selected from the study dataset was employed for exploratory study and determination of the count of clusters (i.e., K value for K-means algorithm). The details of the exploratory dataset and the K value determination analysis was shown in the next subsections.

Moreover, Step 3 produced the cluster label for each sample in the dataset. Consequently, these labels were utilized in Step 4 for profile matching.

Step 4: In this step, for each cluster of the identified clusters, the samples that belonged to that cluster were identified and isolated, and then the correlation-based similarity was calculated among all samples within the cluster, the samples such as profiles were ranked, and the top similar correlated profiles were recommended as the best matching profiles for any selected user.

Dataset

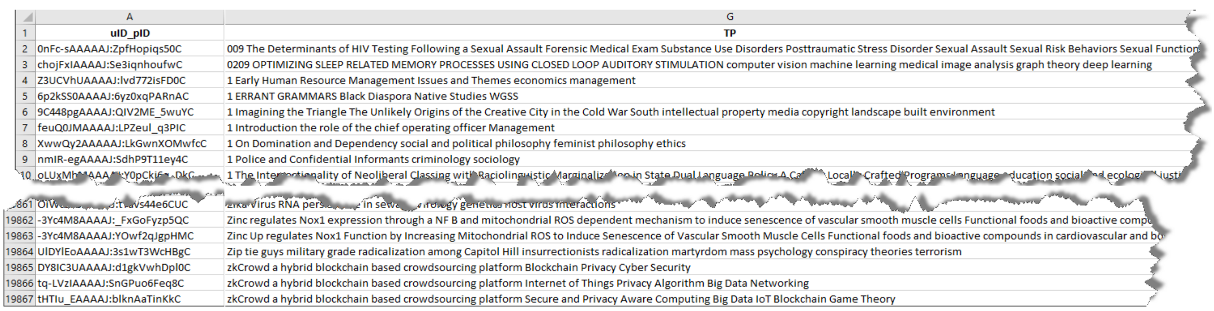
As mentioned in Step 1 description, hundreds of researcher profiles were crawled from GS as Html web pages. The data chunks were extracted from the web pages and filtered. The data chunks that could be utilized are many, such as Researcher's Years of Experience (RYE), h_Index (hI), i10_Index (iI), Publication Age (PA), Publication Citations Count (PCC), Publication Title (PT), and Researcher List of Interests. In addition to the user ID and publication ID (uID:pID) for indexing and matching purposes. However, some of these chunks of data were user-related, such as RYE, hI, iI, and research interests, while others were publication-based, such as PA, PCC, and PT. Therefore, as this study focuses on textual-based categorization and profile matching, the publication-based chunks of data were considered in the dataset, especially the Publication Title (PT). Nevertheless, Regarding the Researcher List of Interests, it was noticed during preprocessing that the keywords included in the List of Interests of researchers contained noise data such as spelling mistakes and

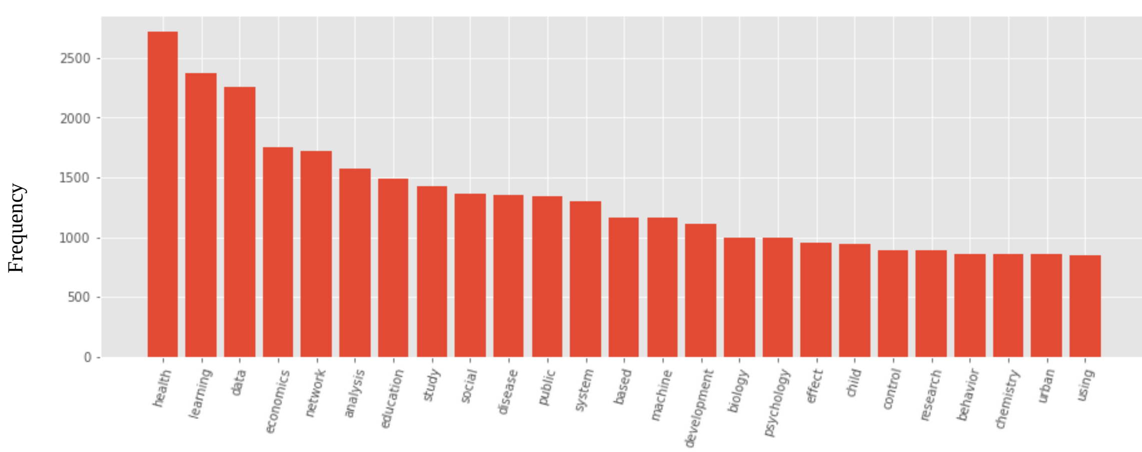
sometimes not. Therefore, the interests' keywords were treated in this work as "text" and added to each publication's title found in the researcher profile. Initially, the dataset represented the data of 1351 researchers from Georgia State University (GSU) and included the data of 22540 publications. However, as a part of data filtering mentioned in Step 1, the Researcher Profiles, which included a very large or very low count of publications (outliers), were eliminated; for experimental purposes, the elimination was performed using a simple query method. The remaining profiles contained a count of publications ranging from 6 up to 2239 publications. Table 2 shows the statistical information of the final dataset. Figure 3 shows a portion of records in CSV format, while Figure 4 and Figure 5 show the most frequent words and Bigrams used in the publication Titles included in the final dataset.

Table 2 Dataset Statistics

Count of Users (Researches)	882	
Count of Publications	19866	
Publication Titles		
Unique Vocabulary in Publication Titles	18350	
Total count of words in publication Titles	269854	
Average words count per Title	14.94	
Publication Title Statistics	Publications Count	Title Length (words)
Average	44.99	143.01
Min.	6	10
Max.	239	312

As seen in Table 2, the final dataset contained the data of 882 Researchers and included the titles of 19866 publications. The eliminated profiles, i.e., the profiles which included a very large or very low count of publications, perform about 34% of the total profiles count. However, the effect of this elimination in terms of computational cost, performance, and time was not studied in this research as this research aimed to prove the concept of the proposed method.


A
G
1 uID_pID TP
2 0nFc-sAAAAAJ:ZpfHopiqs50C 009 The Determinants of HIV Testing Following a Sexual Assault Forensic Medical Exam Substance Use Disorders Posttraumatic Stress Disorder Sexual Assault Sexual Risk Behaviors Sexual Function
3 chojFxIAAAAJ:Se3iqnhoufwC 0209 OPTIMIZING SLEEP RELATED MEMORY PROCESSES USING CLOSED LOOP AUDITORY STIMULATION computer vision machine learning medical image analysis graph theory deep learning
4 Z3UCVhUAAAAJ:lvd772isFD0C 1 Early Human Resource Management Issues and Themes economics management
5 6p2kSS0AAAAJ:6yz0xqPARnAC 1 ERRANT GRAMMARS Black Diaspora Native Studies WGSS
6 9C448pgAAAAJ:QIV2ME_5wuYC 1 Imagining the Triangle The Unlikely Origins of the Creative City in the Cold War South intellectual property media copyright landscape built environment
7 feuQ0JMAAAAJ:LPZeul_q3PIC 1 Introduction the role of the chief operating officer Management
8 XwwQy2AAAAAJ:LkGwnXOMwfcC 1 On Domination and Dependency social and political philosophy feminist philosophy ethics
9 nmIR-egAAAAJ:SdhP9T11ey4C 1 Police and Confidential Informants criminology sociology
19862 -3Yc4M8AAAAJ:_FxGoFyzp5QC Zinc regulates Nox1 expression through a NF B and mitochondrial ROS dependent mechanism to induce senescence of vascular smooth muscle cells Functional foods and bioactive
19863 -3Yc4M8AAAAJ:YOwf2qJgpHMC Zinc Up regulates Nox1 Function by Increasing Mitochondrial ROS to Induce Senescence of Vascular Smooth Muscle Cells Functional foods and bioactive compounds in cardiovascular and
19864 UlDYlEoAAAAJ:3s1wT3WcHBgC Zip tie guys military grade radicalization among Capitol Hill insurrectionists radicalization martyrdom mass psychology conspiracy theories terrorism
19865 DY8IC3UAAAAJ:d1gkVwhDpl0C zkCrowd a hybrid blockchain based crowdsourcing platform Blockchain Privacy Cyber Security
19866 tq-LVzIAAAAJ:SnGPuo6Feq8C zkCrowd a hybrid blockchain based crowdsourcing platform Internet of Things Privacy Algorithm Big Data Networking
19867 tHTIu_EAAAAJ:blknAaTinKkC zkCrowd a hybrid blockchain based crowdsourcing platform Secure and Privacy Aware Computing Big Data IoT Blockchain Game Theory
Figure 3 A Snap of Dataset in CSV Format


Frequency
2500
2000
1500
1000
500
0
health
learning
data
economics
network
analysis
education
study
social
disease
public
system
based
machine
development
biology
psychology
effect
child
control
research
behavior
chemistry
urban
using
Figure 4 Most Frequent Words used in Publication Titles

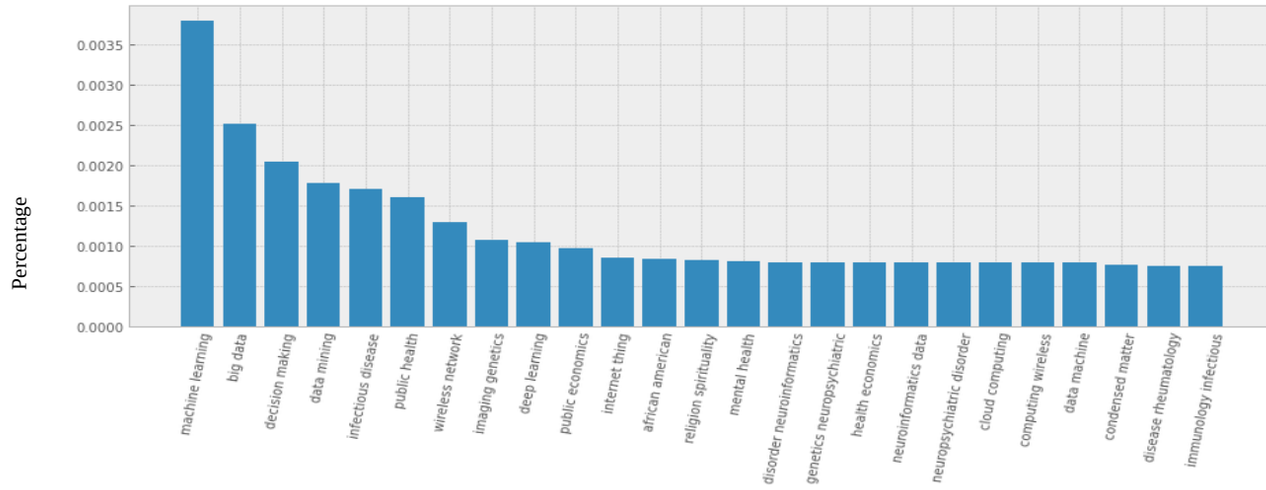


Figure 5 Most Frequent Bigrams used in Publication Titles

As mentioned in Step 2, the PCA method was utilized to reduce the size of feature space generated by the VSM. The application of the PCA algorithm in this step reduced the size of the feature space by about 78%, such that the number of features was reduced from 18350 features in the VSM feature space to 4217 features in the reduced feature space. However, the size of the reduced feature space was selected to represent about 95% of the original feature space. Figure 6 shows the size of reduced feature space after PCA application.

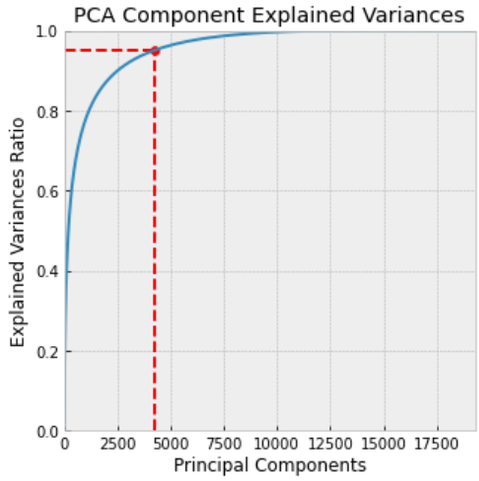


Figure 6 Size of The Reduced Feature Space after PCA Application

As seen from Figure 6, initially, the count of principal components was equal to the count of features in the VSM feature space. However, the curve showed the cumulative variance explained by these components. The cumulative variance of the selected count of components, i.e., 4217, explained 95% of the feature space.

Exploratory Dataset

As the Step 3 discussion mentioned, the researcher employed an exploratory dataset to determine the K-value required for the K-means algorithm. Then selected the exploratory dataset to be representative and informative. Therefore, for each researcher among the 882 researchers included in the study dataset, three publications were selected so that the top three cited publications were included in the exploratory dataset. The selected publications per researcher (i.e., top-cited publications) were expected to be the nearest (or representing the field of study of the researcher). Table 3 showed the statistics of the exploratory dataset. Figure 7 and Figure 8 showed the most frequent words and Bigrams used in the publication Titles included in the exploratory dataset.

Table 3 Exploratory Dataset Statistics

Count of Users (Researches)		882
Count of Publications		2646
Publication Titles		
Unique Vocabulary in Publication Titles		7153
Total count of words in publication Titles		38723
Average words count per Title		14.63
Publication Title Statistics	**Publications Count**	**Title Length (words)**
Average	3	14.63
Min.	3	1
Max.	3	36

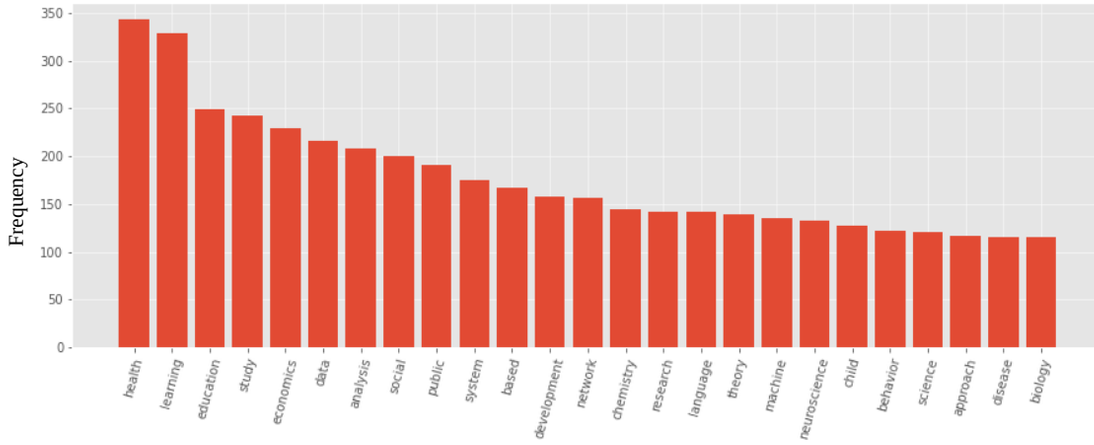


Figure 7 Most Frequent Words used in Publication Titles in The Exploratory Dataset

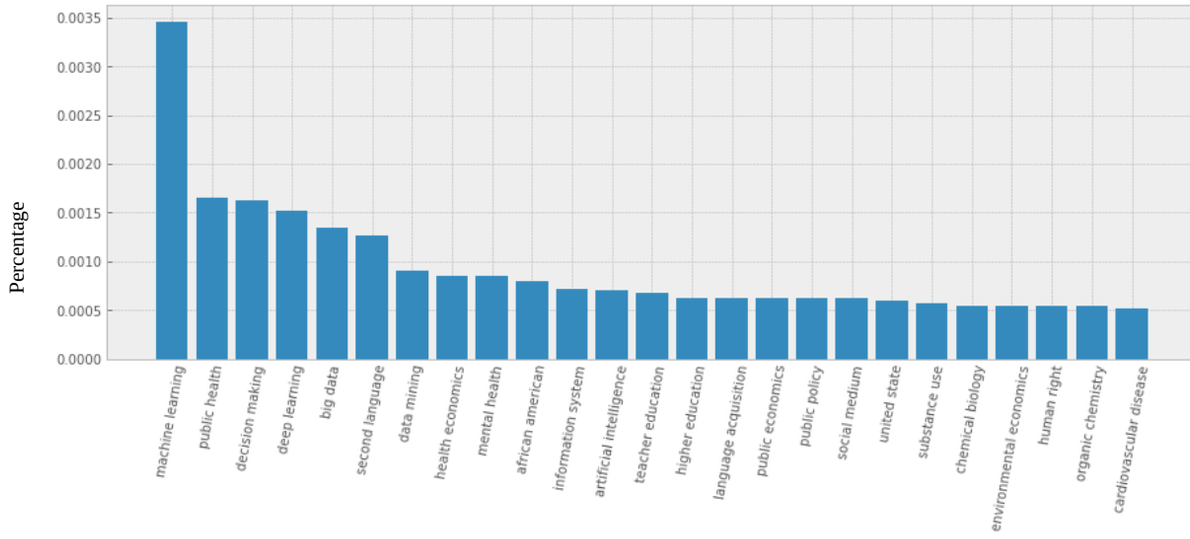


Figure 8 Most Frequent Bigrams used in Publication Titles in The Exploratory Dataset

RESULTS AND DISCUSSIONS

The proposed method was implemented in Python 3.8, while the experiments were conducted under Windows 10 environment, and the results were analyzed using a collection of tools including Orange and MS-Excel. Tasks of clustering and matching were accomplished.

However, as it was known about clustering methods, the evaluation of the clustering was as difficult as the clustering itself (Pfitzner et al., 2008). The proposed method in this work tried to solve the problem of Profile Matching through the application of clustering, an unsupervised machine learning method. Nevertheless, none of the problems - i.e., the profile matching and the clustering- in the domain of consideration had a gold standard dataset to evaluate the results of the proposed method. Therefore, the internal method of evaluation (Feldman and Sanger, 2006) was applied in which the internal clustering quality measures were analyzed, then the corresponding results of the proposed method were benchmarked. Recall that the scope of this work did not include Topic detection, i.e., the method was not responsible for knowing the Research Field of a researcher or publication. However, the clusters or categories in this research represented the Research Fields. Therefore, in this section the clusters were presented by their given numbers: 0, 1 … and so on. Following are the major finding based on the analysis of the results.

As mentioned earlier, the count of clusters considered in this work was determined based on the analysis of the three different clustering quality techniques results on an exploratory dataset; the next subsection shows this analysis' results.

K-value Determination

To determine the optimal value of K, the algorithm was run with K value ranges from 100

to 310. The clustering quality measures; Distortion, the base of Elbow Curve analysis, Davies-Bouldin Index, and Calinski-Harabasz Index, were recorded, scaled, and plotted as shown in Figure 9.

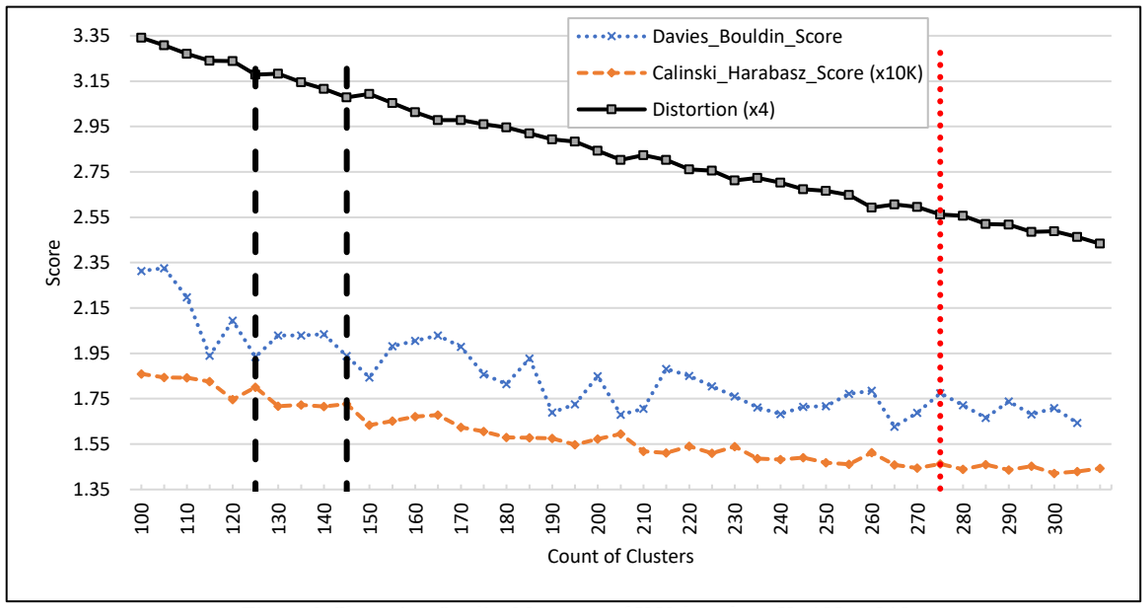


Figure 9 Clustering Quality Measures of K Values from K=100 to 310

Figure 9 showed that several potential K values produce satisfactory quality measures and can be considered as the count of clusters. Based on the assumptions behind these three measures, the value K = 275 was selected as the cluster count in this research. Moreover, a further Kolmogorov Smirnov test (Wilcox, 2017) at that value of K, i.e., 275 showed that the distribution of samples among the clusters fit the normal distribution with a value of $p < 0.05$. The next subsection presented the analysis of results from two-point views: Researchers' Distributions and Publications Distribution against Research Fields, i.e., Categories or Clusters.

Researchers Distributions Analysis

Figure 10 shows the distribution of the count of Researchers among these clusters.

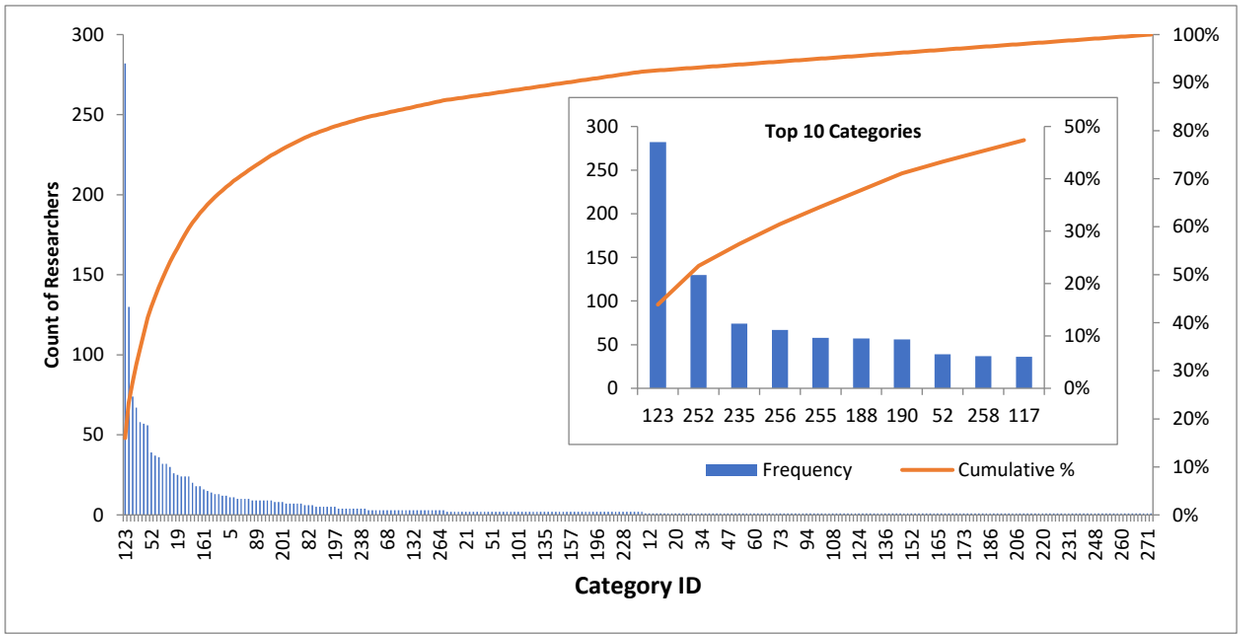


Figure 10 Researchers Distribution among Clusters (Research Fields)

Figure 10 shows the distribution of researchers among the 275 considered clusters. The Frequency columns represented the count of researchers belonging to the corresponding cluster from the horizontal axis. The distribution of the researchers among the "Top 10 Categories" was shown in the internal subplot ("Top 10 Categories"). For example, there were about 282 researchers grouped in "Cluster 123", while less than half of this count of researchers 130 in "Cluster 252", for the remaining clusters, the count of researchers ranges between 1 and 75 researchers. Moreover, the "Top 10 Clusters" included about 50% of the researchers' distribution, whereas the 9th and 10th clusters contained less than 50 researchers each. It is worth mentioning that researchers' categorization (clustering) was based on their scholarly production within the last five years. Therefore, some (if not many) researchers were identified to be included in multiple clusters, which reflected the multidisciplinary nature of many researchers. Figure 11 shows the multidisciplinary distribution identified in the considered dataset.

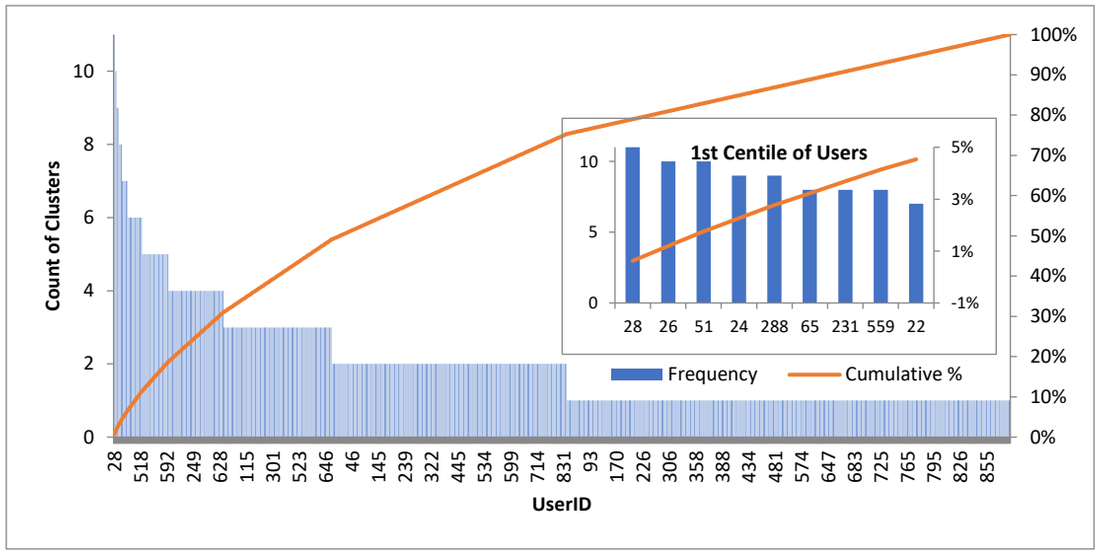


Figure 11 Multidisciplinary Researchers' Distribution

From the multidisciplinary distribution of researchers shown in Figure 11, very few researchers were categorized as involved in abundant research fields; more than 7 fields as revealed in the inner subplot, i.e., the 1st centile of users' multidisciplinary distribution. This portion could be caused by outlier profiles in which a huge number of publications were added automatically to a researcher profile because of the known problem of initials ambiguity of researcher names (Milojević, 2013). Figure 12 shows the clusters per user distribution.

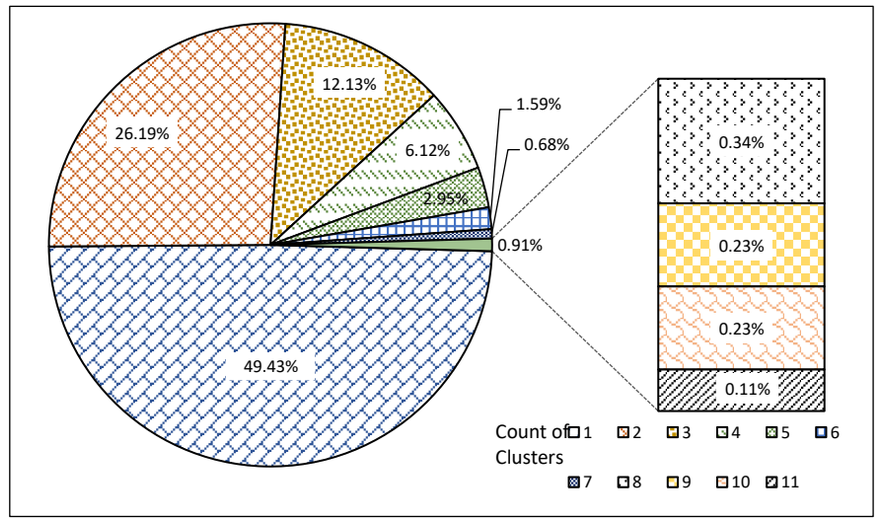


Figure 12 Researchers Distribution among Research Fields

Figure 12 showed that the majority, about 87.5% of researchers, were identified to be working within limited research fields at most 3 disciplines. 49.43% were mono disciplinary researchers, 26.19% were involved in two disciplines, and 12.13% were contributary to three research disciplines. About 11% of researchers were identified to be involved in -4 to 7- research fields and the remaining less than 1% of researchers were involved in abundant research fields as described earlier.

Publications Distribution Analysis

Figure 13 showed the publications distribution among the research fields.

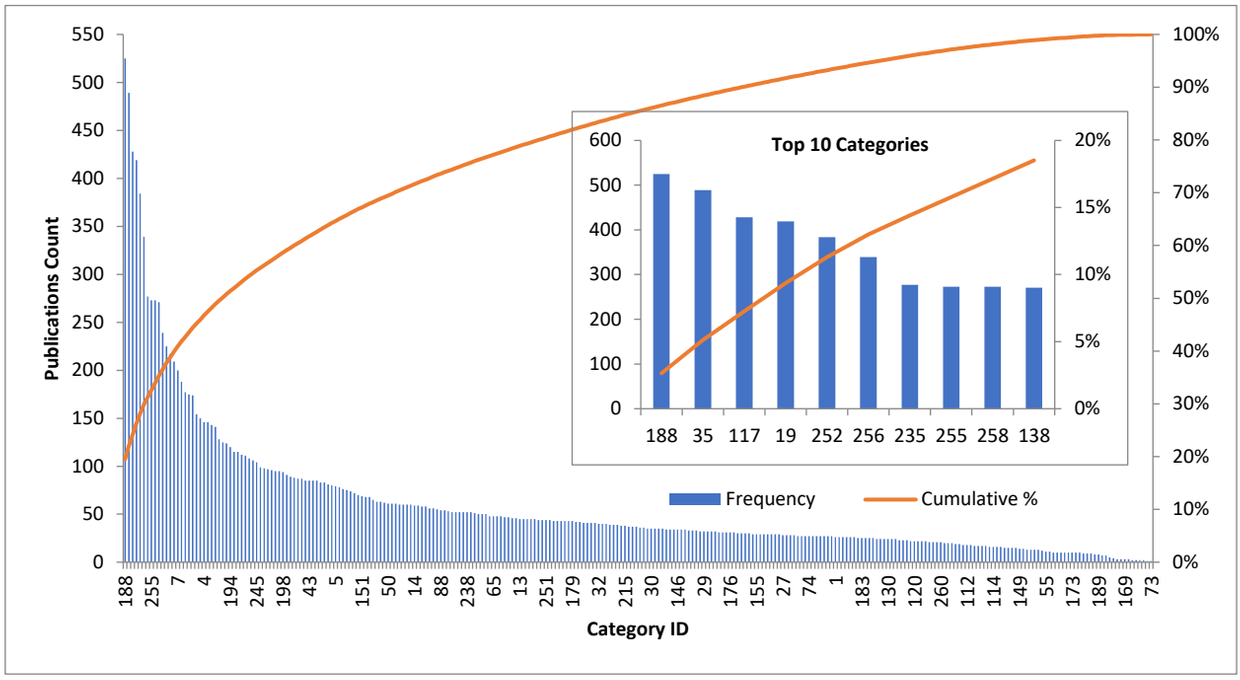


Figure 13 Publications Distribution among Clusters (Research Fields)

Figure 13 shows the distribution of Publications among the 275 Research fields considered in this study. The Frequency columns represented the count of publications that belonged to the corresponding cluster from the horizontal axis. The distribution of publications among the top 10 Categories was shown in the inner subplot Top 10 Categories. For example, there were about 500 publications grouped in the 1st and 2nd clusters of the top 10, i.e., Cluster 188 and Cluster 35. In comparison, the 3rd to 6th clusters contained about 330-430 publications and less than 300 publications per each of the remaining clusters. Moreover, the Top 10 Clusters included about 20% of publications' distribution. It is worth mentioning that the clustering method proposed in this work was not designed to categorize a single publication in more than one category. Therefore, there was no multidisciplinary distribution of publications.

Clustering Quality Test Result

The proposed clustering method in this work was tested on a non-public dataset that suffered from the absence of ground-truth labels; this was because of the lack of such studies in this field. Hence, this case complicated the evaluation of the performance of the clustering method and the proposed profile matching approach. However, the presented "K-value determination" subsection showed that the performance of the clustering method at K=275 is the best among the tested values of K, as well as the statistical Kolmogorov Smirnov test at that value of K, i.e., 275, showed that the distribution of samples among the clusters fits the normal distribution with a value of $p < 0.05$. These results indicated the performance quality of the proposed clustering method. Moreover, the correlation between the resulted clusters was tested. Figure 14 shows a heatmap diagram that visualized the correlation analysis among the identified clusters.

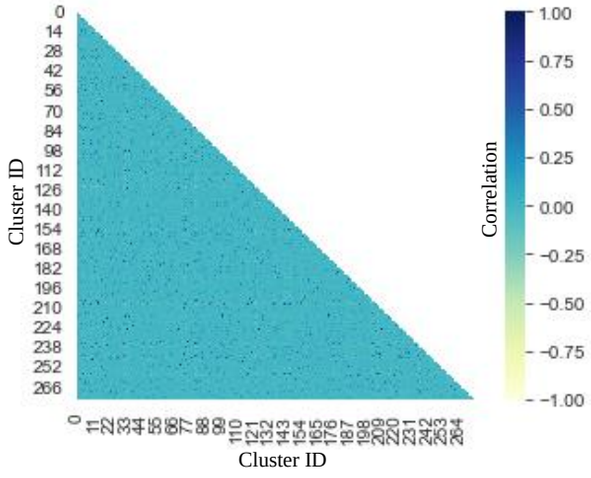


Figure 14 Visualization of Correlation between The Generated Clusters

Figure 14 showed that very few clusters were highly correlated dots in dark blue color. In contrast, the correlation between the majority of clusters ranged between -0.25 – 0.25, which indicated a good separation between clusters.

Profile matching Results

In addition to the clustering process, the proposed method aimed to match the researcher profiles through correlation-based similarity. For each identified cluster, the matrix of correlation between all researchers' publications within the cluster was calculated. The top similar publications were selected, and the researchers were proposed to be the best matching profiles of the selected researcher. As a result of this process, each researcher would be associated with some other researchers based on the similarity of their publication. Figure 15 shows a sample of the results of this process. Table 4 shows the description of the columns in Figure 15 starting from the left, which illustrates the output of the proposed method.

ResearcherProfileID	Group	Rec. Res. ProfileID	Most Similar Publication ID	Similarity	Publication Title
ec72EnIAAAAJ	5	5bnkmZUAAAAJ	5bnkmZUAAAAJ:t7zJ5fGR-2UC	0.899	changing perception harm e cigarette among u adult public health chronic disease tobacco use
		5bnkmZUAAAAJ	5bnkmZUAAAAJ:LO7wyVUgiFcC	0.899	changing perception harm e cigarette v cigarette use among adult u national survey public healtl
		vP6PjG4AAAAJ	vP6PjG4AAAAJ:a0OBvERweLwC	0.856	changing perception harm e cigarette among u adult tobacco global health
		0iqZYjcAAAAJ	0iqZYjcAAAAJ:e5wmG9Sq2KIC	0.310	relationship chronic lung disease status e cigarette use potential influence excessive alcohol use
		0iqZYjcAAAAJ	0iqZYjcAAAAJ:4JMBOYKVnBMC	0.310	dual use electronic cigarette traditional cigarette among adult psychosocial correlate associated r
ec72EnIAAAAJ	123	nmIR-egAAAAJ	nmIR-egAAAAJ:9vf0nzSNQJEC	0.894	benefit working informant criminology sociology
		nmIR-egAAAAJ	nmIR-egAAAAJ:N5tVd3kTz84C	0.894	working informant criminology sociology
		iirGxtEAAAAJ	iirGxtEAAAAJ:dshw04ExmUIC	0.913	protein secondary structure analysis dried blood serum using infrared spectroscopy identify mark
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:efIP2zaiRacC	0.953	response drug resistant epilepsy adult outcome trajectory failure two medication biostatistics
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:08ZZubdj9fEC	0.953	drug resistant epilepsy adult outcome trajectory failure two medication biostatistics
ec72EnIAAAAJ	82	5bnkmZUAAAAJ	5bnkmZUAAAAJ:b1wdh0AR-JQC	0.899	motif perception regarding electronic nicotine delivery system end use among adult mental healt
		vP6PjG4AAAAJ	vP6PjG4AAAAJ:t7zJ5fGR-2UC	0.896	use electronic nicotine delivery system end among chinese adult evidence citywide representati
		vP6PjG4AAAAJ	vP6PjG4AAAAJ:_axFR9aDTf0C	0.896	use electronic nicotine delivery system end china evidence citywide representative survey five cl
		SoO1Xm0AAAAJ	SoO1Xm0AAAAJ:UeHWp8X0CEIC	0.826	motif perception regarding electronic nicotine delivery system end use among adult mental healt
		SoO1Xm0AAAAJ	SoO1Xm0AAAAJ:KlAtU1dfN6UC	0.826	use electronic nicotine delivery system among adult mental health condition
ec72EnIAAAAJ	252	cNsXS68AAAAJ	cNsXS68AAAAJ:isC4tDSrTZIC	0.972	sankofa go back fetch merging genealogy africana study introduction literature humanity african a
		cNsXS68AAAAJ	cNsXS68AAAAJ:RGFaLdJalmkC	0.972	sankofa go back fetch merging genealogy africana study literature humanity african american stud
		SIhUdiEAAAAJ	SIhUdiEAAAAJ:roLk4NBRz8UC	0.806	intercultural communication education sla intercultural competence study abroad culture hyperm
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:kuK5TVdYjLIC	0.214	study statistic knowledge among nurse faculty school research doctorate program biostatistics
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:Y5dfb0dijaUC	0.214	mixed method study contraceptive effectiveness relationship context among young adult primari
EHSVmZ8AAAAJ	2	hEqmxx8AAAAJ	hEqmxx8AAAAJ:MLfJN-KU85MC	0.945	reciprocal relationship depressive symptom employment status labor economics demography ho
		hEqmxx8AAAAJ	hEqmxx8AAAAJ:z_wVstp3MssC	0.945	bilateral relationship depressive symptom employment status labor economics demography hou

Figure 15 Sample of Profile Matching Results

Table 4 Description of Output Data

Column Header	Description
ResearcherProfileID	The researcher Profile ID on GS.
Group	The Category ID(s) (i.e., Research Fields) as detected by the clustering method, note that some researchers are identified to be working in multiple research fields
Rec. Res. ProfileID	The researcher profile IDs whom were detected as top matched researches by the method.
Most Similar Publication ID	The GS id of the publication that belongs to the matched users.
Similarity	The similarity value (correlation) between the identified publication and the publications of the researches.
Publication Title	The publication title form GS.

Figure 15 showed that the proposed method was able to identify the top matched profiles of a Researcher based on the textual analysis of publication titles included in researchers' profiles on GS. The output showed that the publication titles in each group were similar as they had several common words, which were indeed similar to some publications in the Researcher profile under inspection. Additionally, some researchers were categorized in multiple categories where

each category, i.e., groups included similar publications from various user profiles. Furthermore, some identified publications had low similarity values in some groups marked in red color in Figure 15. However, some threshold cut value could be set for such cases to exclude such publication from the group if needed.

RESEARCH CONTRIBUTIONS

This research contributed to the domain by the following:

- The employment of the Unsupervised Machine Learning for solving the Researcher Profiles clustering problem.
- The employment of the correlation-based similarity for solving the Researcher Profiles matching problem.
- The analysis of results revealed hidden information about the scholarly work represented in the considered dataset. However, any institution could reveal such information using the same methods and analysis

CONCLUSIONS

This research aimed to solve the problem of profile matching in Scientific Research and Scholarly Work by employing unsupervised machine learning methods. The Vector Space Model (VSM) based on the term count vectorization and the PCA feature reduction methods were used to represent the data for the proposed machine learning method. Then, the K-mean clustering method was utilized to carry out the task of grouping or clustering the researcher profiles based on the statistical analysis of publication titles of the researchers. The correlation-based similarity was employed for profile matching within the clusters. The method was tested on an extracted dataset from Google Scholar. After preprocessing and filtering, the dataset contains the publication titles of 19866 publications which belong to 882 researchers from Georgia State University (GSU). The publications were categorized into 275 categories, i.e., Research Fields based on the analysis of clustering quality measures Distortion, Davies-Bouldin Index, Calinski-Harabasz Index, and the Kolmogorov Smirnov test. The proposed methods were implemented in python, and the analysis of the results revealed statistical information about the dataset. Moreover, the profile matching results and the clustering quality test result showed that the proposed method accomplished the designed task with high similarity of publications within the clusters and low correlation values among the clusters. The future direction of the research in this field included but was not limited to working on multi-lingual and larger datasets, testing various weighting methods, unsupervised machine learning, quality performance measures or studying the effect of dataset size and quality results generalization.

References

- *Andrews, N. O., and Fox, E. A. (2007). Recent Developments in Document Clustering: Department of Computer Science, Virginia Polytechnic Institute & State ….*
- *Deelers, S., and Auwatanamongkol, S. J. I. J. o. C. S. (2007). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning Along the Data Axis with the Highest Variance. 2(4): 247-252.*
- *Delua, J. (2021). Supervised Vs. Unsupervised Learning: What's the Difference? Artificial intelligence Retrieved 05/09/2021, 2021, from https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning*
- *Erisoglu, M., Calis, N., and Sakallioglu, S. (2011). A New Algorithm for Initial Cluster Centers in K-Means Algorithm. Pattern Recognition Letters. 32(14): 1701-1705.*
- *Eze, B., Kuziemsky, C., and Peyton, L. (2020). A Configurable Identity Matching Algorithm for Community Care Management. Journal of Ambient Intelligence and Humanized Computing. 11(3): 1007-1020.*
- *Feldman, R., and Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press.*
- *Franklin, J. (2005). The Elements of Statistical Learning: Data Mining, Inference and Prediction. The Mathematical Intelligencer. 27(2): 83-85.*
- *Garbade, M. J. (2018). Understanding K-Means Clustering in Machine Learning. Towards Data Science Retrieved 05/09/2021, 2021, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1*
- *Garcia, P. E. (2016). Hybrid Algorithm for Matching Profiles and Social Networks.*
- *Jain, A. K., Murty, M. N., and Flynn, P. J. J. A. c. s. (1999). Data Clustering: A Review. 31(3): 264-323.*
- *Li, S., Lv, X., Wang, T., and Shi, S. (2010). The Key Technology of Topic Detection Based on K-Means. 2010 International Conference on Future Information Technology and Management Engineering. 387-390.*
- *Li, Y., Peng, Y., Zhang, Z., Yin, H., and Xu, Q. (2019). Matching User Accounts across Social Networks Based on Username and Display Name. World Wide Web. 22(3): 1075-1097.*
- *Milojević, S. (2013). Accuracy of Simple, Initials-Based Methods for Author Name Disambiguation. Journal of Informetrics. 7(4): 767-773.*

- *Milojević, S. (2014). Principles of Scientific Research Team Formation and Evolution. Proceedings of the National Academy of Sciences. 111(11): 3984-3989.*
- *Nurgaliev, I., Qu, Q., Bamakan, S. M. H., and Muzammal, M. (2020). Matching User Identities across Social Networks with Limited Profile Data. Frontiers of Computer Science. 14(6): 146809.*
- *Paembonan, S., Manga, A. R., Jusmidah, Atmajaya, D., Waluyantari, A. V., Astuti, W., and Mansyur, S. H. (2018). Combination of K-Means and Profile Matching for Drag Substitution. 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT). 6-7 Nov. 2018. 180-183.*
- *Petrovic, S. (2006). A Comparison between the Silhouette Index and the Davies-Bouldin Index in Labelling Ids Clusters. Proceedings of the 11th Nordic Workshop of Secure IT Systems. 53-64.*
- *Pfitzner, D., Leibbrandt, R., and Powers, D. (2008). Characterization and Evaluation of Similarity Measures for Pairs of Clusterings. Knowledge and Information Systems. 19(3): 361.*
- *Pizzi, C., and Ukkonen, E. (2008). Fast Profile Matching Algorithms — a Survey. Theoretical Computer Science. 395(2): 137-157.*
- *Ray, S., and Turi, R. H. (1999). Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. Proceedings of the 4th international conference on advances in pattern recognition and digital techniques. 137-143.*
- *Redmond, S. J., and Heneghan, C. (2007). A Method for Initialising the K-Means Clustering Algorithm Using Kd-Trees. Pattern Recognition Letters. 28(8): 965-973.*
- *Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., and Fujita, H. (2017). Modified Frequency-Based Term Weighting Schemes for Text Classification. Applied Soft Computing. 58: 193-206.*
- *Santos, R. S., Malheiros, S. M. F., Cavalheiro, S., and de Oliveira, J. M. P. (2013). A Data Mining System for Providing Analytical Information on Brain Tumors to Public Health Decision Makers. Computer Methods and Programs in Biomedicine. 109(3): 269-282.*
- *Sharma, S., and Gupta, V. J. I. J. o. C. A. (2012). Recent Developments in Text Clustering Techniques. 37(6): 14-19.*
- *Sugiarto, I., Diyasa, G. S. M., and Idhom, M. (2021). Profile Matching Algorithm in Determining the Position of Colleagues. Journal of Physics: Conference Series. 1844(1): 012026.*
- *Sun, C., Wan, Y., and Chen, Y. (2009). Dynamics of Research Team Formation in Complex Networks. Complex Sciences. 2009//. Berlin, Heidelberg. 2004-2015.*
- *Tran, N.-Y., Chan, E. K. J. C., and Libraries, R. (2020). Seeking and Finding Research Collaborators: An Exploratory Study of Librarian Motivations, Strategies, and Success Rates. 81(7): 1095.*
- *Wang, X., and Xu, Y. (2019). An Improved Index for Clustering Validation Based on Silhouette Index and Calinski-Harabasz Index. IOP Conference Series: Materials Science and Engineering. 569: 052024.*
- *Wassermann, B., and Zimmermann, G. (2011). User Profile Matching: A Statistical Approach. CENTRIC 2011, The fourth international conference on advances in human-oriented and personalized mechanisms, technologies, and services. 60-63.*
- *Wilcox, R. (2017). Comparing Two Groups. In: R. Wilcox (ed.). Introduction to Robust Estimation and Hypothesis Testing (Fourth Edition) (pp. 145-234): Academic Press.*
- *Yuan, C., and Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. 2(2): 226-235.*
- *Zhang, D., and Li, S. (2011). Topic Detection Based on K-Means. 2011 International Conference on Electronics, Communications and Control (ICECC). 2983-2985.*